REPUBLIKA E SHQIPËRISË

UNIVERSITETI POLITEKNIK I TIRANËS

FAKULTETI I TEKNOLOGJISË SË INFORMACIONIT

DEPARTAMENTI I INXHINIERISË INFORMATIKE

EVIS TRANDAFILI

Për marrjen e gradës
"Doktor"
në "Teknologjitë e Informacionit dhe Komunikimit"
drejtimi "Inxhinieri Informatike"

DISERTACION

NJË ALGORITËM I RI QË KOMBINON METODA STATISTIKORE DHE METODA TË MACHINE LEARNING PËR NXJERRJEN E AUTOMATIZUAR TË FJALËVE KYÇE NGA TEKSTE NË GJUHËN SHQIPE

Udhëheqës Shkencor
PROF. DR. ALEKSANDËR XHUVANI

Tiranë, 2014

NJË ALGORITËM I RI QË KOMBINON METODA STATISTIKORE DHE METODA TË MACHINE LEARNING PËR NXJERRJEN E AUTOMATIZUAR TË FJALËVE KYÇE NGA TEKSTE NË GJUHËN SHQIPE

Disertacioni

i paraqitur në Universitetin Politeknik të Tiranës

për marrjen e gradës

"Doktor"

në

"Teknologjitë e Informacionit dhe Komunikimit"

drejtimi Inxhinieri Informatike

nga

Znj. Evis Trandafili

2014

Disertacioni i shkruar nga

Znj. Evis Trandafili

DIND, Universiteti i Tiranës, Shqipëri, 2008

I aprovuar nga

Juria

_____ Kryetari, Juria e doktoratës

_____ Anëtar, Juria e doktoratës

_____ Anëtar, Juria e doktoratës

_____ Anëtar, Juria e doktoratës

_____ Anëtar, Juria e doktoratës

I pranuar nga

_____ Dekan, Fakulteti i Teknologjisë së Informacionit

Miratuar nga

Akademik Jorgaq Kaçani, Rektor i UPT

_____, Këshilli i Profesorëve, FTI

Abstrakt

NJË ALGORITËM I RI QË KOMBINON METODA STATISTIKORE DHE METODA MACHINE LEARNING PËR NXJERRJEN E AUTOMATIZUAR TË FJALËVE KYÇE NGA TEKSTE NË GJUHËN SHQIPE

Zhvillimet intensive të teknologjisë në dekadat e fundit, kanë bërë që çdokush të ketë në dispozicion sasi gjithmonë e më të mëdha të dhënash që vijnë nga shumë burime të ndryshme. Megjithatë, nëse volumi i të dhënave është rritur eskponencialisht ashtu si edhe larmia e metodave të aksesimit të të dhënave, një tjetër problem ka lindur dhe vijon të bëhet më i vështirë që është transformimi apo interpretimi i të dhënave për të nxjerrë informacion të dobishëm. Disponueshmëria e të dhënave rreth një entiteti jo domosdoshmërisht implikon si rrjedhojë informacion të dobishëm rreth atij entiteti. Përkundrazi, shpesh, volume të mëdha të dhënash sjellin konfuzion në gjetjen e informacionit të duhur.

Në këtë kuadër, problemi bëhet akoma më i vështirë kur të dhënat që duhet të interpretohen vijnë në forma të pastrukturuara siç është p.sh teksti. Shumica e informacionit në qeveri, industri, biznes, apo institucione të tjera, është i ruajtur në formën e Bazave të të Dhënave Tekst. Këto lloje të dhënash po marrin përmasa gjithmonë e më të mëdha si pasojë e rritjes së volumit të tyre në format elektronik kryesisht për shkak të World Wide Web që mund të shihet si një bazë dinamike dhe e ndërlidhur të dhënash tekst. Si rrjedhojë e shpërthimit eksponencial të rritjes së të dhënave tekst në WWW, përpjekje intensive janë bërë nga komuniteti shkencor për t'u përballur me këtë problem. Për arsye të mungesës së strukturës, të dhënat tekst zakonisht menaxhohen nëpërmjet motorrëve të kërkimit, të cilët fokusohen në kërkimin sipas një query me fjalë kyçe. Studimet kryesore në këtë drejtim janë përqendruar në rritjen e efektivitetit dhe të efiçensës së këtyre sistemeve, ashtu si edhe në probleme të tjera si: klasifikim teksti, kategorizim teksti, përmbledhje, etj.

Në këtë kontekst, zhvillimi i metodave që ndihmojnë në filtrimin apo gjetjen e informacionit, merr një rëndësi të veçantë. Një nga keto metoda është edhe ajo e gjetjes së fjalëve kyçe të cilat mund të ndihmojnë në identifikimin e atyre dokumentave që përputhen me kërkesat për informacion të përdoruesit. Kërkime intensive janë bërë për këto metoda, por aktualisht për gjuhën Shqipe kërkimi shkencor në këtë drejtim ka qenë modest për shkak të vështirësive që paraqet kjo gjuhë me specifikat e saj, por edhe për shkak të interesit të pakët që paraqet kjo gjuhë për motorrët e mëdhenj të kërkimit në rrjet të cilet kanë fokusuar kërkimin shkencor për gjuhët më të përhapura.

Në këtë dizertacion trajtohet problemi i nxjerrjes së fjalëve kyçe nga tekste në gjuhën Shqipe. Punimi propozon një algoritëm të ri i cili është projektuar duke respektuar rregullat strukturore të gjuhës Shqipe, por që mund të përshtatet edhe për gjuhë të tjera. Ky algoritëm mbështetet në krijimin e një modeli nëpërmjet klasifikatorit Naïve Bayes i cili përdoret për të rritur efiçensën e modelit statistikor të nxjerrjes së fjalëve kyçe që bazohet në teknikën TFIDF. Klasifikatori jep kontribut në identifikimin e termave kyçe duke vlerësuar nëqoftëse një term i zgjedhur nga modeli statistikor është ose jo determinues për klasifikatorin. Algoritmi është aplikuar mbi dokumenta tekst të shkruara në gjuhën Shqipe dhe të paraproçesuara nga ana linguistike.

Eksperimentet e zhvilluara tregojnë që algoritmi është efikas dhe nxjerr nga dokumentat, fjalë kyçe domethënëse për dokumentin dhe që e dallojnë atë nga dokumenta të tjerë në koleksione të gjera dokumentash të ngjashëm. Vlerësimi i performancës bazohet në një faktor F-measure i cili përdor Precision dhe Recall. Vlera e F-measure për algoritmin tonë rezultoi 36%, rezultat ky që konsiderohet relativisht i mirë duke patur parasysh që edhe midis ekspertëve të ndryshëm që zgjedhin manualisht fjalë kyçe nga i njëjti dokument, përputhshmëria është rreth 40%. Në dizertacion diskutohen gjithashtu edhe përmirësime të mundshme që mund t'i bëhen algoritmit si pjesë e punimeve të ardhshme.

Fjalë kyçe: *Nxjerrje e Automatizuar Fjalësh Kyçe, Nxjerrje Informacioni të Kërkuar, Machine Learning, Text Mining, Paraproçesim Linguistik, TFIDF, Naïve Bayes, Proçesim i Gjuhëve Natyrale.*

Falenderime

Së pari, dua të shpreh mirënjohjen për udhëheqësin shkencor, Prof. Dr. Aleksandër Xhuvani i cili më dha mundësinë të punoj me të në një fushë kaq frymëzuese dhe interesante duke më përkrahur në idetë dhe inisiativat e mia. Mbështetja e vazhdueshme dhe diskutimet e pakursyera me të më kanë motivuar në projektimin dhe implementimin me sukses të këtij algoritmi.

Falenderoj përzemërsisht dekanen e Fakultetit të Teknologjisë së Informacionit, Prof. Dr. Rozeta Miho, për mbështetjen e vazhdueshme të saj dhe ndjekjen me përkushtim të mbarëvajtjes së ciklit të studimeve të doktoratës.

Gjithashtu dua të falenderoj të gjithë kolegët e Departamentit të Inxhinierisë Informatike, dhe në veçanti përgjegjësen e departamentit, Prof. Asoc. Elinda Meçe, për pozitivitetin dhe përkrahjen e vazhdueshme.

Në mënyrë të veçantë falenderoj bashkëshortin tim Marenglen që më ka mësuar durimin dhe pasionin për kërkimin shkencor.

Evis Trandafili
Tiranë, 2014

TABELA E PËRMBAJTJES

ABSTRAKT ..... **III**

FALENDERIME ..... **V**

LISTA E FIGURAVE ..... **IX**

LISTA E TABELAVE ..... **X**

FJALOR TERMINOLOGJIK ..... **XI**

1. HYRJE **1**

1.1. Bazat Shkencore dhe Motivimi **1**

1.2. Objektivat e Studimit **2**

1.3. Struktura e Dizertacionit **3**

2. TEXT MINING: GJETJA E INFORMACIONIT TË STRUKTURUAR DHE MARRJA E INFORMACIONIT TË KËRKUAR **5**

2.1. Text Mining **5**

2.1.1. Hyrje në Text Mining 5

2.1.2. Arkitektura e Sistemeve Text Mining 10

2.1.3. Algoritmat dhe Metodat 12

2.2. Gjetja e Informacionit të Kërkuar (Information Retrieval) **13**

2.2.1. Bazat Matematike 15

2.3. Marrja e Informacionit të Strukturuar (Information Extraction) **25**

2.3.1. Njohja e Entiteteve të Emëruara (angl.Named Entity Recognition) 27

2.3.2. Nxjerrja e Lidhjeve (angl.Relation Extraction) 29

2.3.3. Nxjerrja e Pasupervizuar e Informacionit (angl. Unsupervised Information Extraction) 31

3. GJETJA E FJALËVE KYÇE **34**

3.1. Hyrje **34**

3.2. Metodat Kryesore **35**

3.3. Përkufizimi dhe Identifikimi i Termave **38**

3.4. Karakteristikat e Termave **38**

3.4.1. Karakteristikat Linguistike të Termave 39

3.4.2. Karakteristikat Statistikore të Termave 39
3.4.3. Karakteristikat e Shpërndarjes së Termave 40

3.5. Metodat për Gjetjen e Fjalëve Kyçe 41

3.5.1. Nxjerrja e Fjalëve Kyçe nga Dokumenta Monogjuhësore 41
3.5.2. Nxjerrja e Fjalëve Kyçe nga Dokumenta Multigjuhësore 47

3.6. Sistemet për Gjetjen e Fjalëve Kyçe 47

3.6.1. TERMINO 49
3.6.2. LEXTER 50
3.6.3. Xtract 51
3.6.4. ANA 51
3.6.5. KeyGraph 52
3.6.6. GenEx 53
3.6.7. KEA 53
3.6.8. GlossEx 54
3.6.9. TextRank 56
3.6.10. Hulth 57
3.6.11. CollabRank 57
3.6.12. RAKE 58

3.7. Metodat e Vlerësimit të Sistemeve të Gjetjes së Fjalëve Kyçe 60

3.7.1. Kriteret për Përputhshmërinë e Shprehjeve 60
3.7.2. Përcaktimi i Rrënjëve të Fjalëve (angl. stemming) 61
3.7.3. Precision dhe Recall 61

4. GRAMATIKA FORMALE E SHQIPES 64

4.1. Morfologjia e Gjuhës Shqipe 64

4.1.1. Emrat 64
4.1.2. Mbiemri 65
4.1.3. Përemri 66
4.1.4. Foljet 67
4.1.5. Ndajfoljet 69
4.1.6. Numëruesi 69
4.1.7. Lidhëza 70
4.1.8. Parafjala 70
4.1.9. Pasthirrma 70
4.1.10. Onomatopetë 70
4.1.11. Pjesëza 71

4.2. Sintaksa e Gjuhës Shqipe 71

4.2.1. Tipet e Fjalive 71
4.2.2. Format e Fjalive 72
4.2.3. Pjesët Përbërëse të Fjalisë 72

4.2.4. Formimi i Fjalëve në Gjuhën Shqipe 75

5. ANALIZA DHE PROJEKTIMI I ALGORITMIT 79

5.1. Komponenti Linguistik 79

5.1.1. Eleminimi i Lidhësave/Konektorët e Tekstit 79
5.1.2. Gjetja e Rrënjëve të Fjalëve të Thjeshta 80
5.1.3. Probleme në Gjetjen e Rrënjës së Fjalëve në Gjuhën Shqipe 83
5.1.4. Gjetja e Rrënjëve të Fjalëve të Përbëra 85

5.2. Komponenti Statistikor 87

5.3. Komponenti i Machine Learning 87

5.4. Projektimi i Algoritmit 90

5.5. Detaje të Implementimit 92

6. VLERËSIMI EKSPERIMENTAL 94

6.1. Korpusi i Dokumentave 94

6.2. Metoda e Vlerësimit 94

6.3. Rezultatet 96

6.3.1. Test i Tolerancës ndaj Fjalëve Kyçe të Vendosura nga Autori 97
6.3.2. Testi i Cilësisë së Fjalëve Kyçe Automatike 99
6.3.3. Testi i Ndikimit të Madhësisë së Korpusit Trajnues 100
6.3.4. Testi i Vlerësimit të Shkallës së Përmirësimit ndaj Modelit Statistikor 101
6.3.5. Testi i Fjalëve Kyçe me Peshë Statistikore të Ulët 102
6.3.6. Testi i Vlerësimit Manual të Fjalëve Kyçe 103

7. KONKLUZIONET 105

7.1. Kontributi i Këtij Disertacioni 106

7.2. Puna në të Ardhmen 106

ANEKS A 108

REFERENCA 111

Lista e Figurave

Figura 2.1. Modeli Input-Output i Text Mining

Figura 2.2. Arkitektura funksionale e sistemeve Text Minig në High-Level

Figura 2.3. Kategorizimi i modeleve të sistemeve të GjIK

Figura 2.4. Dy pjesët e indeksit të invertuar. Fjalori ruhet në kujtesë me pointera në secilën listë që ruhet në disk.

Figura 2.5. Ndërtimi i indeksit me anë të renditjes dhe grupimit, bazuar në Përmbledhjen e Shakespeare.

Figura 2.6. Paraqitja e ngjashmërisë së dokumentave nëpërmjet kosinusit të këndit midis tyre.

Figura 2.7. Arkitektura e Sistemeve të Marrjes së Informacionit të Strukturuar

Figura 3.1. Titujt dhe fjalët kyçe të autorit dhe ato të zgjedhura automatikisht për tre artikuj

Figura 3.2. Modulet e proçesit të nxjerrjes së termave

Figura 3.3. Paraqitja e të dhënave nga KeyGraph

Figura 3.4. Proçesi i trajnimit dhe nxjerrjes së Shprehjeve Kyçe

Figura 3.5. Komponentët e Glossex

Figura 3.6. Fjalë kyçe kanditate pas parsimit

Figura 3.7. Matrica e fqinjësisë për fjalët kyçe kandidate

Figura 3.8. Peshat e fjalëve llogaritur nga matrica e fqinjësisë

Figura 3.9. Vlerësimi i algoritmit të nxjerrjes së fjalëve kyçe

Figura 3.10. Matrica e klasifikimit të fjalëve kyçe

Figura 5.1. Metodat e gjetjes së rrënjëve të fjalëve

Figura 5.2. Shembull i një rregulli për nxjerrjen e rrënjës së fjalës

Figura 5.3. Algoritmi i gjetjes së rrënjës së fjalës

Figura 5.4. Grupet e fjalëve të përbëra

Figura 5.5. Algoritmi Naive Bayes Multinomial

Figura 5.6. Trainimi i klasifikatorit për të gjeneruar modelin

Figura 5.7. Skema e algoritmit të gjetjes së fjalëve kyçe të një dokumenti

Figura 5.8. Baza e të dhënave MySQL për dokumentat e organizuar sipas peshave statistikore

Figura 6.1. Dokumentat e trajnimit të organizuar në klasa

Figura 6.2. Një pjesë nga dokumentit që i është dhënë secilit prej ekspertëve

Figura 6.3. Paraqitja grafike e marëdhënies midis madhësisë së korpusit dhe F-measure

Lista e Tabelave

Tabela 6.1. Përmbledhje e të dhënave të korpusit.

Tabela 6.2. Fjalët kyçe të nxjerra manualisht nga dy ekspertë të ndryshëm për të njëjtin dokument.

Tabela 6.3. Vlerësimi i koeficientit K për konkordancën midis autorëve

Tabela 6.4. Precision dhe recall për sistemin automatik të nxjerrjes së fjalëve kyçe

Tabela 6.5. Ndryshimi i F-measure sipas madhësisë së korpusit

Tabela 6.6. Precision dhe recall sipas komponentit statistikor

Tabela 6.7. Precision dhe recall i gjetjes së fjalëve kyçe me metoda statistikore për disa gjuhë të huaja.

Tabela 6.8. Përputhshmëria midis Algoritmit të kombinuar me Machine Learning dhe atij statistikor

Tabela 6.9. Rezultatet e vlerësimit manual të fjalëve kyçe nga një ekspert

Tabela A.1. Tebala e eleminimit të zanoreve

Tabela A.2. Tebala e lidhësave të gjuhës Shqipe

Tabela A.3. Tebala e rregullave për fshirjen e parashtesave

Tabela A.4. Tebala e rregullave për fshirjen e prapashtesave

Fjalor Terminologjik

Analiza sintaksore (angl. syntactic parsing) – Informacioni lidhur me strukturën pemë apo të marrëdhënieve/relacioneve gramatikore të përbërësve.

Data Mining – fushë studimi e cila merret me analizën dhe kërkimin në baza të dhënash ekzistuese për të nxjerrë informacion të ri.

Domain – nje hapesirë ose bashkësi me një karakteristikë të përbashkët.

Etiketimi i pjesëve të ligjëratës (angl. partof-speech tagging)– proçesi i etiketimit të fjalëve në një tekst në bazë të pjesëve koresponduese të ligjëratës. p.sh. emër, mbiemër, folje, lidhëze etj. Është tipi themelor i etiketimit të korpusit linguistik, i cili shërben si pikënisje për llojet e tjera të analizave.

Fjalë kyçe – një fjalë ose shprehje që përfaqëson përmbajtjen e një dokumenti.

Framework – platformë universale software për zhvillimin e aplikacioneve software, produkteve apo shërbimeve.

Gjetje e Informacionit të Kërkuar (GJIK) – angl. Information Retrieval

Korpus – Bashkësi e madhe dokumentash për qëllime trajnimi të modeleve.

Lematizim (angl. lematization) – Reduktimi i fjalëve në korpus në leksemat e tyre përkatëse apo në fjalën përfaqësuese që jepet në fjalor; p.sh. fjalët *jam, je, është, jemi, jeni, janë; isha, ishe, ishte, ishim, ishit, ishin; qeshë, qe, qemë, qetë, qenë; qenkam, qenke, qenka, qenkemi, qenkeni, qenkan; qenkësha, qenkëshe, qenkësh, qenkëshim, qenkëshit, qenkëshin; qofsha, qofsh, qoftë, qofshim, qofshi, qofshin; ji, jini; (të) jem, (të) jesh, (të) jetë, (të) jemi, (të) jeni, (të) janë* – që të gjitha janë trajta të foljes "me qenë", që në fjalorë jepet me trajtën e saj të vetës së parë njëjës të dëftores të së tashmes: jam – dhe kjo është lema e të gjitha këtyre fjalëve.

Marrje e Informacionit të Strukturuar (MIS) – angl. Information Extraction

Metadata – formë specifike të dhënash rreth një dokumenti, si autori, titulli apo data e publikimit.

Ndarja e tesktit në blloqe (angl. Text Chunking) - Fazë e paraproçesimit të tekstit për Proçesim të Gjuhëve Natyrale.

Nxjerrja e Lidhjeve (angl. Relation Extraction) – nxjerra e marëdhënieve semantike midis dy entiteteve nga një teskt i shkruar në gjuhë natyrale.

Query – kërkesë për informacion nga një bazë të dhënash.

Shënimi semantik (angl. semantic annotation) **–** Fjalët, togjet e fjalëve (sintagmat) ose fjalitë kanë etiketa që shënojnë fushën semantike (kuptimore), metaforën etj., me të cilën zgjidhet problemi i homonimisë, si në rastin: *dhe (korja e rruzullit tokësor; dhe i zi) – emër, dhe (jap; ti dhe) – folje, dhe (unë dhe ti) – lidhëse, dhe (dhe atje nuk u mësua) – pjesëz*; dhe problemi i polisemisë (shumëkuptimësisë), si në rastin: *rrënjë (rrënjë e një bime; rrënjë e flokut; prejardhje; rrënjë katrore – në matematikë, rrënjë kimike – në kimi etj.).*

Template – një model ose format i paracaktuar i cili përdoret si standard në krijimin e modeleve të reja.

Text Mining – nën-fushë e Data Mining e cila analizon dhe kërkon në baza të dhënash tekst për të nxjerrë informacion të dobishëm.

Token – njësi leksikore p.sh. *ai, atij; djalë, djem, djemve; jam, janë, isha, qesh; kam, kemi* etj.

1. Hyrje

Në këtë kapitull përshkruhen në terma të përgjithshëm disa nga elementët kryesorë të këtij dizertacioni. Bazat shkencore mbi të cilat bazohet ky studim, përmenden shkurt në terma të thjeshtuar. Qëllimi paraprak është dhënia e një tabloje të përgjithshme të dekompozimit teorik të problemit deri në nivelet me të cilat lidhet puna konkrete e zhvilluar.

1.1. Bazat Shkencore dhe Motivimi

Vitet e fundit, zhvillimet teknologjike në drejtimet hardware dhe software, kanë bërë që çdokush të ketë në dispozicion sasi gjithmonë e më të mëdha informacioni të shumëllojshëm. Studimet e para në Data Mining janë fokusuar në të dhëna të strukturuara, që zakonisht menaxhohen nëpërmjet Sistemeve të Bazave të të Dhënave. Në realitet, një sasi e konsiderueshme informacioni ruhet në Baza të Dhënash Tekst, që konsistojnë në bashkësi dokumentash të marra nga burime të ndryshme si artikuj gazetash, publikime shkencore, libra, biblioteka dixhitale, mesazhe e-mail, faqe web, etj, (Witten and Frank, 2012). Shumica e informacionit në qeveri, industri, biznes, apo institucione të tjera, është i ruajtur në formën e Bazave të të Dhënave Tekst. Këto lloje të dhënash po marrin përmasa gjithmonë e më të mëdha si pasojë e rritjes së informacionit në format elektronik. World Wide Web mund të shihet si një bazë dinamike të dhënash tekst, e ndërlidhur. Për arsye të mungesës së strukturës, të dhënat tekst zakonisht menaxhohen nëpërmjet motorrëve të kërkimit, të cilët fokusohen në kërkimin sipas një query me fjalë kyçe. Studimet kryesore në këtë drejtim janë përqendruar në rritjen e efektivitetit dhe të efiçensës së këtyre sistemeve, ashtu si edhe në probleme të tjera si: klasifikim teksti, kategorizim teksti, përmbledhje, etj, (Aggarwal and Zhai, 2012).

Në këtë kontekst gjetja e informacionit të nevojshëm nga këto burime, merr një rëndësi të veçantë. Përpara se të thellohemi në këtë fushë po bëjmë disa përcaktime të nevojshme: me shprehjen "Gjetje Informacioni të Kërkuar", do t'i referohemi termit në anglisht "Information Retrieval"; me shprehjen "Marrje Informacioni të Strukturuar", do t'i referohemi termit në anglisht "Information Extraction".

Në vitet 90, shumica e njerëzve nuk i preferonin sistemet e Gjetjes së Informacionit të Kërkuar (GjIK). Në dekadën e fundit, optimizimi i efiçensës së GjIK, ka çuar në nivele cilësore të motorrëve të kërkimit në web, dhe në shumicën e rasteve përdoruesit janë të kënaqur me rezultatet e kërkimit. Kërkimi në web ka krijuar një standard, madje konsiderohet si burimi i preferuar i informacionit. GjIK ka për qëllim të gjejë informacion nga bashkësi të mëdha dokumentash në web apo në intranet-et e kompanive, sipas query – ve të bazuara në fjalë kyçe. Sistemet e GjIK ekzistuese kthejnë një listë dokumentash kandidatë të rezultatit të kërkimit. Secili prej tyre duhet

të filtrohet më tej nga përdoruesi për të evidentuar shkallën e përputhshmërisë me kërkimin e tij. Në këtë mënyrë lind nevoja e përdorimit të mjeteve që zvogëlojnë sasinë e tekstit që duhet të filtrojmë/lexojmë për të marrë informacionin që na duhet. Marrja e Informacionit të Strukturuar (MIS), është një nga teknologjitë bazë në optimizimin e bashkësisë së rezultateve të kërkimit. Identifikimi i entiteteve dhe lidhjeve semantike midis tyre, është thelbësor në marrjen e një informacioni të koncentruar. Teknologjitë e MIS nxjerrin në pah faktin që në dokumentin e gjetur duhet të përputhen jo vetëm fjalët e query-t, por edhe klasifikimet semantike dhe lidhjet midis entiteteve (Moens, 2006).

Në MIS, një rol thelbësor luan përcaktimi i fjalëve kyçe, bashkësia e të cilave na jep një paraqitje kompakte të përmbajtjes së një dokumenti. Përdorimi i tyre përmirëson funksionalitet e sistemeve të GjIK. Pavarësisht avantazheve të dukshme në kërkim, shumica e dokumentave nuk kanë fjalë kyçe të paracaktuara. Shumë zgjidhje bazohen në vendosjen manuale të fjalëve kyçe nga profesionistë të fushave ose nga vetë autorët e dokumentave. Rrjedhimisht, kërkimi shkencor është fokusuar në gjetjen e metodave automatike për nxjerrjen e fjalëve kyçe nga dokumentat dhe më pas përdorimin e tyre si lista të sugjeruara për indeksuesit profesionistë ose për gjenerimin e elementëve përmbledhës për ato dokumenta të cilët do të ishin të paaksesueshëm nga sistemet e GjIK, (Berry and Kogan, 2010).

Komuniteti shkencor vazhdon të bëjë përpjekje të konsiderueshme për zvogëlimin e hendekut në komunikimin me kompjuterat në gjuhë natyrale. Duke patur parasysh faktin që mjetet dhe zgjidhjet me të cilat punohet janë të varura nga sintaksa dhe struktura gjuhësore, ky hendek zgjerohet edhe më shumë nëqoftëse i referohemi gjuhës Shqipe. Motivacioni kryesor që na shtyu të punojmë në këtë fushë është plagjiatura e theksuar dhe mungesa e respektit për te drejtat e autorit në vendin tonë. Për t'u admiruar janë përpjekjet që po bëhen në Shqipëri për të kaluar online të gjitha materialet publike, por kjo inisiativë duhet mbështetur me baza shkencore në mënyrë që këto dokumenta të bëhen të aksesueshëm nga algoritmat që përdorin motorrët e kërkimit por, edhe të mbrohen nga shkeljet e të drejtave të autorit. Me anë të kësaj teze do të përpiqemi të japim kontributin tonë në kontrollin e dokumentave në gjuhën Shqipe. Në këtë punim do të trajtojmë problemet që lidhen me MIS nga dokumenta të pastrukturuara tekst duke u përqëndruar në përcaktimin e fjalëve kyçe për dokumentat tekst në gjuhën Shqipe. Suksesi në këtë përpjekje do t'i hapë rrugën zhvillimeve të mëtejshme në Text Mining edhe për gjuhën Shqipe.

1.2. Objektivat e Studimit

Qëllimi i këtij studimi është ndërtimi i një metode efiçente për nxjerrjen e fjalëve kyçe nga dokumenta individuale në gjuhën shqipe. Objektivat kryesorë të studimit janë:

- Vlerësim të situatës aktuale dhe zhvillimeve në teknikat e reja të nxjerrjes së fjalëve kyçe nga bashkësi dinamike dokumentash tekst. Ky studim do të bazohet në metodat e

krijuara për gjuhën angleze, meqënëse këtu është përqëndruar shumica e kërkimit shkencor dhe avancimet kanë qënë më të mëdha.

- Studim të literaturës shkencore për strukturën linguistike dhe morfologjike të gjuhës Shqipe.
- Propozimi i një algoritmi për nxjerrjen e fjalëve kyçe të një dokumeti të shkruar në gjuhën Shqipe dhe ndërtimi i një moduli në Java, i cili të funksionojë si aplikacion stand-alone, dhe të aplikohet mbi bashkësi dinamike dokumentash tekst në gjuhën Shqipe. Ky do të jetë fokusi kryesor i studimit.
- Testim i zgjidhjes së propozuar për verifikimin e saktësisë së rezultatit. Kjo do të realizohet duke vlerësuar përputhshmërinë së rezultatit të marrë në mënyrë të automatizuar, me bashkësitë e fjalëve kyçe të propozuara nga grupe të pavarura profesionistësh të fushës.
- Vlerësimi i performancës së arritur nga zgjidhja e propozuar për grupe të ndryshme tekstesh në gjuhën shqipe, në raport me performancën e arritur nga zgjidhjet për bashkësi dokumentash në gjuhën angleze me anë të metodave të tjera. Të gjitha krahasimet eksperimentale do të bëhen duke përdorur disa nga metodat standarde që përdoren zakonisht në fushën e GjIK për vlerësimin e rezultateve.
- Vlerësim i përgjithshëm i rezultateve të arritura dhe diskutim kritik i çështjeve dhe problematikave që mbeten për t'u zgjidhur në të ardhmen.

1.3. Struktura e Dizertacionit

Struktura e organizimit të këtij disertacioni është si më poshtë:

Kapitulli i parë përshkruan në terma të përgjithshëm disa nga elementët kryesorë të dizertacionit. Bazat shkencore mbi të cilat bazohet ky studim, përmenden shkurt në terma të thjeshtuar. Këtu jepet një tablo e përgjithshme e dekompozimit teorik të problemit deri në nivelet me të cilat lidhet puna konkrete e zhvilluar.

Kapitulli i dytë përmban një trajtim të thelluar të tre fushave shkencore të cilat ndërthuren në punën e këtij dizertacioni që janë: Text Mining, Information Retrieval dhe Information Extraction. Këtu prezantohen konceptet dhe modelet bazë të përpunimit të tekstit në gjuhë natyrale dhe paraqiten disa nga problemet tipike që zgjidhen nga këto modele.

Kapitulli i tretë trajton gjerësisht problemin klasik të nxjerrjes së fjalëve kyçe duke u fokusuar në karakteristikat linguistike, statistikore dhe të pozicionimit. Gjjithashtu, trajtohen metodat kryesore aktuale për nxjerrjen e fjalëve kyçe duke përshkruar disa nga sistemet kryesore dhe mënyrat e vlerësimit të tyre.

Kapitulli i katërt trajton karakteristikat e gjuhës Shqipe të nevojshme për analizën linguistike në kuadër të paraproçesimit të tekstit. Në këtë kapitull trajtohen komponentët morfologjikë dhe sintaksorë të nevojshëm për projektimin dhe zhvillimin e komponentëve të algoritmit.

Kapitulli i pestë paraqet në mënyrë të detajuar komponentët përbërës të algoritmit të gjetjes së fjalëve kyçe të një dokumenti. Ky algoritëm prezantohet si nënproblem i atij të kërkimit të informacionit, të aplikuar në korpuse në gjuhën Shqipe të kategorizuara në klasa. Algoritmi është i organizuar në disa module logjike. Moduli i parë lidhet me elementët linguistikë në gjuhën Shqipe. Komponenti linguistik është një pjesë e rëndësishme e algoritmit të propozuar për gjetjen e fjalëve kyçe i cili i takon fazës së paraproçesimit të dokumentave. Moduli i dytë lidhet me elementet statistikorë të nxjerrë nga trajnimi i korpusit. Komponenti statistikor përdor një metodë statistikore të peshave të fjalëve që përshkruajnë dokumentin. Funksioni i peshave që është përdorur e vë fokusin në metodën TFIDF. Moduli i tretë shfrytëzon kapacitetin e metodave të Machine Learning për të optimizuar përzgjedhjen e fjalëve kyçe. Ndër modelet e Machine Learning të aplikuara për të dhëna në formë teksti është përdorur modeli Naive Bayes. Zgjidhja algoritmike është specifike për gjuhën Shqipe dhe është implementuar në mënyrë të tillë që t'i përshtatet rregullave strukturore të kësaj gjuhe.

Kapitulli i gjashtë paraqet një vlerësim të gjerë eksperimental të algoritmit të propozuar. Eksperimenet janë zhvilluar mbi një korpus (koleksion) dokumentash në gjuhën Shqipe të mbledhura nga fusha heterogjene me anë të të cilave është testuar efikasiteti i algoritmit në nxjerrjen e fjalëve kyçe nga dokumenta të fushave të ndryshme. Pas nxjerrjes së fjalëve kyçe janë zhvilluar disa teste duke marrë për bazë fjalët kyçe të vendosura nga ekspertët. Vlerësimi i performancës është bazuar në një faktor F-measure i cili përdor Precision dhe Recall.

Kapitulli i shtatë paraqet një analizë kritike të punës së zhvilluar duke nxjerrë në pah pikat e forta dhe të dobëta të punimit. Gjithashtu, përmban edhe një sërë drejtimesh për punimet e ardhshme në përmirësimin e metodës së paraqitur.

2. Text Mining: Gjetja e Informacionit të Strukturuar dhe Marrja e Informacionit të Kërkuar

Në këtë kapitull do të trajtohen në mënyrë të plotë të gjitha konceptet shkencore që përbëjnë bazën e studimit tonë.

2.1. Text Mining

Text Mining është një fushë e re kërkimore, e cila e vë theksin në zgjidhjen e problemeve që lindin nga volume shumë të mëdha të dhënash, duke shfrytëzuar teknika nga Data Mining, Machine Learning, Natural Language Processing, Information Retrieval dhe Knowledge Management. Text Mining përfshin proçeset e parapërgatitjes së bashkësive të dokumentave (text categorization, information extraction, term extraction), ruajtjen e të dhënave të ndërmjetme, teknikat për analizën e këtyre të dhënave të ndërmjetme (si distribution analysis, clustering, trend analysis, dhe association rules) dhe në fund paraqitjen e rezultateve, (Feldman and Sanger, 2007).

2.1.1. Hyrje në Text Mining

Text Mining mund të përkufizohet si një proçes që shfrytëzon intensivisht njohuri, në të cilin një përdorues ndërvepron në kohë me një grup dokumentash duke përdorur disa mjete analizuese. Për analogji me Data Mining, Text Mining kërkon të nxjerrë informacion nga burime të dhënash, duke identifikuar dhe eksploruar modele me interes apo të dobishme. Në rastin e Text Mining, burimet e të dhënave janë bashkësi dokumentash tekst, nga të cilët modelet e dobishme nuk gjenden nga rekorde databazash klasike, por nga të dhëna tekst të pastrukturuara. Sigurisht, drejtimet bazë të Text Mining kanë rrjedhur nga kërkimet teorike bazë në Data Mining. Për këtë arsye, sistemet e Text Mining dhe ato të Data Mining kanë shumë ngjashmëri nëqoftëse i shohim nga pikëpamja e arkitekturës së nivelit të lartë. Për shembull, të dy tipet e sistemeve bazohen në: paraproçesim, algoritma të evidentimit të modelit dhe elementë të paraqitjes për të rritur lexueshmërinë e rezultateve. Për më tepër, shumica e tipeve specifike të modeleve që përdoren në veprimet bazë të zbulimit të njohurive në Text Mining, janë zhvilluar dhe verifikuar gjatë kërkimeve në Data Mining. Meqënëse në Data Mining supozohet që të dhënat janë të ruajtura në format të strukturuar, shumica e punës paraproçesuese fokusohet në dy detyra kritike: normalizimi i të dhënave dhe krijimi i bashkimeve të tabelave. Në të kundërt, në sistemet e Text Mining, veprimet e paraproçesimit bazohen në identifikimin dhe nxjerrjen e karakteristikave përfaqësuese të dokumentave të shkruara në gjuhë natyrale. Këto veprime paraproçesimi, transformojnë të dhënat e pastrukturuara, të ruajtura në bashkësi dokumentash, në një format të ndërmjetëm të strukturuar, i cili nuk është i nevojshëm për sistemet e Data Mining. Për më tepër, për shkak të vendit që zënë tekstet në gjuhë natyrale në Text Mining, avancimet në këtë fushë

kanë sjellë zhvillime edhe në disiplina të tjera të shkencave kompjuterike që bazohen në gjuhë natyrale. Vlen të përmendim këtu faktin që Text Mining zhvillon teknika dhe metodologji në drejtim të fushave të: Gjetje Informacioni të Kërkuar (GJIK), Marrje Informacioni të Strukturuar (MIS) dhe linguistikës komputacionale të bazuar në korpuse (databaza per gjenerim automatik modelesh) (Feldman and Sanger, 2007).

2.1.1.1. Dokumenti në bazë të Text Mining

Elementi kyç në Text Mining është fokusi në bashkësinë e dokumentave. Një bashkësi dokumentash mund të jetë çdo koleksion i çfarëdoshëm dokumentash tekst. Zgjidhjet që jepen në Text Mining i adresohen gjetjes së modeleve në koleksione shumë të mëdha dokumentash, që variojnë në numër nga disa mijëra në dhjetra milionë. Koleksionet e dokumentave mund të jenë statike, (te gjithë dokumentat mbeten gjithmonë të pandryshuar), ose dinamike (bashkësia karakterizohet nga hyrje dokumentash të rinj ose të ndryshuar). Duke qënë se kemi të bëjmë me koleksione shumë të mëdha dinamike, sfida kryesore mbetet performanca e çdo komponenti të sistemit Text Mining (Feldman and Sanger, 2007).

Pavarësisht se dokumentat tekst, në përgjithësi, konsiderohen si të dhëna pa strukturë, nga disa këndvështrime ato mund të shihen si objekte të strukturuara. Në këndvështrimin linguistik, edhe dokumenti tekst më i thjeshtë, mbart një strukturë komplekse semantike dhe strukturore, pavarësisht se kjo strukturë konsiderohet implicite dhe në një farë mënyre është e fshehur në përmbajtje. Nga ana tjetër, elementë si shenjat e pikësimit, shkronjat e mëdha e të vogla, numrat, karakteret speciale – sidomos të kombinuara me elementë të paraqitjes si: hapësirat, nënvizimet, tabelat, kolonat, etj – mund të shërbejnë si një lloj gjuhe etiketimi (soft markup language), duke mbartur informacion të rëndësishëm për identifikimin e nënkomponentëve të dokumentit si paragrafët, titujt, data e publikimit, emrat e autorëve, rekordet e tabelave, headerat dhe footnotes. Gjithashtu, edhe sekuenca e fjalëve është një element i rëndësishëm strukturor për dokumentin. Nga ana tjetër, dokumenta pa strukturë, janë edhe ata në formatin HTML, të cilët që në gjenezë të tyre posedojnë metadata të integruara në formën e tageve markup të formalizuara. Dokumentat që kanë relativisht pak strukturë – si për shembull, artikuj shkencorë, raporte biznesi, memorandume ligjore, lajme – cilësohen si dokumenta *pa format* ose *me strukturë të dobët.* Ndërsa, dokumentat me elementë të zgjeruar dhe konsistentë të formatit, në të cilët fushat e tipeve metadata mund të aksesohen lehtësisht – si e-mail, faqet Web HTML, skedarët PDF dhe skedarët nga editorë tekst me model të përcaktuar ose rregulla style-sheet – konsiderohen si dokumenta *pjesërisht të strukturuar.*

Veprimet e paraproçesimit në Text Mining, tentojnë të theksojnë elementët që ndodhen në një dokument të shkruar në gjuhë natyrale, në mënyrë që të lehtësohet transformimi i tij nga paraqitja me strukturë implicite dhe jo të rregullt në një paraqitje me strukturë eksplicite. Duke marrë parasysh numrin e madh të fjalëve, shprehjeve, fjalive dhe elementëve të tjerë që mund të ketë

edhe një dokument mjaft i thjeshtë – pa përmendur, numrin e kuptimeve të ndryshme që mund të ketë secili prej këtyre elementëve në kontekste të ndryshme, – detyra thelbësore për shumicën e sistemeve Text Mining është identifikimi i një nënbashkësie të thjeshtuar të karakteristikave/elementëve të dokumentit, të cilat mund të përdoren për ta përfaqësuar atë dokument sa më mirë. Kjo nënbashkësi karakteristikash quhet model perfaqësues i dokumentit. Pavarësisht përpjekjeve për të ndërtuar modele përfaqësuese efiçente, çdo dokument tekst në një koleksion, përbëhet nga një numër i madh karakteristikash të cilat ndikojnë nga dizenjimi deri te performanca e sistemeve Text Mining. Problemet që lidhen me dimensionin e lartë të karakteristikave, (madhësia dhe shkalla e kombinimeve të vlerave të secilës karakteristikë për të dhënat), zakonisht janë më të mëdha në sistemet Text Mining se sa në sistemet klasike Data Mining. Edhe për dokumentin tekst më modest, numri i karakteristikave në nivel fjale i nevojshëm për të përfaqësuar dokumentin mund të jetë jashtëzakonisht i lartë. Për shembull, në një koleksion me 15 000 dokumenta të marra nga rubrikat e lajmeve, mund të identifikohen më shumë se 25 000 rrënjë fjalësh. Edhe në rastin kur punohet me tipe karakteristikash më të optimizuara, përsëri mund të identifikohen dhjetëra mijëra nivele konceptuale karakteristikash për një fushë të vetme aplikative. Ndërkohë që numri i atributeve të analizuara në një proçes Data Mining në nje databazë relacionale, është ndjeshëm më i vogël. Në mënyrë indirekte, hapësira e dimensioneve kontribuon në kriteret që diferencojnë sistemet Text Mining nga ato Data Mining duke përqëndruar përpjekjet më të mëdha në veprimet paraproçesuese në sistemet Text Mining.

Meqënëse algoritmet në Text Mining aplikohen mbi *modelin përfaqësues* të dokumentit, (bazuar në karakteristikat e paraqitjes), dhe jo drejtpërdrejtë në dokumentat tekst, rezultojnë dy objektiva kryesorë: I pari është kalibrimi korrekt i sasisë dhe nivelit semantik të karakteristikave që të mbartë kuptimin e dokumentit. Kjo tendencë e çon paraproçesimin në Text Mining drejt zgjedhjes dhe nxjerrjes së një sasie gjithmonë e më të madhe karakteristikash përfaqësuese të dokumentit tekst. Objektivi i dytë është identifikimi i modelit në mënyrë që të jetë më efiçent dhe praktik nga ana komputacionale, i cili është një proçes që thekson kompresimin e bashkësisë së karakteristikave ne bashkësi përfaqësuese të vogla të cilat t'i korrespondojnë karakteristikave të pasura nga ana semantike (Feldman and Sanger, 2007).

Pavarësisht se për të përfaqësuar dokumentin tekst mund të përdorim karakteristika të shumëllojshme, më të përdorshmet janë katër tipet e mëposhtme:

- ***Karakteret**:* shkronjat, numrat, karakteret speciale dhe hapësirat përbëjnë themelin e karakteristikave semantike si fjalët, termat apo konceptet. Një përfaqësim në nivel karakteri mund të përfshijë të gjithë bashkësinë e karaktereve të dokumentit ose një nënbashkësi të filtruar të saj. Përfaqësimet në nivel karakteri të cilat nuk përmbajnë informacion pozicional janë pothuajse të padobishme në aplikacionet e Text Mining, ndërkohë që në rastin kur shoqërohen me informacion mbi pozicionin, (psh. Bigrame ose trigrame), janë disi më të dobishme dhe të përdorshme. Në përgjithësi, përfaqësimet

e bazuara në karaktere nuk janë shumë të optimizuara, por nga ana tjetër shihen si përfaqësimi më i plotë i një dokumenti tekst.

- ***Fjalët**:* fjalët specifike të përzgjedhura direkt nga dokumenti, mund të përshkruhen si niveli bazë semantik. Në përgjithësi, një karakteristikë në nivel fjale, duhet t'i korespondojë vlerës së jo më shumë se një token-i linguistik. Një dokument tekst mund të përfaqësohet nga një karakteristikë për çdo fjalë brenda atij dokumenti, gjë që çon në krijimin e paraqitjeve në nivel fjalësh që kanë qindra mijëra fjalë unike në hapësirën e karakteristikave të koleksionit të dokumentave. Megjithatë, paraqitjet e dokumentave në nivel fjalësh realizohen duke përdorur të paktën disa optimizime minimale për të filtruar karakteristikat përfaqësuese nga elementë si shenja pikësimi, karaktere simbolike apo numra pa ndonjë kuptim.
- ***Termat***: janë shprehje që përbëhen nga një fjalë e vetme ose grup fjalësh që përzgjidhen direkt nga dokumenti nëpërmjet metodologjive të nxjerrjes së termave. Karakteristikat në nivel termash përbëhen nga fjalë dhe shprehje specifike të gjetura në dokumentin bazë, për të cilin janë përfaqësuese. Pra, një paraqitje e bazuar në terma, përbëhet nga një nënbashkësi termash në ate dokument. Për shembull nëqoftëse një dokument ka një fjali:

  Shkrimtari Ismail Kadare është dekoruar nga Presidenti i Republikës së Shqipërisë me Urdhrin "Nderi i Kombit" dhe nga shteti francez me urdhrat "Kryqi i Legjionit të Nderit" e "Oficer i Legjionit të Nderit".

  Një listë termash që përfaqësojnë dokumentin përfshijnë forma me një fjalë si "Kadare", "dekoruar", "presidenti", "shteti", "francez", "urdhrat" dhe forma me shumë fjalë si "Shkrimtari Ismail Kadare", "Presidenti i Republikës së Shqipërisë", "Nderi i Kombit", "shteti francez", "Kryqi i Legjionit të Nderit", "Oficer i Legjionit të Nderit".

  Shumica e metodave të nxjerrjes së termave mund të konvertojnë tekstin e papërpunuar në një dokument, në terma të normalizuara – sekuenca të një ose më shumë formave të fjalëve të tokenizuara (ang. tokenized) dhe lematizuara (ang. lemmatized) dhe të shoqëruara me etiketa të pjesëve të fjalive (part of speech tags). Këto metoda gjithashtu, përdorin qasje për gjenerimin dhe filtrimin e një liste të shkurtuar dhe koncentruar të termave kandidatë për përfaqësimin e dokumentit. Ky proçes rezulton në një përfaqësim më të vogël, por më të pasur nga ana semantike të dokumentit në krahasim me perfaqësimin e dokumentit në nivel fjalësh.
- ***Konceptet***: janë karakteristika të dokumentit të gjeneruara në mënyrë manuale, statistikore, të bazuar ne rregulla ose metodologji hibride. Aktualisht, konceptet nxirren nga dokumenti duke përdorur operacione komplekse paraproçesimi të cilat identifikojnë fjalë, shprehje, paragrafë ose njësi më të mëdha sintaksore që lidhen me *identifikatorë konceptesh*. Për shembull, një koleksion dokumentash në lidhje me makinat sportive, mund të mos e përmbajë fjalën "automobilistike" ose shprehjen "garë

provë”, megjithatë këto koncepte mund të gjenden midis bashkësisë së koncepteve që përdoren për të identifikuar dhe paraqitur atë koleksion dokumentash. Shumë metoda kategorizimesh përmbajnë edhe një referencë me një burim të jashtëm informacioni; për disa metoda statistikore, ky burim mund të jetë thjesht një bashkësi e caktuar dokumetash për testim. Metodat manuale të kategorizimit të bazuara në rregulla, referenca të jashtme dhe validime, zakonisht ndërveprojnë me një standard referimi si për shembull një ontologji e fushës, një leksikë ose një koncept formal.

2.1.1.2. Kërkimi i Modeleve

Sistemet Text Mining bazohen në algoritma dhe heuristika që marrin në konsideratë konceptet e shpërndarjes, frekuencave dhe lidhjeve të ndryshme në nivel grupi dokumentash (angl. Interdocument level), me qëllim zbulimin e natyrës dhe lidhjes së koncepteve që reflektohen nga e gjithë bashkësia e dokumentave. Për shembull, në një koleksion artikujsh, një numer i madh artikujsh për një politikan X dhe një skandal Y, mund të shërbejë si indikacion për prishje të imazhit te politikanit X, ose një numër në rritje i artikujve për kompaninë Z dhe artikullin T mund të tregojë ndryshim të interesave të kompanisë Z. Në një shembull tjetër do të nxjerrim lidhjen potenciale midis proteines P1 de P2 nga modelet (a) disa artikuj ilustrojnë lidhjen e proteines P1 me enzimën E1, (b) disa artikuj përshkruajnë ngjashmëri funksionale midis enzimës E1 dhe enzimës E2, dhe (c) disa artikuj e lidhin enzimen E2 me proteinën P2. Në të tre shembujt, informacioni nuk merret nga një dokument individual, por nga bashkësia e disa prej tyre. Metodat Text Mining për analizën e modeleve, tentojnë të gjejnë lidhje të ndërthurura të koncepteve që rezultojnë nga bashkësia e koncepteve ne shqyrtim. Metodat e Text Mining – zakonisht të bazuara në kërkime brute-force – prodhojnë një bashkësi të madhe modelesh, gjë që rezulton në një problem teprie informacioni. Një nga detyrat kryesore në sistemet Text Mining është lejimi i përdoruesve që të limitojnë teprinë e modeleve rezultat, duke mundësuar vendosjen e indekseve të rëndësisë për rezultatet e kërkimit. Të tilla indekse ndihmojnë përdoruesit e sistemit që të marrin informacionin më të rëndësishëm. Megjithatë mund të themi që problemi i teprisë së informacionit ekziston në të gjithë aktivitetet e zbulimit të njohurive (angl. Knowledge discovery activities), ai thjeshtë theksohet më shumë nëqoftëse ndërveprohet me koleksione të mëdha dokumentash. Për këtë arsye, veprimet e Text Mining duhet të projektohen jo vetëm që të japin informacionin e dëshiruar, por edhe që ky informacion të jetë një bashkësi e menaxhueshme për përdoruesin.

2.1.1.3. Paraqitja e Rezultatit

Funksionaliteti më i rëndësishëm i paraqitjes së rezultateve nga sistemet Text Mining është shfletimi i rezultateve (angl. browsing). Sistemet e fundit Text Mining suportojnë shfletimin i cili është njëkohësisht dinamik dhe i bazuar në përmbajtje (angl. Content-based), meqënëse shfletimi bazohet në përmbajtjen e një koleksioni të caktuar dokumentash dhe jo nga strukturat e

paracaktuara në mënyrë strikte. Shfletimi lehtësohet nga paraqitja grafike e modelit të koncepteve në formë hierarkike.

Sistemet Text Mining e vënë përdoruesin përballë një bashkësie jashtëzakonisht të madhe konceptesh që merren nga koleksionet e gjëra të dokumentave. Rrjedhimisht, duhet t'i mundësohet përdoruesve të kërkojnë nëpër këto koncepte duke patur gjithmonë mundësi të zgjedhin midis pamjes së përgjithshme të koleksionit, ose të detajohen në koncepte të veçanta të identifikuara. Mjetet e vizualizimit përdoren shpesh nga sistemet Text Mining për të lehtësuar kërkimin dhe eksplorimin e modeleve konceptuale. Ato përdorin elementë grafikë për të shprehur lidhjet komplekse midis të dhënave. Në të shkuarën mjetet e vizualizimit për Text Mining gjeneronin harta statike, grafe të paraqitura si fotografi e çastit e modelit ose raporte të shfaqura në ekran ose të printuara. Sistemet e sotme Text Mining bazohen totalisht në paraqitjen grafike interaktive të rezultateve të kërkimit që lejojnë përdoruesin të tërheqë (angl. drag), shtyjë (angl. pull), klikojë ose të ndërveprojë direkt me paraqitjen grafike të modelit konceptual (Feldman and Sanger, 2007).

2.1.2. Arkitektura e Sistemeve Text Mining

Në nivel abstrakt, një sistem Text Mining merr input, (dokumenta të papërpunuara), dhe gjeneron tipe të ndryshme outputi, (psh. modele, harta lidhjesh, trende, etj). Ne Figurën 2.1 ilustrohet ky paradigëm bazë. Nga këndvështrimi i përdoruesit, zbulimi i njohurive na çon drejt një paradigmi input-output më kompleks për Text Mining. Në këtë paradigëm përdoruesi është pjesë e asaj që mund të quhet cikël interaktiv i kërkimit (angl. querying), shfletimit dhe perfeksionimit që rezulton në një bashkësi rezultat e cila nga ana tjetër e drejton përdoruesin drejt një serie iterative veprimesh kërkimi, shfletimi dhe rafinimi.

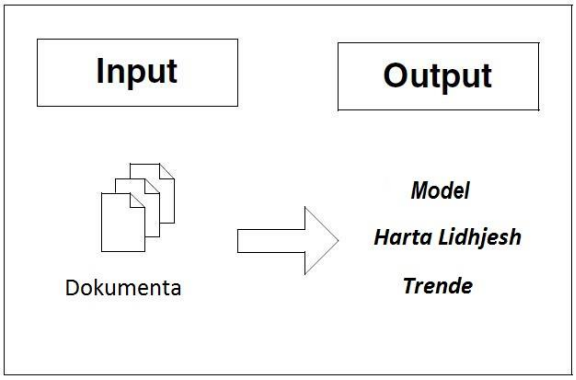

Figura 2.1. Modeli Input-Output i Text Mining (Feldman and Sanger, 2007)

Në nivel funksional, një sistem Text Mining ndjek modelin klasik te aplikacioneve Data Mining, dhe si i tillë ndahet në katër fusha kryesore. (a) paraproçesimi (b) veprimet bazë të

kërkimit (c) komponentët e paraqitjes dhe funksionalitetet e shfletimit dhe (d) teknikat e përmirësimit të rezultatit.

- **Paraproçesimi** përfshin të gjitha proçedurat, proçeset dhe metodat që nevojiten për të pregatitur të dhënat për veprimet bazë të zbulimit të njohurive në një sistem Text Mining. Këto detyra bazohen në veprimet e paraproçesimit dhe të kategorizimit të të dhënave burim (angl. data source). Veprimet e paraproçesimit zakonisht e konvertojnë secilin burim fillestar informacioni në format kanonik para se të aplikojnë mbi to metoda të ndryshme të nxjerrjes së karakteristikave duke krijuar një koleksion të ri dokumentash të cilët përfaqësohen plotësisht nga koncepte. Gjithashtu, nëqoftëse mundet, veprimet paraproçesuese mund të nxjerrin ose aplikojnë rregulla për vendosjen e formatit të datës.
- **Veprimet Bazë të Kërkimit** janë zemra e një sistemi Text Mining dhe përfshijnë gjetjen e modeleve, analizën e trendit dhe algoritmet inkrementuese të zbulimit të njohurive (angl. Knowledge Discovery). Midis modeleve të ndryshme që përdoren për zbulimin e njohurive në të dhënat tekst jane shpërndarjet (angl. distributions), frekuencat e bashkësive të koncepteve dhe lidhjet (angl. associations). Veprimet bazë të zbulimit mund të përdorin edhe krahasimet – ose identifikimit të nivelit të prioritetit - me disa prej modeleve. Sistemet e avancuara, ato të orientuara nga domain-i ose të dyja mund të theksojnë cilësinë e veprimeve të ndryshme që ato kryejnë duke shfrytëzuar kontekstin (angl. background knowledge). Për këto veprime bazë të një sistemi Text Mining mund t'i gjejmë të referuara si proçese të *distilimit të njohurive* (angl. knowledge distillation processes).
- **Komponentët e paraqitjes** përfshijnë GUI dhe funksionalitete shfletimi mbi modelet, po ashtu edhe aksesin në gjuhën për query. Mjetet e vizualizimit dhe editorët për ndërtimin dhe optimizimin e query-ve gjithashtu janë pjesë e kësaj arkitekture. Komponentët në nivelin e paraqitjes mund të përfshijnë mjete grafike ose të bazuara në karaktere për të krijuar ose modifikuar bashkësi koncepte (angl. concept clusters) ose për të krijuar profile të cilësuara për modele ose koncepte specifike (Feldman and Sanger, 2007).
- **Teknikat e përmirësimit të rezultatit**, në rastin më të thjeshtë përfshijnë metoda që filtrojnë informacionin dhe grupojnë të dhënat e ngjashme, ndërsa kur këto teknika janë më të plota, ofrojnë qasje në ngjeshje, renditje, prerje, përgjithësime dhe grupime me qëllim optimizimi të rezultateve. Këto teknika njihen si *postproçesim*.

Paraproçesimi dhe veprimet bazë kërkuese janë dy pjesët më kritike në çdo sistem Text Mining siç paraqiten në figurën e mëposhtme.

Figura 2.2. Arkitektura funksionale e sistemeve Text Minig në High-Level (Feldman and Sanger, 2007)

2.1.3. Algoritmat dhe Metodat

Në këtë çështje do të trajtojmë problemet kryesore që lindin në kontekstin e Text Mining.

- **Marrja e Informacionit të Strukturuar nga Dokumenta Tekst.**
  Marrja e Informacionit të Strukturuar nga dokumenta tekst është një nga problemet kyçe në Text Mining, i cili shërben si element fillestar për shumë algoritme në Text Mining. Për shembull, nxjerrja e entiteteve dhe lidhjeve midis tyre nga teksti, mund të na japë me shumë informacion semantik se sa një "grumbull me fjalë", prandaj zakonisht përdoret për të suportuar inferencat rreth njohurive të fshehura në të dhëna tekst. Marrja e Informacionit të Strukturuar do te jetë një çështje e cila do të trajtohet më gjerësisht në vazhdim.
- **Përmbledhje e Tekstit**
  Një funksion shumë i rëndësishem i nevojshën në shume aplikacione Text Mining është formulimi i një informacioni përmbledhës për një dokument të madh tekst ose një grup dokumentash teskt të së njëjtës fushë. Teknikat përmbledhëse i përkasin dy kategorive: përmbledhje me përzgjdhje, ku përmbledhja bazohet në fragmente të përzgjedhura nga teksti fillestar dhe përmbledhje sintetizuese ku përmbledhja përmban informacion të sintetizuar i cili jo domosdoshmërisht gjendet në tesktin fillestar (Nenkova & McKeown, 2012).
- **Metoda Automatike e të Mësuarit (angl. Unsupervised Learning Method)**
  Metodat Automatike të të Mësuarit nuk kërkojnë të dhëna për t'u trajnuar, prandaj mund të aplikohen mbi çdo tekst pa kosto shtesë të veprimeve manuale. Këto metoda ndahen në dy kategori të cilat janë grupimi (angl. clustering) dhe modelimi i fushës (angl. topic modeling). Grupimi konsiston në ndarjen e bashkësisë në grupe dokumentash sipas ngjashmërisë me fushat/temat e paracaktuara. Modelimi i fushës konsiston në grupimin e dokumentave tekst bazuar në modele probabilitare. Kjo lidhet ngushtë me problemin e zvogëlimit të dimensionve ku fushat janë dimensione konceptuale ndërsa dokumentat tekst paraqiten si kombinacione lineare probabilitare të të gjitha fushave.
- **LSI dhe Reduktimi Dimensional**

Problemi i reduktimit të dimensioneve studiohet gjërësisht në kërkimet që trajtojnë bazat e të dhënave ku të dhënat fillestare përdoren në format të kompresuar për ekstraktim dhe indeksim (Jolliffee, 2002). Indeksimi semantik është një formë e reduktimit të dimensioneve e cila përdoret në të dhëna tekst në mënyrë që të përcaktohen aspektet semantike të tekstit duke e bërë atë më të përshtatshëm për aplikacione të ndryshme kërkimi. Një shembull i përdorimit të kësaj teknike është reduktimi i efekteve të zhurmës të sinonimisë dhe polisemisë. Një tjetër formë e reduktimit të dimensioneve janë modelet probabilitare si PLSA, LDA të cilat e realizojnë reduktimin dimensional sipas një përqasjeje probabilitare në të cilën fushat/temat krijohen duke u bazuar në shpërndarjen e fjalëve.

- **Metoda e të Mësuarit të Kontrolluar**
  Metodat e të mësuarit të kontrolluar përdorin të dhëna për trajnim në mënyrë që të gjejnë një klasifikator ose funksion regresioni i cili më pas do të përdoret në të dhënat e reja. Ky njihet edhe si problem klasifikimi. Për të gjetur zgjidhje të problemeve të ndryshme në Text Mining përdoren klasifikatorë si: decision trees, rule-based, nearest neighbor, probabilistic dhe maximum margin.
- **Transferimi i të Mësuarit me të Dhëna Tekst**
  Transeferimi i të Mësuarit i takon një grupi tjetër metodash në të Mësuarit e Kontrolluar ku nuk përdoren të dhëna për trajnim. Problemi i transferimit të të mesuarit konsiston në nxjerrjen/ekstraktimin dhe transferimin e njohurive të mësuara midis domain-eve të ndryshme. Për shembull, në një bashkësi me të dhëna tekst dhe multimedia, preferohet të përdoren metodat e transferimit të njohurive siç janë cross-media transfer midis domain-eve.
- **Teknika Probabilitare të Text Mining**
  Ekzistojnë teknika të ndryshme probabilitare duke përfshirë edhe Modelet Automatike dhe Metodat e Kontrolluara të të Mësuarit të cilat janë algoritmat më të përdorshme në Text Mining. Këto teknika përdorin teorinë e probabilitetit për të modeluar ambiguitetin e të dhënave. Këto modele probabilitare përdorin një bashkësi mundësish shpërndarjesh probabilitare që lidhen me të dhënat e observuara dhe që përdoren për të gjetur shpërndarjen probabilitare më të dobishme (Feldman and Sanger, 2007).

2.2. Gjetja e Informacionit të Kërkuar (Information Retrieval)

Kuptimi që lidhet me termin Gjetje e Informacionit të Kërkuar (GjIK), është shumë i gjerë, megjithatë si një fushë akademike ajo mund të përkufizohet si më poshtë:

GjIK është gjetja e materialit, (zakonisht dokumenta) pa strukturë, (teskt) që kënaq nevojën për informacion nga koleksione të mëdha (të ruajtura në kompjuter).

Ky përkufizim tregon që GjIK ka qënë një aktivitet i kryer nga vetem pak kategori personash, si p.sh. punonjësit e bibliotekave, asistentët ligjorë dhe profesione të tjera kërkimi. Në ditët e sotme gjërat kanë ndryshuar dhe shumë njerëz punojnë çdo ditë me sisteme GjIK kur përdorin motorrët e kërkimit në web ose kur kërkojnë në emailin e tyre. GjIK po bëhet forma kryesore e aksesimit të informacionit, duke lënë pas kërkimin klasik në databaza.

Termi "të dhëna pa strukturë" i referohet të dhënave që nuk kanë një strukturë semantike të qartë dhe të lehtë për t'u kapur nga një kompjuter. Shembull i kundërt i të dhënave pa strukturë janë bazat e të dhënave relacionale, të cilat kompanitë i përdorin për të mbajtur inventarin e produkteve apo të dhënat e personelit. Në të vertetë asnjë e dhënë nuk është plotësisht pa strukturë, meqënëse çdo tekst përmban edhe një strukturë sintaksore të fshehur të gjuhës dhe strukturë përbërëse që lidhet me paragrafët, titujt, etj.

Fusha e GjIK mbulon edhe suportin e përdoruesve në eksplorimin dhe filtrimin e bashkësive të dokumentave ose në proçesimin e mëtejshëm të dokumentave të gjetur. Nëqoftëse jepet një bashkësi dokumentash, krijimin e clusterave është ndarja e dokumentave në disa grupe të caktuara bazuar në përmbajtjen e tyre. Nga ana tjetër, nëqoftëse na jepen një bashkësi kategorish (psh, grupmoshat e lexuesve të librave), klasifikimi është përcaktimi i klasës së cilës i përket çdo dokument ose bashkësi dokumentash. Shpesh herë kjo arrihet duke klasifikuar në mënyrë manuale disa dokumenta që më pas të mundemi të klasifikojmë dokumentat e rinj.

Sistemet e GjIK dallohen nga njëri tjetri nga shkalla e të dhënave mbi të cilën ato veprojnë. Këtu mund të veçojmë tre shkallë. Në *motorrët e kërkimit web,* sistemi duhet të mundësojë kërkimin në miliarda dokumenta të ruajtur në miliona kompjutera. Çështja që lind këtu është mbledhja e dokumentave për indeksim, duke bërë të mundur ndërtimin e sistemeve që funksionojnë normalisht me sasi të tilla të dhënash që menaxhojnë edhe aspekte të veçanta të web siç është shfrytëzimi i hypertext dhe mos lejimi i manipulimit të faqeve nga providerat të cilët tentojnë të manipulojnë përmbajtjen e faqes me qëllim rritjen artificiale të rëndësisë së saj nga motorrët e kërkimit. Nga ana tjetër është *Gjetja e Informacionit Personal*. Sistemet operative në ditët e sotme vijnë të integruar me GjIK. Programet për email mundësojnë jo vetëm kërkimin, por edhe klasifikim teksti: përmbajnë të paktën filtër për spam dhe ndonjëherë edhe mjete automatike ose manuale për klasifikimin e emailit në menyrë që të dërgohet direkt në direktori të caktuara. Disa çështje që lindin këtu përfshijnë trajtimin e llojit të ndryshëm të dokumentave që ndodhen zakonisht në një kompjuter personal dhe thjeshtimin e mirëmbajtjes së sistemit të kërkimit. Midis këtyre shkallëve është kërkimi institucional, ku kërkimi realizohet në dokumentat e brendshëm të një korporate, një bazë të dhënash ose artikuj shkencorë. Në këtë rast, dokumentat ruhen në sisteme qëndrorë skedarësh dhe një ose disa kompjutera të dedikuar do të mundësojnë kërkimin në kolesionin e dokumentave (Manning et al, 2009).

Për të gjetur dokumentat e duhur nga një teknikë e GjIK, dokumentat transformohen në një paraqitje të përshtatshme. Çdo strategji gjetjeje, aplikon një model specifik me qëllim krijimin e paraqitjes së nevojshme të dokumentit (Turtle and Croft, 1992). Modelet më të njohura të paraqitjes kategorizohen në dy dimensione: bazat matematike dhe karakteristikat e modelit, siç paraqitet në figurën e mëposhtme (Kuropka, 2004).

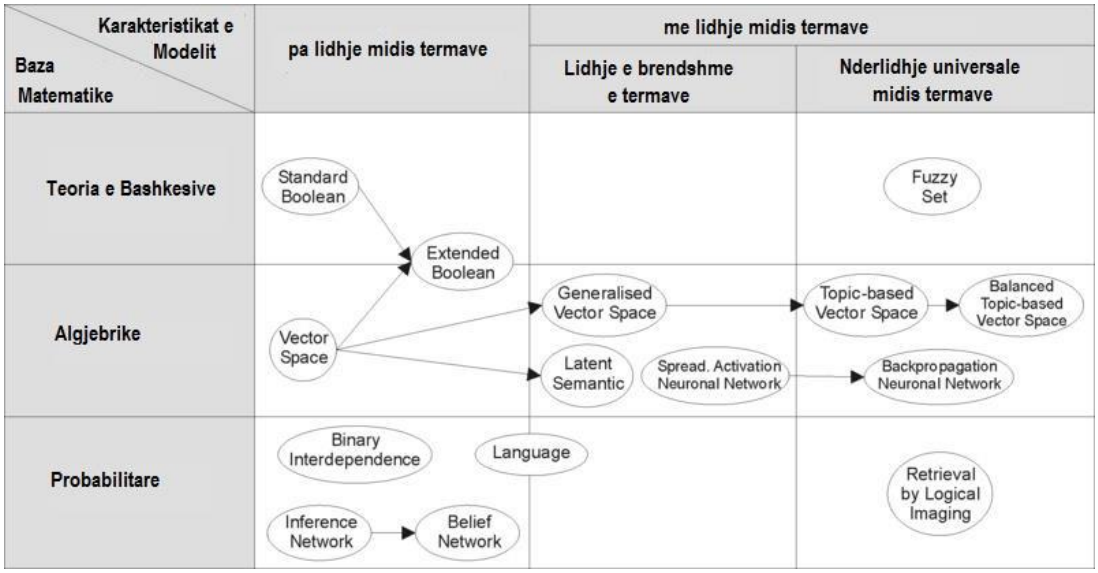

Figura 2.3. Kategorizimi i modeleve të sistemeve të GjIK (Kuropka, 2004)

2.2.1. Bazat Matematike

Modelet bazë që përdoren në sistemet GjIK janë modeli Boolean, modeli vektorial dhe modeli probabilitar të cilat përfshihen në dimensionin e parë, Bazat Matematike (Baeza-Yates and Ribeiro-Neto, 1999). Në modelin Boolean dokumentat dhe query-t paraqiten si një bashkësi termash indeks, pradaj këto janë modele që i takojnë Teorisë së Bashkësive. Në modelin e hapësirës vektoriale (angl. Vector Space Model, VSM), dokumentat dhe query-t paraqiten si vektorë në hapësira shumë-dimensionale, prandaj modelet janë algjebrike. Në modelin probabilitar, framework-u për të modeluar dokumentat dhe paraqitjen e query-ve bazohet në teorinë e probabilitetit , prandaj modeli është probabilitar (Manninget al, 2009).

2.2.1.1. Teoria e Bashkësive

Modeli i gjejtes Booleane është një model për GjIK në të cilin ne mund të shtrojmë çdo lloj query në formën e një shprehjeje Booleane termash, në të cilën termat kombinohen me anë të operatorëve AND, OR dhe NOT. Modeli e sheh çdo dokument si një bashkësi termash. Çdo term do ta konsiderojmë si: një fjalë të vetme psh "*aeroplan*"; një n-gram, psh "*aer*", "*roplan*", "*plan*", etj; një togfjalësh psh "*republikë parlamentare*", "*Fakulteti i Teknologjisë së Informacionit*", etj.

Qëllimi i këtij modeli është të ndërtojë një sistem që të adresojë veprimet e gjetjes ad-hoc. Ky është veprimi standard në sistemet e GjIK i cili ka për qëllim të gjejë dokumentat që plotësojnë nevojën e përdoruesit për informacion. Çdo nevojë për informacion i komunikohet sistemit nëpërmjet një query të formuluar nga përdoruesi. Një dokument konsiderohet që përmbush nevojën e përdoruesit, nëqoftëse ka informacion me vlerë, bazuar në kërkesën e marrë paraprakisht.

Për të ilustruar këtë model do të marrim një shembull konkret. Supozojmë që kemi N = 1 milion dokumenta. Kësaj bashkësie do t'i referohemi si *koleksion* ose *korpus*. Supozojmë që çdo dokument përmban rreth 1000 fjalë dhe çdo fjalë kërkon afërsisht 6 byte për t'u shprehur. Në total do të kishim një korpus me madhësi 6 GB. Në një bashkësi të tillë llogaritet të ketë rreth 500 000 terma/tokens të ndryshëm. Shifrat e mësipërme na krijojnë një ide mbi vlerat me të cilat do të punojmë. Matrica term-dokument duhet të mbajë ndodhjen ose jo në secilin dokument për të gjithë termat e ndryshëm të bashkësisë. Konkretisht nuk do të mund ta ndërtonim këtë matrice me përmasa 500K x 1M dhe gjysëm trilioni 0 dhe 1-sha sepse është e pamundur të ruhet në kujtesën e një kompjuteri.

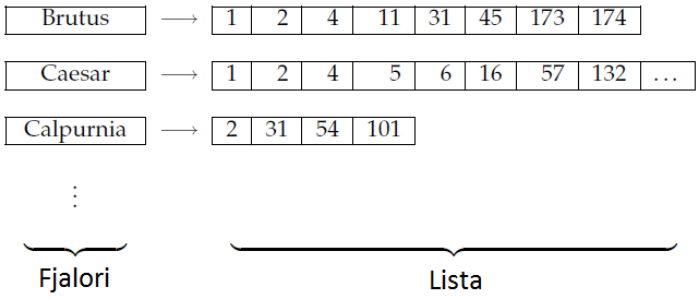

Figura 2.4. Dy pjesët e indeksit të invertuar. Fjalori ruhet në kujtesë me pointera në secilën listë që ruhet në disk (Manninget al, 2009).

Në këtë rast, meqënëse matrica është e rrallë ruajmë vetëm ndodhjet e termave, pra vetëm pozicionet e 1-shave. Ky është një nga konceptet bazë të GjIK që lidhet me *indeksin e invertuar*, term ky i përdorur gjërësisht në GjIK për referencën nga termi drejt pjesës së dokumentit ku ky term ndeshet. Për secilin term të koleksionit – do ta quajmë fjalor të korpusit, ruhet lista e dokumentave në të cilat ndeshet ky term. Fjalori renditet sipas alfabetit dhe secila listë renditet sipas ID së dokumentave, siç tregohet në Figurën 2.4.

Për të optimizuar kohën e kërkimit, indeksimi i dokumentave bëhet në avancë. Hapat kryesorë të tij janë:

1. Mblidhen dokumentat që duhet të indeksohen.
2. Tokenizohet teksti, duke e kthyer çdo dokument në një listë me token.

3. Realizohet paraproçesimi linguistik, i cili gjeneron një listë me token të normalizuar që do të shërbejnë si terma indeksimi.
4. Indeksohet çdo dokument ku ndeshet secili term duke krijuar indeksin e invertuar i cili konsiston në një fjalor dhe një bashkësi listash.

Paraproçesimi ka si veprime kryesore heqjen e fjalëve të përgjithshme si: "dhe", "të", "atëherë", "një" , etj, meqënëse këto nuk mbartin informacion të lidhur me përmbajtjen e dokumentave dhe gjetjen e rrënjës së fjalëve mënyrë që indeksimi të ketë sa më pak variacione të panevojshme. Çdo dokument në korpusin e dokumentave ka një numër serial unik, docID. Gjatë fazës së indeksimit mund të caktojmë numra të njëpasnjëshëm për çdo dokument të ri. Si input shërben një listë me token të normalizuar për secilin dokument, si një listë çiftesh term – docID siç tregohet ne Figurën 2.5.

Përmbledhja e Shakespeare-t (angl. Shakespeare's Collected Works) është një volum libri që përmban të gjitha veprat dhe poemat e William Shakespeare. Konsiderohet i përshtatshëm për t'u përdorur për teste në Text Mining sepse përmban numër shumë të madh fjalësh, mund të sigurohet lehtë në format elektronik dhe zotërohet nga një numër i madh njerëzish. Kjo përmbledhje është përdorur si bazë të dhënash në shumë algoritme të GjIK.

Hapi kryesor i indeksimit është renditja e kësaj liste sipas rendit alfabetik të termave. Ndodhjet e shumëfishta të të njëjtit term në një dokument eleminohen, ndërsa ndodhjet e të njëjtit term në dokumenta të ndryshëm bashkohen duke formuar ciftin term – listë duke zvogëluar nevojat e indeksit për kujtesë. Fjalori mund të mbajë edhe disa statistika, si psh, numrin dokumentave në të cilët ndodhet secili term – kjo vlerë është edhe gjatësia e listës së secilit term. Ky informacion shtesë rrit efiçensën e motorrëve të kërkimit Booleanë. Për proçesimin e çdo query boolean mbi këtë indeks bazohemi në të gjitha rregullat e Teorive të Bashkësive. P.sh. nqs query mbi këtë bashkësi do të ishte "Brutus AND Calpurnia", së pari do të lokalizonim në fjalor termin Brutus, e më pas do të merrnim listën e këtij termi. Të njëjtën gjë do të kryenim edhe për termin e dytë. Rezultati i query-it do të ishte prerja e dy listave. Në mënyrë analoge mund të vepronim për çdo lloj query, duke kombinuar veprimet e prerjeve dhe bashkimeve të bashkësive. Është e rëndësishme që këtu të kemi parasysh edhe optimizimin e query-ve sipas rasteve.

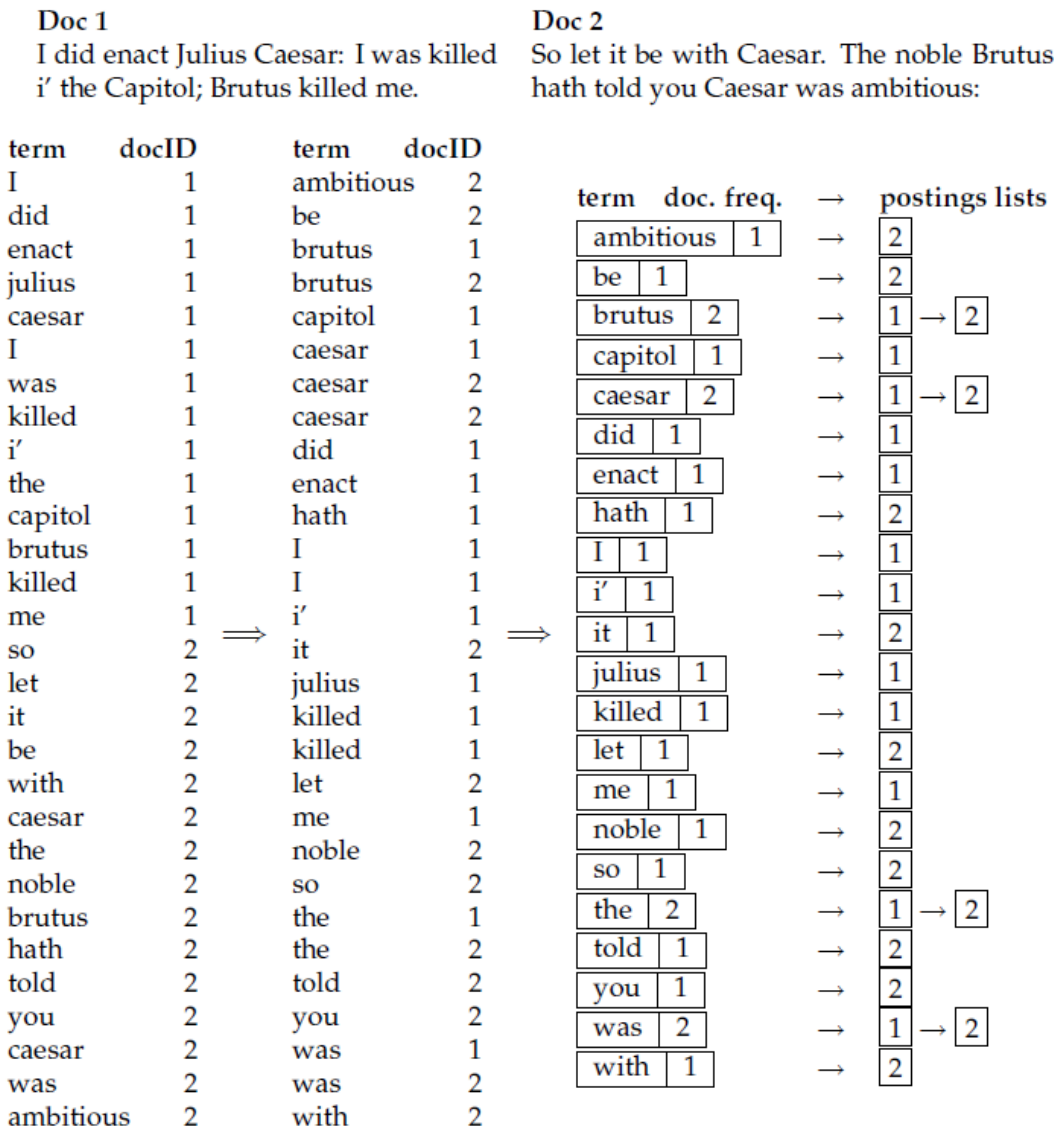

Figura 2.5. Ndërtimi i indeksit me anë të renditjes dhe grupimit, bazuar në Përmbledhjen e Shakespeare (Manninget al, 2009).

Modeli i gjejtjes Booleane ndryshon ndjeshëm nga modelet e rankuara të gjetjes siç është Modeli i Hapësirës Vektoriale (angl. Vector Space Model), në të cilin përdoren query në format teksti të thjeshtë, domethënë pa përdorur një gjuhë specifike me operatorë të përshtatshëm për shprehjet e query-ve. Pavarësisht avantazheve të modeleve të rankuara te gjejtjes, modelet Booleane kanë qënë më të përdorurat nga providerat komercialë të informacionit për tre dekada, deri në vitet 90. Këto sisteme përveç operatorëve Booleanë bazë, kishin edhe disa operatorë të tjerë shtesë, si operatori i afërsisë së termit. Operatori i afërsisë është një mënyrë për të vendosur që dy terma duhet të ndeshen në afërsi me njëri-tjetrin në një dokument. Afërsia matet duke limituar numrin e fjalëve që gjenden midis ose duke iu referuar një njësie strukturore si fjalia ose paragrafi. Këto ndryshime kanë krijuar modelin e zgjeruar të gjetjes Booleane. Modelet Booleane të zgjeruara

tentojnë të përmirësojnë rezultatin e kërkimit duke përdorur konceptin e Bashkësive të Paqarta (angl. fuzzy sets) të propozuara nga Zadeh në 1965. Sipas kësaj teorie një element ka një shkallë pranie dA në një bashkësi A që bie ne intervalin [0,1], ndryshe nga modeli tradicional ku një element e ka praninë në një dokument 0 ose 1 – është element i bashkësisë apo nuk ështe element i bashkësisë (Frakes And Baeza-Yates, 1992).

Çështjet kryesore që trajtohen në këto modele janë:

- Të përcaktohet më mirë bashkësia e termave në fjalor dhe të mundësohet gjetja me tolerance ndaj gabimeve të drejtshkrimit apo zgjedhjes jo të mirë të fjalëve të kërkimit.
- Shpesh kërkohet sipas një fjalie ose shprehjeje për të identifikuar një koncept si psh, sistem operativ. Për t'iu përgjigjur query-ve të tilla, indeksimi duhet të zgjerohet me afërsitë e termave në dokumenta.
- Një model Boolean ruan vetëm praninë ose mungesën e termave në nje dokument, ndërkohë që ne mund të na duhet t'i japim më shumë peshë një dokumenti që e përmban termin disa herë në krahasim me një dokument që e ka atë vetëm një herë. Në këtë rast na duhet informacion mbi frekuencën e termave në listë.
- Query-t Booleane kthejnë një bashkësi me dokumentat që plotësojnë kërkesën, dhe nuk e rendit bashkësinë rezultat sipas rëndësisë. Realizimi i rankimit të rezultateve kërkon një mekanizëm për të vendosur një sasi pikësh për dokumentin për të mbajtur informacion sa mirë e plotëson query-n ai dokument.

2.2.1.2. Algjebra

Modelet algjebrike përfaqësojnë dokumenta dhe query në formë vektorësh, matricash ose rekordesh (angl. tuples). Në këto modele ngjashmëria midis vektorit të query dhe vektorit të dokumentit paraqitet si vlerë skalare. Deri tani e kemi parë një dokument si sekuencë termash, pa marrë parasysh strukturën e dokumentit. Dokumentat elektronikë përmbajnë edhe metadata të cilat kanë fusha të tilla si data e krijimit, formati i dokumentit, autori apo titulli. Vlerat e një fushe konsiderohen si vlera të fundme. Query "*Gjej të gjithë dokumentat që janë shkruar nga William Shakespeare në 1601 dhe përmbajnë shprehjen 'Të jesh a të mos jesh'*" do të kthejë një rezultat duke shfrytëzuar metadata. Nga ana tjetër, nëqoftëse, një term i query-t ndeshet disa herë në një dokument, atëherë ai dokument ka të bëjë më shumë me atë query, prandaj duhet t'i caktohet peshë më e lartë. Një mekanizëm për renditjen e rezultateve të një query sipas rëndësisë do të ishte llogaritja e shumës së termave të kërkimit. Për të realizuar këtë, çdo termi në dokument i vendoset një peshë e cila varet nga frekuenca e termit në atë dokument. Qëllimi në këtë rast, është të caktohet një vlerë skalare për termin *t* të query-t në dokumentin *d*, bazuar në peshën e *t* në *d*. Zgjidhja më e thjeshtë është të vendoset si peshë e termit numri i herëve që ai term gjendet në dokumentin *d*. Kjo skemë peshash njihet si *Frekuencë Termash* (angl. term frequency) dhe shënohet $tf_{t,d}$ për termin *t* në dokumentin *d*. Vendosja e peshave sipas një funksioni që varet nga

frekuenca e fjalëve konsiderohet si përmbledhje sasiore e dokumentit, ndërsa renditja e fjalëve brenda dokumentit nuk merret parasysh. Pesha/frekuenca e llogaritur në këtë mënyrë i konsideron të gjithë termat e dokumentit njëlloj të rëndësishëm përballë përmbushjes së query-t, rrjedhimisht disa terma kanë pak ose aspak fuqi diskriminuese, psh termi "*makinë*" ndeshet pothuajse në të gjithë dokumentat që flasin për industrinë e makinave. Në këtë mënyrë ai nuk ka fuqi përcaktuese dhe diskriminuese. Për të realizur këtë skemë peshash duhet të përdoret frekuenca e dokumentave df_t, e cila tregon numrin e dokumentave në korpus që e përmbajnë termin *t*. Për të llogaritur peshën e një termi sipas kësaj logjike përdoret formula e mëposhtme, e cila quhet *frekuencë inverse e dokumentit* (angl. inverse document frequency).

$$idf_t = \log \frac{N}{df_t} \qquad \text{ku N është numri total i dokumentave në korpus}$$

Mund të themi që *idf* e një termi që nuk haset shpesh ne korpus është e lartë, ndërsa *idf* e një termi të shpeshtë është e ulët. (Manning et al, 2009).

Nga sa thamë më sipër, kombinimi i frekuencës së termit, *tf*, dhe frekuencës inverse të dokumentit, *idf*, do të na japë rëndësinë relative të termit në një dokument. Ky kombinim përbën skemën e peshave TFIDF, e cila i cakton termit *t* në dokumentin *d*, peshën:

$$\text{TFIDF}_t = tf \; x \; idf_t$$

Kjo peshë ka karakteristikat e mëposhtme:

- Është e lartë kur *t* ndeshet shumë herë në një numër të vogël dokumentsh, duke pasur fuqi të madhe diskriminuese në këtë grup dokumentash.
- Është e ulët kur *t* ndeshet pak herë në një dokument ose ndeshet në shumë dokumenta, duke mos patur fuqi përcaktuese.
- E ulët kur termi *t* ndeshet pothuajse në të gjithë dokumentat.

Çdo dokument shihet si një vektor me një komponent korrespondues për secilin term në fjalor me peshën TFIDF. Për termat që nuk ndeshen në një dokument, pesha është zero. Për një query q, mund të themi që vlera sasiore treguese e plotësimit të query-t merret si shumë e tf-idf për çdo term të query-t.

$$\text{Pra,} \quad \text{pesha}(\text{q}, \text{d}) \; = \sum_{t \in q} \text{tf} - \text{idf}_{t,d}$$

Paraqitja e një grupi dokumentash si vektorë në një hapësirë vektoriale njihet si *Modeli i Hapësirës Vektoriale* dhe është shumë i rëndësishëm në veprimet e GjIK duke filluar nga renditja e dokumentave sipas një query, klasifikimi i dokumentave dhe grupimi i dokumentave. Query konsiderohet si një dokument i vogël dhe ruhet në të njëjtën hapësirë vektoriale si korpusi i dokumentave. Mbetet që në këtë hapësirë vektoriale të llogaritet ngjashmëria midis dy vektorëve, vektori dokument dhe vektori query. Mënyra standard për të caktuar një vlerë sasiore për

ngjashmërinë midis dy dokumentave është llogaritja e kosinusit të këndit që formojnë vektorët e dokumentave $V(d_1)$ dhe $V(q)$ (angl. cosine similarity). Ngjashmëria ka vlerë 1 kur dy vektorët kanë të njëjtin drejtim ose janë shumë të ngjashëm; ka vlerë 0 kur vektorët janë pingule ose themi të pavarur; ka vlere -1 kur vektorët kanë drejtim të kundërt . Në kontekstin tonë ngjashmëria do të jetë midis 0 dhe 1 sepse peshat tf-idf nuk mund të jenë negative prandaj këndi që formojnë dy vektorët e dokumendave nuk mund të jetë më i madh se 90°.

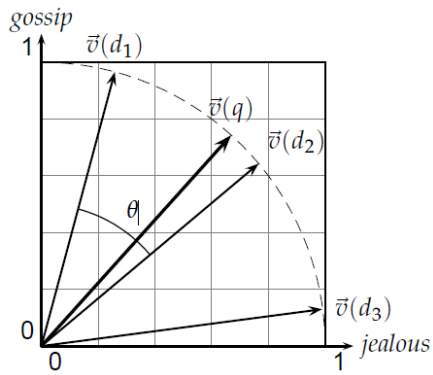

Figura 2.6. Paraqitja e ngjashmërisë së dokumentave nëpërmjet kosinusit të këndit (Manning et al, 2009).

$$ngjashmëria(d1, \mathrm{d2}) = \cos\theta = \frac{\mathrm{V(d1)} \ . \ \mathrm{V(d2)}}{|\mathrm{V(d1)}| \ \ |\mathrm{V(d2)}|}$$

Ku: numëruesi paraqet prodhimin skalar të vektorëve, ndërsa emëruesi paraqet prodhimin vektorial. Vektorët normalizohen në vektorin njësi në mënyrë që dy dokumenta, prerja e të cilëve është njëri prej tyre, të kenë ngjashmërinë maksimale, rrjedhimisht prodhimi skalar dy vektorëve të normalizuar do të japë ngjshmërinë midis dy dokumentave (Manning et al, 2009).

Pamja e koleksioneve të N dokumentave si një bashkësi vektorësh na jep një matrice term-dokument të rendit MxN ku rreshtat paraqesin M termat e N kolonave që paraqesin dokumentat. Ketu duhet të kemi parasysh që termat duhet të jenë proçesuar paraprakisht dhe të jenë kthyer në rrenjën e fjalës në mënyrë që për shembull fjalët "armik" dhe "armiq" të konisderohen si një term i vetëm dhe të shprehen në të njëjtin dimension te matricës (Manning, et al, 2009).

Për të konkretizuar sa thamë më sipër, po marrim një korpus me tre dokumenta të vegjël dhe po zbatojmë modelin e hapësirës vektoriale.

D1: gazeta shqiptare online

D2: kërkim online i shpejtë

D3: informacion nga gazeta shqiptare

Numri i dokumentave N = 3. Disa terma ndodhen vetëm në një dokument ndërsa disa të terë në dy dokumenta. Vlerat *idf* për termat e korpusit, të renditur alfabetikisht janë:

gazeta $log_2(3/2) = 0.584$
informacion $log_2(3/1) = 1.584$
kërkim $log_2(3/1) = 1.584$
online $log_2(3/2) = 0.584$
shpejtë $log_2(3/1) = 1.584$
shqiptare $log_2(3/2) = 0.584$

Për të gjitha termat në korpus llogaristim vlerat e tf si më poshtë:

	gazeta	*informacion*	*kërkim*	*online*	*shpejtë*	*shqiptare*
d1	1	0	0	1	0	1
d2	0	0	1	1	1	0
d3	1	1	0	0	0	1

Duke shumëzuar vlerat *tf* me vlerat *idf* për secilin term, marrim matricen dokument-term si më poshtë:

	gazeta	*informacion*	*kërkim*	*online*	*shpejtë*	*shqiptare*
d1	0.584	0	0	0.584	0	0.584
d2	0	0	1.584	0.584	1.584	0
d3	0.584	1.584	0	0	0	0.584

Supozojmë se jepet query: “gazeta gazeta online”, ateherë llogarisim edhe për të vektorin *TFIDF* dhe më pas gjejmë ngjashmërinë relative të çdo dokumenti në korpus me query-in. Kur llogarisim vlerat e *TFIDF* për termat e query-t pjesëtojmë frekuencën me frekuencën maksimale (2) dhe e shumëzojmë me vlerat e *idf*.

q	(2/2)*0.584=0.584	0	0	(1/2)*0.584=0.292	0	0

Llogarisim gjatësinë (vlerë absolute e vektorit) e secilit dokument dhe të query-t.

Gjatësia e *d1* $= \sqrt{0.584^2 + 0.584^2 + 0.584^2} = 1.011$

Gjatësia e *d2* $= \sqrt{1.584^2 + 0.584^2 + 1.584^2} = 2.316$

Gjatësia e *d3* $= \sqrt{0.584^2 + 1.584^2 + 0.584^2} = 1.786$

Gjatësia e *q* $= \sqrt{0.584^2 + 0.292^2} = 0.652$

Vlera e llogaritur e ngjashmërisë është:

ngjashmëria(d1,q) = (0.584*0.584 + 0*0 + 0*0 + 0.584*0.292 + 0*0 + 0*0)/(1.011*0.652)=0.776

ngjashmëria (d2,q)=(0*0.584 + 0*0 + 1.584*0 + 0.584*0.292 + 1.584*0 + 0*0)/(2.316*0.652)=0.112

ngjashmëria (d3,q)=(0.584*0.584 + 1.584*0 + 0*0 + 0*0.292 + 0*0 + 0.584*0)/(1.786*0.652)=0.292

Në bazë të vlerave të ngjashmërisë, rezultati i plotësimit të query-t i renditur sipas rëndësisë është: *d1, d3, d2*.

2.2.1.3. Probabiliteti

Modelet probabilitare e trajtojnë proçesin e GjIK si një inferencë probabilitare. Ngjashmëritë llogariten si probabiliteti me të cilin një dokument plotëson query-n. Shpesh këto modele përdorin teorema të teorisë së probabilitetit. Renditja e dokumentave sipas rëndësisë që kanë në kërkim bazohet në principet bazë të probabilitetit, pra renditja e dokumentave bëhet mbi bazën e probabilitetit që ai dokument plotëson kushtin me të cilin po kërkohet (Belkin and Croft, 1992). Mbi këtë parim merret në konsideratë që ka pasiguri në paraqitjen e informacionit të kërkuar dhe të dokumentave. Nëqoftëse jepet një query, sistemi i GjIK ka një kuptim jo të qartë të infrmacionit që po kërkohet. Teoria e probabilitetit përmban principet bazë për arsyetimin në kushte pasigurie. Me fjalë të tjera, teoria e probabilitetit formalizon llogaritjen e mundësisë së një dokumenti për të qënë ai i kërkuari. (Manning et al, 2009).

Sistemi i propozuar nga Turtle dhe Croft ne 1991 përdor rrjetat Bayesiane të inferences për të renditur dokumentat duke përdorur probabilitetin me kusht P(Info kërkuar|dokumenti) që një informacion i kërkuar gjendet në një dokument të dhënë. Rrjeti i inferencës konsiston në një graf të orientuar aciklik, ku nyjet paraqesin lidhjet midis kushteve që përfaqësohen nga nyjet. Rrjeti i inferencës është një bashkësi dokumentash, një rrjet konceptual i paraqitjes që përfaqëson fjalorin e indeksimit, dhe rrjeti i query që përfaqëson informacionin e dëshiruar. Rrjeti konceptual i paraqitjes është inferenca midis dokumentit dhe query-t. Për të renditur dokumentat, inicialiozohet rrjeti i inferencës dhe probabilitetet shpërndahen në rrjet deri sa merret probabiliteti i nyjes që përfaqëson informacionin e kërkuar. Këto probabilitete përdoren për të renditur dokumentat.

Modelet statistikore kanë avantazhet e mëposhtme:

1. U sigurojnë përdoruesve një renditje sasiore të dokumentave që kënaqin kërkesën e tyre. Në këtë mënyrë përdoruesit mund të kontrollojnë outputin që u serviret atyre duke vendosur një minimum saktësie, ose duke pëcaktuar numrin e rezultateve që do t'i shfaqet.
2. Query-t bëhen më të thjeshta për t'u shkruar meqënëse shkruhen në gjuhë natyrale.
3. Mund të paraqitet edhe pasiguria që vjen me zgjedhjen e koncepteve të query-t.

Nga ana tjetër modelet statistikore kanë edhe disavantazhet si më poshtë:

1. Kanë fuqi shprehëse të kufizuar. Operatori NOT nuk mund të shprehet sepse peshat janë pozitive. Një rast tjetër është query Boolean ((A and B) or (C and D)) nuk mund të paraqitet me një query të hapësirës vektoriale. Si përfundim mund të themi që, modelet statistikore nuk kanë funqinë shprehëse të modeleve Booleane.
2. Modelet statistikore e kanë të kufizuar strukturën për të paraqitur elementë statistikorë të rëndësishëm si shprehjet. Edhe afërsitë janë të vështira për t'u shprehur, element ky që është shumë i rëndësishëm në kërkimin e avancuar.
3. Llogaritja e probabiliteteve me qëllim renditjen e rezultateve është një veprim i kushtueshëm nga ana llogaritëse.
4. Një listë e renditur i krijon përdoruesit një pamje të kufizuar të hapësirës së informacionit, dhe nuk arrihet të merret drejtpërdrejtë sugjerimi si duhet të modifikohet query nëqoftëse është e nevojshme.
5. Query duhet të përmbajë një numër të madh fjalësh për të rritur performancën e kërkimit. Përdoruesit duhet të përballen me zgjedhjen e fjalëve të përshtatshme që gjenden në dokumentat me interes në kërkim.

- ***Modeli i Pavarësisë Binare***

Ky model i konsideron dokumentat si vektorë binarë, pra mban informacion vetëm për prezencën apo mungesën e termave në një dokument. Termat janë të shpërndarë në mënyrë të pavarur në bashkësinë e dokumentave rezultat. Paraqitja është një bashkësi e renditur variablash booleane. Paraqitja e një dokumenti dhe e query është një vektor me një element boolean për çdo term në konsideratë. Nëqoftëse dokumeti do të paraqitej si vektori d = (x_1, ..., x_m) ku x_t=1 nëqoftëse termi t është prezent në dokumentin d dhe x_t=0 nëse jo. Shumë dokumenta mund të kenë të njëjtën paraqitje vektoriale. Pavarësia tregon që termat në dokument konsiderohen të pavarur nga njëri tjetri dhe nuk modelohet asnjë lidhje midis termave të një dokumenti. Kjo pavarësi është rasti "naiv" i Klasifikatorit Naiv Bayes (angl. Naïve Bayes classifier), ku karakteristikat që implikojnë njëra tjetrën konsiderohen thjeshtë si të pavarura. Kjo parantezë na lejon që ta trajtojmë paraqitjen e modelit si një rast të Modelelit të Hapësirës Vektoriale (Manning et al, 2009).

- ***Modelet e Gjuhës***

Modeli statistikor i gjuhës i cakton probabilitetin një sekuence prej m fjalësh P(w_1, ..., w_m) nëpërmjet shpërndarjes së probabiliteteve. Modelet e gjuhës përdoren në shumë aplikacione te Proçesimit të Gjuhëve Natyrale (angl. natural language processing) si njohja e gjuhës, përkthimi automatik (angl. machine translation), part-of-speech tag, parsimi dhe GjIK. Në njohjen e gjuhës dhe në kompresimin e të dhënave, ky model tenton të kapë karakteristikat e gjuhës dhe të parashikojë fjalën pasuese në sekuencë. Kur përdoret në GjIK, modeli i gjuhës lidhet me një

dokument në një bashkësi. Pasi merr si input query-n q, dokumentat output renditen sipas probabilitetit që modeli i gjuhës do të gjenerojë për termat e query-t $P(q|M_d)$. Modelet unigram janë më të përdorshmet në GjIK sepse janë të mjaftueshëm për të përcaktuar temën e një pjese teksti. Këto modele llogarisin probabilitetin e gjetjes së një fjale të vetme, pa konsideruar influencat nga fjalët para ose pas. Kjo çon në modelin e Çantës së Fjalëve (angl. bag of words) dhe gjeneron një shpërndarje multinormale të fjalëve. Gjetja e probabiliteteve bëhet e vështirë në korpuse të mëdha në të cilat shprehjet ose fjalitë mund të jenë të gjata dhe nuk ndeshen gjatë trajnimit të modelit të gjuhës. Për këtë arsye këto modele përafrohen me modelet N-gram (Manning et al, 2009).

2.3. Marrja e Informacionit të Strukturuar (Information Extraction)

Qëllimi kryesor i sistemeve MIS është zbulimi i informacionit të strukturuar nga tekste pa strukturë ose gjysëm të strukturuar. Për shembull nëqoftëse do të marrim fjalinë:

"Në 1998, Larry Page dhe Sergey Brin themeluan Google Inc".

Nga kjo fjali mund të nxjerrim informacionin e mëposhtëm:

FounderOf(Larry Page, Google Inc.),

FounderOf(Sergey Brin, Google Inc.),

FoundedIn(Google Inc., 1998).

Ky informacion mund të përdoret direkt nga përdoruesi, ose në përgjithësi mund të përdoret nga sistemet kompjuterike si motorrët e kërkimit apo Sistemet e Menaxhimit të Bazave të të Dhënave për të siguruar shërbime më të mira për përdoruesit që kërkojnë informacion.

Sistemet e Marrjes së Informacionit të Strukturuar, MIS, pavarësisht llojeve të tyre, bazohen në faktorët e mëposhtëm:

- Kërkesa për Informacion
- Përgjigja e kërkesës së bërë gjendet në të dhëna të pastrukturuara si tekst apo imazhe.
- Është e pamundur për njerëzit që të proçesojnë kaq shumë informacion
- Edhe një kompjuter nuk mund ta adresojë informacionin e kërkuar nëpërmjet një query sepse ai nuk është i ruajtur në format të strukturuar.

MIS është një disiplinë e Inteligjencës Artificiale, e cila përpiqet t'i gjejë zgjidhje problemeve të këtij lloji. Një sistem MIS përbëhet nga dy komponente të veçanta: Komponenti i trajnimit dhe komponenti i zhvillimit. Në përfundim të trajnimit, sistemi duhet të ketë nxjerrë të gjitha modelet e nevojshme.

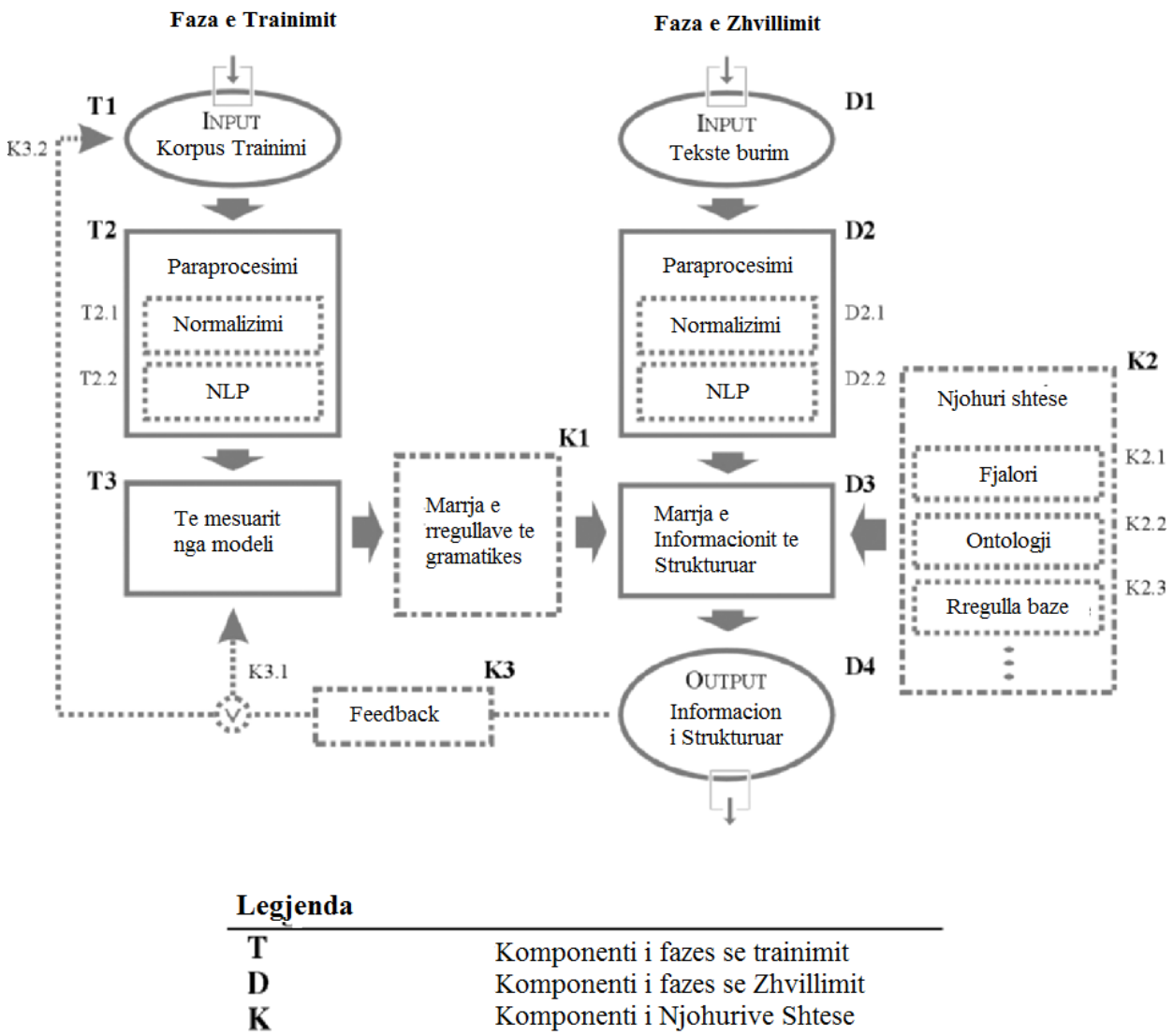

Figura 2.7. Arkitektura e Sistemeve të Marrjes së Informacionit të Strukturuar (Moens, 2006).

Si hap i parë i kësaj faze, duhet të zgjidhet një bashkësi tekst përfaqësuese e detyrës që duhet të kryejë sistemi. Më pas faza e paraproçesimit (T2) merret me normalizimin e karakteristikave formale të tekstit gjatë së cilës realizohet pasurimi i tekstit me metadata linguistike të cilat do të përdoren si parametra në proçesin e marrjes (T2.2). Në këtë pikë përdoren disa mjete të Proçesimit të Gjuhëve Natyrale. Gramatika e nxjerrë nga trajnimi, është një funksion matematik i cili mund të parashikojë klasën për një shembull tjetër. Në komponenten e ndërtimit (D1-4), sistemi i MIS do të identifikojë informacion semantik për tekste të tjera që nuk u përfshinë më parë në trajnim. Faza e paraproçesimit duhet të jetë e ngjashme me atë të përdorur gjatë trajnimit. Më pas teksti input i kalohet fazës së nxjerrjes (D3), e cila përdor gramatikën e zbuluar gjatë trainimit/mësimit (K1). Sistemi i nxjerr elementët tekstualë, i klasifikon dhe i kthen në format të strukturuar (D4).

Disa sisteme më të sofistikuara kanë edhe mekanizmin vetërregullues (K3), në të cilin rezultati përfundimtar përdoret për të ri-trainuar komponentin e të mësuarit në menyrë inkrementuese. Për këtë arsye në sistemet reale komponenti i zhvillimit shpesh quhet Faza e Vlerësimit dhe Testimit. Detyrimisht, veprimi kryesor në sistemet MIS është nxjerrja e informacionit semantik nga tekstet, veprim i cili realizohet në avancë. Lloji i informacionit semantik që nxirret, varet nga njësitë linguistike që do të targetohen në proçesin e nxjerrjes dhe përmbajtja linguistike që mbulohet nga sistemi (Moens, 2006).

2.3.1. Njohja e Entiteteve të Emëruara (angl. Named Entity Recognition)

Një entitet i emëruar është një sekuencë fjalësh që i referohet një entiteti të botës reale, si psh "Shqipëria", "Naim Frashëri" apo "OKB". Qëllimi i Njohjes së Entiteteve të Emëruara – shkurt NEE – është identifikimi i entiteteve të emëruara nga një tekst i thjeshtë dhe klasifikimi i tyrë në tipe të paracaktuara si person, organizatë dhe vend. Kjo nuk mund të realizohet duke krahasuar stringje me një fjalor të indeksuar, sepse entitetet e emëruara për një tip entiteti nuk formojnë bashkësi të fundme, rrjedhimisht çdo përpjekje për të ndërtuar një fjalor entitetesh të indeksuar sipas tipeve do të ishte i paplotë. Për më tepër tipi i entitetit varet nga konteksti ku ai ndodhet. Për shembull, "Nënë Tereza" mund t'i referohet një personi ose një vendi- *aeroporti Nënë Tereza*. Për të përcaktuar tipin e entitetit, në këtë rast duhet të merret për bazë konteksti në dokumentin ku ndodhet ky entitet (Aggarwal and Zhai, 2012).

NEE përbën veprimin bazë në sistemet e MIS. Nxjerrja e strukturave komplekse si relacione dhe ngjarje kërkon saktësi në NEE si fazë të paraproçesimit. Në sistemet që kanë për qëllim përgjigjen e pyetjeve, stringjet e përgjigjes janë entitete të emëruara të cilat duhet që më parë të identifikohen dhe klasifikohen. Entitetet e emëruara më të përdorura janë *person*, *organizate* dhe *vend*. Megjithatë, edhe paraqitja e datës, koha, vlerat monetare dhe përqindjet kanë filluar të përdoren, megjithëse këto shprehje nuk janë entitete të mirëfillta. Përveç tipeve të përgjithshme të entiteteve përdoren edhe tipe të tjera të përcaktuara për domaine dhe aplikacione specifike. Kërkimet më të fundit në NEE përdorin metoda statistikore të Machine Learning (Aggarwal and Zhai, 2012).

2.3.1.1. Modeli i Bazuar në Rregulla (angl. Rule-Based)

Ky model i NEE përcakton në mënyre manuale ose automatike një bashkësi rregullash. Çdo token i tekstit, i cili përfaqësohet nga disa karakteristika të tij, krahasohet me rregullat derisa përcaktohet rregulli i përgjithshëm. Një rregull përbëhet nga një model (angl.pattern) dhe një veprim. Modeli është një shprehje e rregullt mbi modelet e tokens-ave. Veprimi është etiketimi i një sekuence tokens si një entitet ose i disa entiteteve njëkohësisht. Për shembull, për të etiketuar si entitet të tipit *person* çdo sekeuncë me tokens të formës "Znj. Y" ku Y është një fjalë që fillon me shkronjë të madhe, duhet të përcaktohet një rregull i formës:

(token = "Znj." tipi ortografik = Shkronjë e parë e madhe) → person

Krijimi manual i këtyre rregullave të NEE kërkon shumë ekspertizë dhe gjithashtu shumë punë. Për këtë arsye është kushtuar vemendje e veçantë në ndërtimin e metodave automatike të mësimit të rregullave (Aggarwal and Zhai, 2012).

2.3.1.2. Modeli Statistikor i të Mësuarit

Shumica e kërkimit shkencor në NEE bazohet në modelet statistikore të Machine Learning. Pjesa më e madhe e këtyre algoritmeve e trajtojnë këtë si një problem të etiketimit të sekuencave. Etiketimi i sekuencave është një problem shumë i njohur në Machinë Learning dhe përdoret për të modeluar probleme të Proçesimit të Gjuhëve Natyrale (angl. Natural Language Processing), si etiketimi i pjesëve të ligjëratës (angl. partof-speech tagging), ndarjen e tekstit në blloqe (angl. chunking) dhe NEE.

Ky model mund të formulohet si në vijim:

Jepet një sekuencë $x = (x_1, x_2, \ldots, x_n)$ që i korepondon vektorit të karakteristikave. Qëllimi është të etiketohen të gjjithë x_i me një y_i. Etiketa y_i mund të parashikohet duke u bazuar në klasifikimin standard i cili merr për bazë vetëm x_i koresponduese, ose etiketimin sekuencial në të cilin supozohet që etiketa y_i varet jo vetëm nga x_i, por edhe nga vlerat dhe etiketat e tjera të sekuencës. Kjo varësi kufizohet në vlerat dhe etiketat e një bashkësie të ngushtë fqinjësh nga pozicioni korent i. Për të lidhur Njohjen e Entiteteve të Emëruara me problemin e etiketimit sekuencial, çdo fjalë në një fjali trajtohet si një vlerë e veçantë e sekuencës. Klasat e etiketave duhet të identifikojnë kufijtë dhe tipet e entiteteve të emëruara në sekuencë. Paraqitja BIO (Finkel et al, 2004), e cila fillimisht u përdor për ndarjen e tekstit në blloqe (angl. text chunking), përdoret gjerësisht për etiketimin e sekuencave. Për çdo tip entiteti T, krijohen dy etiketa, B-T dhe I-T. Një token i etiketuar me B-T është fillimi i një entiteti të emëruar të tipit T, ndërsa një token i etiketuar me I-T është i brendshëm në një entitet të emëruar të tipit T. Etiketa O përdoret për token-at që janë jashtë (angl. outside) një entiteti të emëruar, (Aggarwal and Zhai, 2012). Më poshtë jepet një shembull i një sekuence NEE.

Steve	Jobs	ishte	themeluesi	i	Apple	Inc.
B-PER	I-PER	O	O	O	B-ORG	I-ORG

Etiketimi sekuencial është zhvilluar në disa drejtime. Po përmendim disa prej tyre.

- **Modelet e Fshehta Markoviane** (angl. Hidden Markov Models): sekuenca më e mirë e etiketave $y = (y_1, y_2, \ldots, y_n)$ për një sekuencë të vrojtuar $x = (x_1, x_2, \ldots, x_n)$, është ajo që maksimizon probabilitetin me kusht $p(y|x)$ ose probabilitetin e bashkimit $p(x, y)$. Në këtë rast supozohet që një proçes Markovian i cili gjeneron një etiketë të re bazuar vetem në një ose

disa etiketa dhe/ose vlera të mëparshme. Nëqoftëse y trajtohet si një gjendje e fshehur, atëherë kemi një model Markovian të fshehur.

- **Modelet Markoviane me Entropi Maksimale** (angl. Maximum Entropy Markov Models): këto modele kombinojnë karakteristikat e modeleve të fshehura markoviane dhe modeleve të entropisë maksimale. MMEM është një model i kushtëzuar i cili supozon që vlerat e panjohura janë pjesë të një zinxhiri Markovian dhe nuk janë të pavarura nga njëra tjetra.
- **Fushat e Rastësishme të Kushtëzuara** (angl. Conditional Random Fields): i përkasin njërit prej modeleve të kushtëzuara. Dallimi kryesor nga modelet MMEM është që një etiketë varet jo vetëm nga etiketat paraardhëse, por gjithashtu nga ato pasardhëse dhe nga konteksti i sekuencës. Modelet e rastësishme të kushtëzuara modelohen si grafe pa drejtim, ndërsa modelet MMEM dhe modelet e fshehura Markoviane modelohen si grafe të drejtuara.

2.3.2. Nxjerrja e Lidhjeve (angl. Relation Extraction)

Nxjerrja e Lidhjeve është një nga detyrat më të rëndësishme në MIS. Kjo fushë studimi ka për qëllim të identifikojë dhe nxjerrë lidhjet semantike midis dy entiteteve në një tekst të dhënë. Lidhjet midis dy entiteteve quhet lidhje binare. Tipet kryesore të këtyre marrëdhënieve janë: fizike personale, sociale ose përkatësie. Disa shembuj lidhjesh binare janë: *ndodhet-në(UPT, Tirana), babai*-i(Abdyl Frashëri, Mithat Frashëri). Nxjerrja e Lidhjeve synon të nxjerrë lidhjet/marrëdhëniet midis entiteteve duke analizuar të gjitha evidencat në tekst dhe jo vetëm të atyre lidhjeve që përmenden shprehimisht. Nxjerrja e lidhjeve realizohet nëpërmjet tre grupeve të metodave: modelet me supervizim, modelet pjesërisht të supervizuara dhe metodat e rendit të lartë.

2.3.2.1. Nxjerrja e Supervizuar e Lidhjeve (angl. Supervised Methods)

Modelet e supervizuara e trajtojnë problemin e Nxjerrjes së Lidhjeve si një problem klasifikimi. Sipas këtij modeli, lidhjet midis entiteteve mund të identifikohen duke u bazuar në lidhjet e vendosura paraprakisht. Në këtë mënyrë vështirësohet identifikimi i një tipi lidhjeje i cili nuk është ndeshur më parë. Një tjetër problem në këtë model është që kërkimi bëhet kryesisht brenda fjalisë, duke bërë të pamundur zgjerimin e lidhjeve në nivel më të gjerë (përtej fjalisë). Shumica e teknikave që i përkasin modeleve të supervizuara kërkojnë një bashkësi të madhe të dhënash për trajnim (Bach and Badaskar, 2007).

2.3.2.1.1. Modelet e bazuara në Karakteristika (angl. Feature-based Classification)

Nëqoftëse jepen shembuj lidhjeje pozitivë dhe negativë, mund të nxirren nga teksti karakteristika sintaksore dhe semantike. Këto karakteristika shërbejnë si shenja për të vendosur nëqoftëse entitetet në një fjali janë ose jo të lidhura. Karakteristikat sintaksore që nxirren nga fjalia janë: 1) vetë entitetet, 2) tipet e të dy entiteteve, 3) sekuenca e fjalëve midis entiteteve, 4)

numri i fjalëve midis entiteteve dhe 5) rruga në pemën e parsimit që përmban dy entitetet. Karakteristikat semantike dhe sintaksore që nxirren i jepen klasifikatorit për trajnim në formën e një vektori karakteristikash. Disa modele që përdoren për identifikimin e tipeve të ndryshme të lidhjeve midis entiteteve janë modelet log-lineare dhe modelet që përdorin SVM (Support Vector Machines) të trajnuara mbi karakteristika duke përdorur kernel polinomial dhe linear. Disa karakteristika janë indikatorë të përshtatshëm për të treguar lidhjet e entiteteve, ndërsa të tjerë nuk janë, prandaj është e rëndësishme të zgjidhen ato karakteristika të cilat kanë peshë në problemin konkret. Metodat e bazuara në karakteristika aplikohen nëpërmjet zgjedhjeve heuristike, ndërsa karakteristikat zgjidhen sipas modelit *trial-and-error* për të maksimizuar performancën. Meqënëse aplikimet në Proçesimin e Gjuhëve Natyrale, (angl. NLP) në përgjithësi, dhe Nxjerrja e Lidhjeve në veçanti kërkojnë një paraqitje të strukturuar të të dhënave input, është e vështirë të merret një nënbashkësi optimale e karakteristikave të përshtatshme. Për të zgjidhur problemin e zgjedhjes së karakteristikave të përshtatshme, janë dizenjuar kernel të specializuar për Nxjerrjen e Lidhjeve që eksplorojnë plotësisht të dhënat input në formën e pemëve të parsimit. Këto kernel eksplorojnë në hollësi paraqitjen e inputit në mënyrë implicite në një hapësirë me shume dimensione. Më poshtë po përmendim disa nga metodat e kernel për Nxjerrjen e Lidhjeve (Bach and Badaskar, 2007).

2.3.2.1.2. Modelet e Bazuara në Kernel

Ky është një drejtim i rëndësishëm i Nxjerrjes së Lidhjeve. Metodat e Bazuara në Kernel realizohen nëpërmjet vektorit të suportit, të prodhimit skalar të dy shembujve, pra krahasohen midis tyre dy instanca dhe merret një rezultat bazuar në ngjashmërinë midis tyre. Niveli i ngjashmërisë bazohet në numrin e substringjeve të ngjashme dhe llogaritet në mënyrë dinamike. Një nga avantazhet e metodave të kernel është që instancat nuk kërkohet të jenë të etiketuara për t'u llogaritur. Tre teknikat më të përdorura të metodave të bazuara në kernel janë: kernel i bazuar në sekuenca, kernel i bazuar në pemë dhe kernel i përbërë, (Bach and Badaskar, 2007).

- **Kernel i bazuar në sekuenca**: kjo metodë bazohet në një graf të drejtuar i cili llogarit rrugën më të shkurtër që ekziston midis dy argumentave. Rrugët janë të ngjashme nëqoftëse gjatësia e tyre është e njëjtë dhe përmban nyje të ngjashme, ku një nyje është një fjalë ose string. Në këtë mënyrë mund të kemi dy rrugë të cilat do të kishin ngjashmëri të ndryshme nga zero nëqoftëse të dyja do të ishin të tipit *Vend -> Folje -> Ndërtesë*, edhe nëqoftëse nuk do të kishin asnjë fjalë të përbashkët. Problemi kryesor në këtë strukturë është që rrugët me gjatësi të ndryshme kanë ngjashmërinë zero. Bunescu dhe Mooney propozuan llogaritjen e ngjashmërisë së sekuencave bazuar në ngjashmërinë e substringjeve të tyre. Çdo nyje në një sekuencë përfaqësohet nga një vektor karakteristikash dhe ngjashmëria midis dy nyjeve është podhimi skalar i vektorëve të tyre. Ngjashmëria midis dy nënsekuencave të të njëjtës gjatësi merret si prodhim i ngjashmërisë së secilit çift të nyjeve në të njëjtin pozicion. Më pas ngjashmëria midis sekuencave llogaritet si shumë e peshuar e ngjashmërive të të gjithë

nënsekuencave të të njëjtës gjatësi nga dy seuencat. Përdorimi i sistemit të peshave minimizon efektin e stringjeve të zakonshme/të shpeshta. Fillimisht ky sistem u përdor për të zbuluar ndërveprimet midis proteinave (Bach and Badaskar, 2007).

- **Kernel i bazuar në pemë**: Argumenti kryesor në ketë model është: nëqoftëse dy pemë parsimi kanë shumë struktura nënpemësh të përbashkëta, atëherë dy instancat janë të ngjashme me njëra tjetrën. Kjo metodë është efiçente aty ku është e vështirë të merren në konsideratë të gjithë elementët brenda një njësie të vetme kerneli. Para se të kombinohen metodat në një kernel të përbërë, duhet të përcaktohen disa argumenta si pathi i kernel dhe pathi i marrëdhënieve . Metodat e Kernel dhe metodat e bazuara në karakteristika janë dy drejtimet kryesore të modeleve të supervizuara, te cilat gjithashtu kanë anët e tyre pozitive dhe negative. Metodat e bazuara në karakteristika nuk kanë kompleksitet të lartë veprimesh, por kërkojnë që karakteristikat që duhet të nxirren të përcaktohen pas analizës së tektit. Karakteristikat 'e mira' gjenden nëpërmjet eksperimenteve. Nga ana tjetër kemi metodat e kernel që kanë kompleksitet të lartë veprimesh por nuk kërkojnë gjenerimin e bashkësisë së karakteristikave për të llogaritur ngjashmëritë mbi një koleksion të madh të dhënash. Modelet e supervizuara kanë disa kufizime në raport me flesibilitetin e zgjerimit të kërkimit në një tip tjetër entiteti relacional. Sasia e veprimeve është gjithashtu problem kur rritet sasia e të dhënave që duhet të proçesohen, prandaj kërkohet që të dhënat të strukturohen në formën e një peme parsimi para se të përdoret si input, (Bach and Badaskar, 2007).
- **Kernel i përbërë**: metodat me supervizim të pjesshëm janënjë model i rëndësishëm në nxjerrjen e lidhjeve. Dallimi kryesor i këtyre metodave nga ato të diskutuara më sipër është që të dhënat input të përdorura për trajnim nuk etiketohen manualisht. Dy metodat kryesore në këtë kategori janë bootstraping dhe supervizimi në distancë. Bootstraping fillon punën me një dataset të vogël instancash duke gjetur modele të lidhjeve. Supervizionimi në distancë funksionon mbi lidhjet e njohura midis instancave ekzistuese (Bach and Badaskar, 2007).

2.3.3. Nxjerrja e Pasupervizuar e Informacionit (angl. Unsupervised Information Extraction)

Njohja e Entiteteve të Emëruara, NEE dhe Nxjerrja e Lidhjeve përdornin tipe lidhjesh dhe entitetesh të mirë përcaktuara në avancë, në varësi të aplikacionit. Për të trajnuar modele të tilla kërkohen sasi të mëdha informacioni të etiketuar dhe indeksuar paraprakisht. Kjo ka kosto të lartë meqënëse kërkon kohë dhe ekspertizë njërëzore. Për të thjeshtuar këtë problem, janë bërë hapa përpara në nxjerrjen pa supervizim të informacionit nga bashkësi të mëdha dokumentash. Më poshtë do të analizojmë drejtimet kryesore të studimit në këtë fushë. Së pari, do të shohim zbulimin e lidhjeve dhe të template-ve, me qëllim zbulimin e tipeve kryesore të lidhjeve dhe templateve në një domain të dhënë duke grupuar entitetet bazuar në karakteristikat kontekstuale leksiko-sintaksore (Aggarwal and Zhai, 2012).

2.3.3.1. Zbulimi i Lidhjeve dhe Nxjerrja e Template-ve

Ndryshe nga Nxjerrja e Supervizuar e Lidhjeve, ku entitetet që kërkohet të zbulohen njihen në avancë, ka raste kur nuk kemi ndonjë tip specifik lidhjeje, por duam të zbulojmë lidhjet më të rëndësishme nga një korpus i dhënë. Për shembull, nëqoftëse na jepen disa artikuj gazetash që flasin për përmbytjet e ndodhura vitet e fundit, do të na interesonte të zbulonim lidhjen "*prek*" e cila jep lidhjen midis përmbytjeve dhe vendeve të përmbytyra.

Shinyama dhe Sekine në 2006, realizuan një studim të cilit iu referuan si Zbulim i Pakufizuar i Lidhjeve. Ata mblodhën shumë artikuj nga gazetat në Web dhe më pas përdorën grupimin e thjeshtë (angl.clustering), bazuar në ngjashmëritë leksikore, për të gjetur artikujt që trajtojnë të njëjtën ngjarje. Në këtë mënyrë ata pasuruan paraqitjen e karakteristikave të një entiteti duke përdorur ndodhjet e shumëfishta në artikuj të ndryshëm. Më pas, ata përdorën parsimin sintaksor për të nxjerrë entitetet e emëruara nga këto artikuj. Çdo entitet i emëruar mund të paraqitej si një bashkësi modelesh sintaksore së bashku me karakteristikat e veta. Për shembull, një model mund të tregojë që entiteti është kryefjalë e foljes "prek", (shembulli i përmbytjeve). Në fund, duke përdorur paraqitjen e karakteristikave, ata grupuan çifte entitetesh që ndodheshin në të njëjtin artikull. Rezultati përfundimtar ishte një tabelë në të cilën rreshtat përfaqësonin artikujt e ndryshëm, ndërsa kolonat i korrespondonin roleve të ndryshme në një relacion (Shinyama and Sekine, 2006).

Rosenfeld dhe Feldman janë marrë me Zbulimin pa Supervizim të Relacioneve në mënyrë më të përgjithshme. Supozohet që inputi i problemit përbëhet nga çifte entitetesh bashkë më kontekstet e tyre. Algoritmi i zbulimit të relacioneve, i grupon këto çifte entitetesh në bashkësi të ndryshme, ku çdo grup paraqit një lidhje të veçantë semantike. Një grup i përgjithshëm përdoret për çiftet e entiteteve pa relacion ose ato me relacione të parëndësishme. Konteksti i çdo çifti entitetesh përmban kontekstin e secilit entitet dhe kontekstin e çiftit. Një çift entitetesh përfaqësohet nga një bashkësi karakteristikash që derivohen nga konteksti. Rosenfeld dhe Feldman morën në konsideratë vetëm karakteristikat e modeleve bazë. Për shembull, "arg_1 bazohet në arg_2", është një model nga i cili mund të nxirret konteksti i përbashkët midis dy entiteteve. Për grupim, u morën në konsideratë grupimet hierarkike dhe grupimi me K-means. Kjo metodë ishtë në gjendje të zbulonte relacione si *QytetNëShtet* dhe *PunësuarNë*, (Rosenfeld and Feldman 2007).

Zbulimi i relacioneve merr në konsideratë vetëm relacionet binare. Do të ishte me shumë vlerë zgjidhja e një detyre më komplekse ku të gjendej një template me hapësira për shumë role semantike. Zgjidhja me e drejtpërdrejtë do të ishte identifikimi i roleve kandidatë dhe më pas grupimi i këtyre kandidatëve në cluster. Ky model i thjeshtuar nuk merr parasysh që një dokument i vetëm mund të mbulojë shumë hapësira rolesh. Për zgjidhjen e këtij problemi, Marx propozoi një algoritëm grupimi me ndërthurje dokumentash për nxjerrje të pasupervizuar të

informacionit. Ky algoritëm cakton një kandidat nga një dokument për një grupim, bazuar në ngjashmëritë e kandidatit me kandidatët e dokumentave të tjerë. Algoritmi i ndan kandidatët e të njëjtit dokument në grupe të ndryshme (Marx et al, 2002).

Këto studime përdorin një model/template të vetëm dhe nuk i etiketojnë hapësirat e gjetura. Chambers dhe Jurafsky dhanë një metodë të plotë për gjetjen e disa templateve nga një korpus dhe vendosjen e etiketave të përshtatshme. Metoda e tyre vepron në dy hapa grupimi, njëri grupon modelet leksikore që përshkruajnë të njëjtin tip ngjarjeje, dhe i dyti grupon rolet kandidate në hapësirat për çdo tip ngjarjeje. Një hapësire etiketohet duke përdorur modelet sintaksore të hapësirave korresponduese. Për shembull, një hapësirë e zbuluar nga kjo metodë, etiketohet automatikisht si "person/organizatë që arreston, merr në pyetje, zbulon, investigon". Një person lehtësisht mund të kuptojë që ky përshkrim i referohet hapësirës "*polic*" (Chambers and Jurafsky, 2011).

2.3.3.2. Nxjerrja e Hapur e Informacionit (angl. Open Information Extraction)

Qëllimi kryesor i Nxjerrjes së Hapur të Informacionit lidhet me gjetjen e fakteve dhe relacioneve më të dobishme nga një korpus heterogjen, si për shembull Web-i (Banko and Etzioni, 2008). Në Nxjerrjen e Hapur të Informacionit nuk merret në konsideratë asnjë tip relacioni paraprak. Algoritmi i realizimit të nxjerrjes së relacioneve në fillim kryen hapin e vizitës së korpusit dhe më pas përpiqet të nxjerrë sa më shumë kandidatë për relacion. Meqënëse nuk janë specifikuar tipe relacionesh paraprake, një pjesë e rezultatit të nxjerrë është një shprehje që përshkruan relacionin e nxjerrë. Me fjalë të tjera, Nxjerrja e Hapur e Relacioneve gjeneron treshen $< arg_1, rel, arg_2 >$. Modeli i nxjerrjes së relacioneve duhet të trajnohet në mënyrë që të nxjerrë relacione arbitrare duke mos përfshirë karakteristika leksikore në model. Për të përmirësuar cilësinë e relacioneve të gjetura mund të përdoren heuristika shtesë si: një relacion në formë shprehjeje duhet të fillojë me një folje, të përfundojë me një parafjalë dhe të jetë sekuencë e vazhdueshme fjalësh në një fjali; ose një shprehje relacionale binare e cila duhet të gjendet në një korpus të madh, e shoqëruar me një numër minimal argumentash të ndryshëm. Përdorimi i këtyre heuristikave shtesë çon në rezultate më efektive (Aggarwal and Zhai, 2012).

3. Gjetja e Fjalëve Kyçe

Fjalët kyçe përkufizohen si një sekuencë prej një ose më shumë fjalësh të cilat japin një përmbledhje të kondensuar të përmbajtjes së një dokumenti. Ato përdoren shpesh për të formuluar query-t e nevojshëm në sistemet e GjIK, meqënëse janë të thjeshtë për t'u lexuar, kuptuar, rishikuar dhe ndarë. Nëpërmjet fjalëve kyçe bëhet më efiçent kërkimi dhe përzgjedhja, meqënëse duke u bazuar në to, përdoruesi mund të vendosë nëqoftëse do ta lexojë ose jo një dokument. Ato mund të përdoren në shumë sisteme të proçesimit të gjuhëve, si kategorizim teksti, nxjerrje informacioni, etj. Fjalët kyçe janë të pavarura nga bashkësia e dokumentave dhe në përgjithësi edhe të pavarura nga gjuha e shkruar në dokumenta (Berry and Kogan, 2010).

3.1. Hyrje

Pavarësisht rëndësisë që kanë fjalët kyçe për analizën, indeksimin dhe kërkimin, shumica e dokumentave të disponueshëm nuk kanë fjalë kyçe të vendosura paraprakisht. Mënyra më e zakonshme e vendosjes së fjalëve kyçe bazohet në zgjedhjen manuale të tyre nga profesionistë të fushës ose nga vetë autori i dokumentit. Për këtë arsye, kërkimi shkencor në këtë fushë fokusohet në gjejtjen e metodave automatike për nxjerrjen e fjalëve kyçe nga dokumentat. Këto fjalë kyçe mund të perdoren më pas si një bashkësi sugjerimesh për profesionistët e fushës që do të zgjedhin manualisht fjalët kyçe ose si përmbledhje e karakteristikave të dokumentit që mund ta bëjnë atë më te aksesueshëm. Pavarësisht rëndësisë që paraqet nxjerrja e fjalëve kyçe në algoritmet për proçesimin e gjuhëve natyrale, ajo është një detyrë e vështirë. Është vënë re qe përputhshmëria e fjalëve kyçe të gjetura nga indeksues profesionistë (veprim manual), varion nga 20 deri në 80%. Megjithëse metodat automatike paraqesin vështirësi në vlerësimin e performancës, ato janë më konsistente se sa teknikat manuale të indeksimit.

Kërkimi në këtë fushë fillimisht u drejtua në vlerësimin e statistikave corpus-oriented për secilën fjalë. Çdo fjale i jepet një vlerë metrike pozitive nëqoftëse nga ana statistikore është diskriminuese në raport me korpusin (Berry and Kogan, 2010). Fjalët kyçe që ndeshen në shumë dokumenta në korpus, nuk konsiderohen si diskriminuese nga ana statistikore, prandaj nuk zgjidhen si fjalë kyçe kandidate.

Metodat korpus-oriented veprojnë në një fjalë të vetme. Ky fakt kufizon njëhsimin e fjalëve si diskriminuese sepse çdo fjalë individuale mund të përdoret në kontekste të shumta. Për të eleminuar këtë problem mund të studiohet nxjerrja e fjalëve kyçe në dokumenta individuale dhe jo të orientuara nga korpusi të cilit ky dokument i përket. Në këtë mënyrë, metodat e orientuara nga dokumentat janë të pavarura nga konteksti, dhe janë të përshtatshme për korpuse në ndryshim të vazhdueshëm, si për shembull, bashkësi abstraktesh teknike apo artikuj lajmesh. Metodat e orientuara nga dokumentat kombinojnë modelet e proçesimit të gjuhëve natyrale për identifikimin

e pjesëve të ligjëratës (angl. part-of-speech tag, POS) të kombinuara me mësim të supervizuar, algoritma të Machine Learning ose metoda statistikore.

Disa çështje që duhen zgjidhur para se sa të përcaktohen fjalët kyçe janë:

- Sa terma/fjalë kyçe duhet të gjenden?
- Cilat terma do të përdoren si fjalë kyçe?
- Fjalët kyçe do të derivohen apo përzgjidhen nga bashkësia e fjalëve të dokumentit?
- Mbi çfarë baze do të renditen fjalët kyçe?
- A duhet të përgjithësohen konceptet?
- Cilat rregulla duhet të ndjekë proçedura e përzgjedhjes?

Për këto pyetje nuk ka përgjigje të prera. Për tre pyetjet e para mund të marrim një përgjigje nga përdoruesi. Në shumicën e aplikimeve përdoruesi nuk përfshihet në indeksimin e fjalëve kyçe, prandaj termat e përzgjedhur duhet të janë sa më të përgjithshëm në mënyrë që të jenë të dobishëm për përdorues të ndryshëm në kohë. Me fjalë të tjera, kërkohet të gjendet bashkësia e termave që përshkruajnë dokumentin, (O'Conor, 1996).

3.2. Metodat Kryesore

Metodat ekzistuese për nxjerrjen automatike të fjalëve kyçe mund të ndahen në katër kategori kryesore:

Modeli statistikor: këto metoda janë shumë të thjeshta dhe nuk kanë nevojë për trajnim paraprak të të dhënave. informacioni statistikor i fjalëve përdoret për identifikimin e fjalëve kyçe në dokument. Disa modele të tilla përdorin informacion statistikor nga N-gram, disa metoda të tjera përdorin frekuencën e fjalëve, TFIDF, co-occurrence të fjalëve, etj, (Zhang, 2008).

Modeli Linguistik: këto modele përdorin elementë linguistikë të fjalëve, kryesisht në nivel fjalish dhe dokumentash. Këto modele përfshijnë analizën leksikore, analizën sintaksore dhe analizën e kontekstit.

Modelet e Machine Learning: nxjerrja e fjalëve kyçe mund të shihet si mësim i supervizuar në Machine Learning. Në këtë rast përdoren fjalët kyçe që i janë caktuar një bashkësie dokumentash për trajnim, për të krijuar një model i cili të aplikohet në gjetjen e fjalëve kyçe për dokumenta të tjerë. Këto modele përdornin Naïve Bayes, Support Vector Machines, etj.

Modele të tjera: Për gjetjen e fjalëve kyçe mund të përdoren metoda të kombinuara të atyre më sipër ose mund të përdoren heuristika si pozicioni, gjatësia, paraqitja e fjalëve, taget html rreth fjalëve, etj.

Nëqoftëse na jepet një bashkësi e madhe dokumentash trajnimi ku për secilin janë caktuar paraprakisht fjalët kyçe përkatëse, teknikat e kategorizimit të tekstit mund të përdoren për të caktuar fjalë kyçe për një dokument të ri. Fjalët kyçe që i caktohen dokumentave të rinj përzgjidhen nga lista e fjalëve kyçe që i janë vendosur dokumentave në bashkësinë e trajnimit. Kjo quhet vendosje fjalësh kyçe me fjalor të kontrolluar. Për çdo fjalë kyçe zgjidhen dokumentat korespondues nga bashkësia e dokumentave trajnues. Për këtë detyrë përdoren teknikat e Machine Learning me anë të të cilave krijohet një klasifikator duke përdorur dokumentat e lidhura me fjalën kyçe si shembuj pozitive dhe të tjerët si shembuj negative. Kur jepet një dokument i ri ai klasifikohet me anë te klasifikatorit të secilës fjalë kyçe. Fjalët kyçe që e klasifikojnë pozitivisht janë fjalë kyçe të dokumentit të ri. Ky proçes njihet si *Caktim Fjalësh-kyçe*, sepse disa fjalë kyçe nga bashkësia e dhënë i caktohen dokumentit të ri.

Një metodë e ndryshme përdoret për të gjetur fjalët kyçe nga dokumenti pa patur më parë një bashkësi fjalësh kyçe të paracaktuar nga dokumentat trajnues. Në këtë rast, listohen të gjitha shprehjet/fjalët që ndeshen në një dokument dhe përdoren heuristika të ndryshme për të përcaktuar fjalët kyçe më përfaqësuese për dokumentin. Meqënëse shumica e fjalëve kyçe janë emra ose shprehje emërore, mund të përdoren teknika sintaksore për t'i identifikuar ato dhe për t'u siguruar që lista e zgjedhur e fjalëve kyçe përmban vetëm shprehje emërore. Heuristikat që përdoren variojnë nga ato më të thjeshtat, si pozicioni i ndeshjes së parë të fjalës në dokument deri në ato më komplekset si frekuenca e fjalës në dokument në raport me frekuencën e saj në dokumentat e tjerë të korpusit. Bashkësia e dokumentave trajnues përdoret për të rregulluar vlerat e parametrave që balancojnë këto faktorë të ndryshëm. Kësaj metode i referohemi si *Nxjerrje Fjalësh-kyçe*.

Duke përdorur metodat e *Caktimit të Fjalëve-kyçe*, të vetmet fjalë kyçe që mund të përdoren për çdo dokument të ri, janë ato që janë përdorur për dokumentat e përdorura për trajnim. Kjo zgjidhje ka avantazhin që të gjitha fjalët kyçe janë të formuluara mirë, por disavantazhin që nuk mund të funksionojë për dokumenta me tematikë të re – te ndryshme nga ajo e korpusit. Për këtë arsye, bashkësia e dokumentave trajnues duhet të jetë mjaftueshëm e madhe për të mbuluar një tematikë sa më te gjerë. Nga ana tjetër, *Nxjerrja e Fjalëve-kyçe* është e hapur në zgjedhjen e fjalëve kyçe nga e gjithë bashkësia e fjalëve të dokumentit. Kjo metodë mund të gjenerojë fjalë kyçe të formuluara. Ajo nuk ka nevojë për bashkësi të mëdha dokumentash trajnues, mëqënëse ato përdoren vetëm për të caktuar parametrat e algoritmit.

Algoritmi për gjetjen e fjalëve kyçe funksionon si në vijim. Mbi dokumentin aplikohen teknikat e tepricave leksikore për të hequr shenjat e pikësimit, lidhëzat dhe fjalët e përgjithshme pa kuptim semantik dhe nxirret një listë me fjalë kandidate për fjalë kyçe. Pas kësaj, për çdo fjalë llogaritet një vlerë skalare që përsonifikon një karakteristikë , si për shembull sa shpesh ndeshet fjala në dokument (e normalizuar me frekuencën në dokumentat e tjerë në korpus), sa shpesh është përdorur si fjalë kyçe në dokumenta të tjerë, nëse ndodhet në titull, nëntituj, abstrakt, etj.

Dokumentat e bashkësisë trajnuese përdoren për të krijuar një model që merr këto karakteristika dhe parashikon nëqoftëse një fjalë kandidate do të shërbejë si fjalë kyçe apo jo. Modeli i gjetur përdoret për të nxjerrë fjalë kyçe të ngjashme nga dokumenta të rinj. Kjo teknikë përdoret gjerësisht për të gjetur fjalët kyçe në dokumentacione teknike dhe përdor skema të përshtatshme të Machine Learning si Naïve Bayes.

Për të ilustruar rëndësinë që kanë teknikat e Machine Learning në zgjidhjen e këtyre problemeve, po ilustrojmë një shembull në figurën e mëposhtme., ku tregohen titujt e tre artikujve shkencorë dhe dy bashkësi fjalësh kyçe për secilin (Frank et al., 1999). Njëra bashkësi ka fjalët kyçe që ka caktuar autori i artikullit, ndërsa tjetra është gjeneruar në menyrë automatike nga teksti i plotë.

Protocols for secure, atomic transaction execution in electronic commerce		Neural multigrid for gauge theories and other disordered systems		Proof nets, garbage, and computations	
anonymity	*atomicity*	disordered systems	disordered	*cut-elimination*	cut
atomicity	*auction*	*gauge fields*	gauge	linear logic	*cut elimination*
auction	customer	*multigrid*	*gauge fields*	*proof nets*	garbage
electronic commerce	*electronic commerce*	neural multigrid	interpolation kernels	sharing graphs	*proof net*
privacy	intruder	neural networks	length scale	typed lambda-calculus	weakening
real-time	merchant		*multigrid*		
security	protocol		smooth		
transaction	*security*				
	third party				
	transaction				

Figura 3.1. Titujt dhe fjalët kyçe të autorit dhe ato të zgjedhura automatikisht për tre artikuj (Frank et al., 1999).

Shprehjet e përbashkëta midis dy bashkësive tregohen me shkronja të pjerrëta. Në secilin rast fjalët kyçe të autorit dhe ato të gjeneruara automatikisht kanë element të përbashkët, megjithatë është e thjeshtë të evidentohet cilat janë fjalët kyçe të caktuara nga autori. Mund të themi që skema e Machine Learning zgjedh fjalë kyçe të mira, por edhe fjalë kyçe që autori nuk do t'i kishte përdorur. Pavarësisht anomalive, lista e gjeneruar e fjalëve kyçe na jep disa karakteristika të arsyeshme të dokumentit. Nëqoftëse, autori nuk do të kishte vendosur fjalë kyçe për artikullin e tij, atëherë ato të gjeneruara automatikisht mund të ishin të dobishme për dikë që kërkon informacion (Witten, 2005).

Fjalët kyçe përdoren edhe për të rritur funksionalitetet të sistemeve te GjIK. Jones dhe Paynter kanë dhënë idenë e përdorimit të një sistemi që liston të gjithë dokumentat e lidhur me bashkësinë e fjalëve kyçe të një dokumenti të dhënë. Gjithashtu, në këtë sistem te cilin e quajtën Phraiser, ata

realizuan paraqitjen e fjalëve kyçe brenda dokumentit në formën e linkeve për të mundësuar lidhjen me dokumentat e lidhuar me atë fjalë kyçe (Jones and Paynter, 2000).

3.3. Përkufizimi dhe Identifikimi i Termave

Në terminologjinë klasike, një term përkufizohet si një shprehje – ose paraqitje – e një koncepti. Kjo nënkupton që për të modeluar problemin e Gjetjes së Fjalëve kyçe duhet të përkufizojmë konceptin dhe të përcaktojmë nëqoftëse një segment teksti përfaqëson ose jo një koncept. Në lidhje me këtë përkufizim do te ndjekim një zgjidhje pragmatike siç propozohet edhe nga (Jacquemin and Borigault, 2003).

Në terminologjinë komputacionale të bazuar në korpus, proçesi i nxjerrjes së termave mund t'i shërbejë qëllimeve të ndryshme si: ndërtimi i ontologjive, indeksimi i dokumentave, validimi i përkthimit, etj. Aplikacione të ndryshme kanë të përbashkët identifikimin e termave nga jo-termat, ose termat specifike të domainit nga termat e përgjithshën të fjalorit (Justeson and Katz, 1995). Proçesi i automatizuar i nxjerrjes së termave pasohet nga një proçes manual i validimit të tyre. Për këtë arsye, outputit të proçedurës së nxjerrjes së termave i referohemi si *termat kandidatë*, disa nga të cilët gjatë fazës së validimit marrin statusin *terma*. Termat kandidatë duhet të jenë në formën e një liste të renditur sipas peshës së secilit term. Në proçesin e validimit dërgohen vetëm termat të cilët kalojnë një vlerë limit të peshës.

Natural Language Processing e bazuar në terma ndahet në katër sub-domaine siç tregohet në Figurën 3.2.

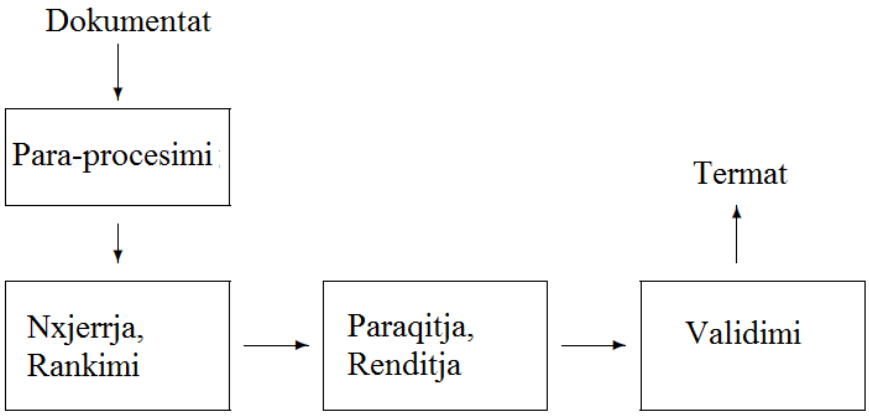

Figura 3.2. Modulet e proçesit të nxjerrjes së termave (Justeson and Katz, 1995).

3.4. Karakteristikat e Termave

Karakteristikat përcaktuese të termave janë të varura nga qëllimi konkret për të cilin ato do të përdoren. Megjithatë ekzistojnë disa karakteristika që mund të shfrytëzohen nga algoritmet për

gjetjen e fjalëve kyçe. Këto karakteristika klasifikohen si linguistike, statistikore dhe të shpërndarjes.

3.4.1. Karakteristikat Linguistike të Termave

Nxjerrja e termave fokusohet në shprehje ose grupe emërore meqënëse shumica e termave janë nominale. Megjithatë, pavarësisht se foljet dhe ndajfoljet marrin më pak vëmendje, edhe ato mund të përcaktojnë domainin. Për një gjuhë të dhënë – ose aplikacion – duam të nxjerrim strukturën linguistike të termave. (Justeson and Katz, 1995) përcaktuan strukturën e termave në anglisht si në (1) ndërsa (Daille, 2000) përcaktoi strukturën e termave ne Frëngjisht si në (2).

(1) ((Adj|Noun)+|((Adj|Noun)_(NounPrep)?)(Adj|Noun)_)Noun

(2) Noun1(Adj|(Prep(Det)?)?Noun2|V Inf)

Zbulimi i rregullave të tilla kërkon që në fazën e paraproçesimit të dokumentave të behet edhe Etiketimi i Pjesëve të Ligjëratës. Pavarësisht se rregulla të tilla do të na ndihmonin të identifikonim termat kandidatë, ato gjithsesi nuk do të eleminonin termat që nuk mund të jenë fjalë kyçe. Vihet re që disa emra dhe ndajfolje janë shumë të pazakontë si fjalë kyçe. Për shembull, *interesante, paraprin, shembull, drejtim*, etj. Këto fjalë klasifikohen në listën negative (angl. stop list). Përveç pjesëve të ligjëratës, edhe relacionet midis tyre janë shumë të rëndësishme. Një term kompleks mund të analizohet si një term kryesor bashkë me një ose me shumë shoqërues. Për shembull, në gjuhën Angleze termi kryesor është emri i fundit që nuk pasohet nga një ndajfolje, si: *database manager* ose *non-European Union national*. Në gjuhën Frënge (2), termi kryesor është emri i parë, pjesa tjeter janë terma shoqërues.

Një aspekt tjetër linguistik është struktura morfologjike. Në gjuhën Suedeze, ndryshe nga Anglishtja, shprehjet emërore shkruhen si një fjalë e vetme, rrjedhimisht duhet të njihen si të tilla dhe të përcaktohet gjithsesi termi kryesor dhe termat shoqërues. Për më tepër fjalët me origjinë latine ose greke ndeshen shpesh në përdorim si fjalë kyçe dhe përmbajnë morfema karakteristike për termat si *meta-, orto-, -graf, -logi*, etj. Rrënjët e fjalëve, ose lemat e fjalëve kanë përdorim të limituar për nxjerrjen e fjalëve kyçe, megjithatë ato janë shumë të dobishme në paraqitje dhe validim sepse reduktojnë numrin e termave kandidatë. Njëlloj si në proçesin e përcaktimit të pjesëve të ligjëratës, edhe identifikimi i rrenjës së fjalëve ka nevojë për paraproçesim linguistik.

3.4.2. Karakteristikat Statistikore të Termave

Karakteristika statistikore bazë e një termi është frekuenca e tij në një korpus. Do të përdoret FA për frekuencën e një njësie ose ngjarjeje A në një korpus dhe N për madhësinë e korpusit. Frekuencat bazë kombinohen për të matur praninë e njëkohshme të termave në korpus. Disa teknika të matjes së ko-okurrencës së fjalëve janë: koefiçienti Dice, Pointwise Mutual Information (PMI) dhe Raporti Log-Likelihood (angl. Log-Likelihood Ratio, LLR), si më poshtë:

$$Dice = \frac{2 \times f_{AB}}{f_A + f_B}$$
$$PMI = \log f_{AB} - (\log f_A + \log f_B)$$
$$LLR = \log L(f_{AB}, f_A, f_B/N) + \log L(f_B - f_{AB}, N - f_A, f_B/N)$$
$$- \log L(f_{AB}, f_A, f_{AB}/f_A) - \log L(f_B - f_{AB}, N - f_A, (f_B - f_{AB})/N - f_A)|$$

AB përdoret për praninë e njëkohshme të *A* dhe *B*. $L(k, n, x)$ është funksioni $x^k(1 - x)^{n-k}$.

Një tjetër informacion statistikor është prania e termave të ndërthurur, pra rasti kur një term kandidat ka prerje jo boshe me nje tjetër term kandidat, më të gjatë ose më të shkurtër. Situata të tilla janë relativisht të shpeshta. Për shembull, janë vënë në dukje shprehjet *floating point* dhe *floating point arithmetic*, të dy terma ndërkohë që bigrami *point arithmetic* nuk është term. Për të tilla raste janë të rëndësishme frekuencat e mëposhtme:

- Frekuencat e termave kandidatë që janë substringje të një tjetër kandidati.
- Frekuenca e termave kandidatë si term kryesor.
- Numrin e termave kandidatë të gjatë që kanë pjesë të tyre një tjetër kandidat.
- Gjatësia e termit kandidat me numër fjalësh.

Këto frekuenca të thjeshta mund të kombinohen në njësi/pesha të termave në menyra të ndryshme. Për shembull, frekuenca e lartë e një kandidati si nënstring i një tjetër kandidati, është tregues pozitiv për kandidatin.

3.4.3. Karakteristikat e Shpërndarjes së Termave

Karakteristikat e shpërndarjes mund të trajtohen si një rast i veçantë i karakteristikave statistikore, meqënëse edhe ato maten me vlera numerike. Ekzistojnë disa karakteristika të shpërndarjes së termave në tekst të cilat janë me interes për problemin e gjetjes së fjalëve kyçe. Një karakteristikë e shpërndarjes është rasti kur termat ndeshen të përqëndura lokalisht dhe kur përsëriten te shoqëruara në grup. Një tjetër karakteristike është shpërndarja e termave në nivel korpusi, përtej nivelit të shpërndarjes brenda dokumentit dhe karakteristika e tretë është shpërndarja në një domain specifik. Një matës shumë i përdorur në Information Retrieval është TFIDF – term frequency, inverse document frequency.

Pesha e termit në një koleksion dokumentash merret si prodhimi *tf* * *idf*. Në Information Retrieval, TFIDF përdoret për të rankuar dokumentat, por mund të shërbejë edhe për të rankuar fjalët dhe sekuencat e fjalëve në një dokument si terma kandidatë për të. Frekuenca e lartë dhe përqëndrimi i lartë i një termi në një dokument tregon qe ai term mund të jetë karakterizues për atë dokument. Fjalët që ndeshen shpesh në një dokument apo grup të vogël dokumentash kanë vlera TFIDF më të larta nga fjalët e zakonshme nëpër tekste si ndajfoljet apo lidhëzat. Megjithëse proçedura formale e TFIDF ndeshet me variacione të vogla në aplikimet e saj, në vija të

përgjithshme ajo funksionon si vijon. Nëqoftëse jepet një koleksion dokumentash *D*, një fjalë *W*, dhe një dokument individual $d \in f$, ne llogarisim:

$$w_d = f_{w,\,d} * \log\,(|D|/f_{w,\,D})$$

Ku $f_{w,\,d}$ është numri i herëve që W ndeshet në *d*, $|D|$ është madhësia e korpusit dhe $f_{w,\,D}$ është numri i dokumentave në *D* në të cilët ndeshet fjala *W* (Salton and Buckley, 1988), (Berger et al. 2000). Për të pasqyruar mbështetjen e këtij modeli mbi baza matematike po marrim disa raste të veçanta në varësi të vlerave të $f_{w,\,d}$, $|D|$ dhe $f_{w,\,D}$.

Supozojmë që $|D| \sim f_{w,\,D}$, pra madhësia e korpusit është afërsisht e barabartë me frekuencën e *W* në *D*. Nëqoftëse, $1 < \log\,(|D|/\,f_{w,\,D}) < c$, për ndonjë konstante *c*, atëherë w_d do të jetë më e vogël se $f_{w,\,d}$ megjithëse pozitive. Kjo presupozon që *W* është e zakonshme të ndeshet në korpusin *D* pavarësisht se mbart një vlerë pozitive të rëndësisë. Ky mund të jetë rasti nëqoftëse TFIDF do të llogaritet për një grup dokumentash me temë: *Të Drejtat e Fëmijëve*. Fjala *fëmijë* do të ishte fjalë e ndeshur shumë shpesh në të gjithë këto dokumenta, prandaj nuk do të ishte përcaktuese për ndonjërin nga dokumentat e këtij korpusi në veçanti. Në të tilla raste, TFIDF do të merrte një vlerë shumë të ulët.

Nga ana tjetër mund të supozojmë që $f_{w,\,d}$ është e madhe dhe $f_{w,\,D}$ është e vogël. Në këtë rast, $\log\,(|D|/\,f_{w,\,D})$ do të kishte vlerë të lartë, po ashtu si w_d . Ky do të përbënte një rast me interes për ne meqënëse fjalët me w_d të lartë janë fjalë me rëndësi në d që ndeshen pak në *D*. Termi *w* thuhet që ka fuqi të lartë diskriminuese në dokumentin *d*.

3.5. Metodat për Gjetjen e Fjalëve Kyçe

Për identifikimin dhe indeksimin e dokumentave nëpërmjet fjalëve kyçe janë përdorur metoda nga më të ndryshmet. Pavarësisht ndryshimeve midis këtyre metodave, shumica e tyre kanë të njëjtat qëllime duke synuar t'i realizojnë ato nëpërmjet përdorimit të heuristikave si: distanca midis fjalëve (frekuenca e fjalëve ose lidhje.relacione të paracaktuara midis fjalëve), përcaktimi i bashkësive të fjalëve që përshtaten me një temë të caktuar ose përshkrimi i informacionit që ndodhet në tekst.

3.5.1. Nxjerrja e Fjalëve Kyçe nga Dokumenta Monogjuhësore

Punimet shkencore të para në nxjerrjen e fjalëve kyçe nga korpuse një-gjuhësore janë të bazuara në karakteristika linguistike dhe kufizohen me shprehjet emërore. Shprehjet emërore mund të identifikohen duke përdorur modelet e pjesëve të ligjëratës, identifikimin e kufijve të shprehjeve ose nëpërmjet rregullave të caktuara gramatikore. Një shembull i tillë është sistemi LEXTER për shprehjet emërore. Termat e përbërë nga një fjalë e vetme identifikohen nëpërmjet gramatikës bazë, për shembull janë analizuar termat një-fjalësh të gjuhës Angleze. Gjithashtu ai

përdori edhe lista me lidhëza për filtrim – fjalë apo morfema që nuk mund të jenë pjesë e një termi (Ananiadou, 1994). Sistemi TERMS kombinon modelet e shprehjeve emërore me listën e lidhëzave dhe një prag frekuence me vlerë default 2 (Justeson and Katz, 1995). Sistemi ACABIT ishte ndër të parët që i rankoi termat kandidatë në bazë të njësive statistikore, duke identifikuar frekuencë absolute dhe raportin loglikelihood si njësi matëse për rankimin e kandidatëve (Daille, 1996).

3.5.1.1. Metoda C-value/NC-value

Në metodën C-value/NC-value përdoren karakteristikat linguistike dhe statistikore për termat e përbërë nga grup fjalësh. C-value është një njësi matëse bazë për termhood të një kandidati a me frekuencë f_a dhe gjatësi $|a|$ i cili i referohet bashkësisë së termave kandidatë që përmbajnë a si pjesë të saj (T_a), dhe kardinalitetin e asaj bashkësie $(|T_a|)$ (Frantzi et al. 1998).

$$C = \log_2 |a| f_a - \frac{1}{|T_a|} \sum_{b \in T_a} f_b$$

Ideja kryesore është të marrim nga vlera e kandidatit bazë bazuar në frekuencën dhe gjatësinë e njësisë, frekuencën mesatare të kandidatëve më të gjatë në të cilët kandidati i dhënë bën pjesë. Nëqoftëse *a* është maksimale, termi i dytë do të jetë zero. C-value aplikohet vetëm për stringjet grupe fjalësh që kanë kaluar filtrin linguistik. Filtri linguistik bazohet në etiketat e pjesëve të ligjëratës dhe në listën e lidhëzave. C-value do të jetë e lartë për stringje pa ndërthurje që kanë frekuencë absolute të lartë në korpusin e dhënë. Nga ana tjetër, kanditatët jo maksimalë që janë pjesë të kandidatëve më të gjatë me frekuencë të lartë do të kenë një C-value të ulët.

Rankimi sipas C-value është faza e parë e metodës. Në fazën e dytë identifikohet bashkësia e fjalëve në kontekstin e termit. Termat nuk karakterizohen vetëm nga struktura e tyre e brendshme dhe frekuenca, por edhe nga ndodhja e tyre në kontekste të caktuara të termit. Çdo fjale në kontekstin e termit i vendoset një peshë e cila varet nga numri i termave me të cilët ajo shoqërohet. Kjo peshë shënohet si: : $\frac{t_w}{T}$ ku t_w është numri i termave që shfaqen së bashku me fjalën dhe *T* është numri total i termave të marra në konsideratë. Peshat kumulative të kontekstit mund të llogariten si:

$$CCW = \sum_{b \in C_a} b_w f_{b/[a]}$$

C_a është bashkësia e fjalëve të kontekstit me peshë pozitive për a, dhe $f_{b/[a]}$ është frekuenca e b si një fjalë e kontekstit për a me peshë b_w. Për të identifikuar fjalët në kontekstin e termit kërkohet validimi i një bashkësie termash kandidatë nga lista e rankuar sipas C-value. Fjalët në kontekstin e termit zgjidhen nga foljet, ndajfoljet dhe emrat që ndodhen në një dritare të vogël rreth termave të validuar. Pas kësaj, rillogaritet rankimi i termave kandidatë sipas C-value, duke marrë në konsideratë për çdo term kandidat NC-value:

$$NC(a) = 0.8C(a) + 0.2CCW(a)$$

NC-value është një modifikim i C-value që zëvendëson disa pjesë të C-value me një vlerësim të bazuar në kontekstin e fjalëve. Peshat 0.8 dhe 0.2 janë caktuar në mënyrë empirike (Frantzi, et al. 1998). Metoda C-value/NC-value u aplikua për herë të parë në një korpus me rekorde mjekësore të patologjive të syrit me rreth 810 000 fjalë. Në këtë korpus, NC-value dhe rezultate me saktësi më të lartë se C-value dhe të dyja së bashku patën saktësi më të lartë se frekuenca absolute.

3.5.1.2. Peshat e Krahasuara (angl. contrastive)

Modeli C-value/NC-value përdor evidencat nga nje domain i vetëm. Nëqoftëse do të kishim edhe një domain tjetër mund të përdornim njësi të tjera matjeje si Peshat e Krahasuara (CW) që marrin në konsideratë shpërndarjen e termave kandidatë brenda dhe jashtë domainit. Po përmendim disa nocione.

- D është domaini target.
- TC është bashkësia e termave kandidatë $\{a_1, a_2, ..., a_n\}$,
- f_{TC} është shuma e frekuencave të të gjithë $a_i \in TC$ në të gjitha domainet
- $f_{a|D}$ është frekuenca e kandidatit a në një domain D
- $f_{a|\neg D}$ është frekuenca e kandidatit a në domainin e krahasuar për D
- për një term kandidat kompleks, a_h është koka e tij dhe M_a bashkësia e operatorëve
- ma është bashkësia e operatorëve të një termi kompleks a që është edhe term kandidat, $m_a = M_a \setminus TC$
- M_{fa} është një faktor modifikimi që aplikohet në operatorët në bashkësinë m_a që përkufizohet më poshtë.

Për terma të thjeshtë peshat e krahasuar përcaktohen si:

$$CW(a) = log f_{a|D} \left(log \frac{f_{TC}}{\sum_J f_{a|J}} \right)$$

Peshat e krahasuara të një termi kandidat kompleks llogariten me anë të frekuencës dhe peshës së krahasuar të kokës së tij.

$$CW(a) = f_{a|D} \times CW(a^h)$$

Për terma të ndryshëm kompleksë që kanë të njëjtën kokë, pesha krahasuese e kokës është e njëjta, prandaj rankimi do të varet nga frekuenca e vetë kandidatit. Për terma të ndryshëm kandidatë me të njëjtën frekuencë brenda domainit, rankun do ta përcaktojnë peshat krahasuese të kokave të tyre. Ky model merr parasysh strukturën linguistike të një termi kompleks, por jo në atë

shkallë që nevojitet. Në këtë model nuk merren parasysh shpërndarja e operatorëve në domainin target dhe në atë krahasues (Wong et al. 2007).

3.5.1.3. ***Peshat Diskriminuese***

Peshat diskriminuese janë prodhim i dy faktorëve që janë: dominanca e domainit (DP) që aplikohet mbi përdorimin e termit brenda domainit dhe tendenca e domainit (DT) që aplikohet në tendencën në prirjen e përdorimit të termit kandidat në domainin target. Frekuenca e lartë brenda domainit nuk është e mjaftueshme si indikator i termhood në domain, ndërsa dominanca dhe tendenca e domainit janë. Për një term kandidat të thjeshtë dominanca e domainit përcaktohet si:

$$\text{DP(a)}= log_{10}(f_{a|D}+10)log_{10}\left(\frac{f_{TC}}{f_{a|D}+f_{a|\neg D}}+10\right)$$

Ky përkufizim garanton qe DP(a) është gjithmonë më e madhe se 1, rritet me rritjen e frekuencave brenda domainit dhe zvogëlohet me frekuenca të larta në domainin krahasues. Për një term kandidat kompleks DP përkufizohet si:

$$DP(a) = log_{10}(f_{a|D}+10)log_{10}DP(a^h)MF_a$$

ku

$$MF_a = log_2\left(\frac{\sum_{m\in m_a} f_{m|D}+1}{\sum_{m\in m_a} f_{m|\neg D}+1}+1\right)$$

Ky përkufizim jep vlera më të mëdha se 1, që rriten me frekuencën brenda domainit të kandidatëve, me DP e kokës, dhe me faktorin e modifikimit. Faktori i modifikimit i jep pikë më të larta termave kompleksë me kokë të paqartë ku operatori ul nivelin e paraqrtësisë. Për shembull H5N1 është sinjal më i fortë për domainin mjekësor sesa për fjalën virus. Faktori i modifikimit është shumë i ngjashëm me njësinë e tendencës së domainit DT(a) që përkufizohet si:

$$DT(a) = log_2\left(\frac{f_{a|D}+1}{f_{a|\neg D}}+1\right)$$

Një term kandidat që ka të njëjtën frekuencë si në domainin target ashtu edhe në atë krahasues do të ketë një vlerë të DT afër vlerës 1 dhe një vlerë dukshëm më të madhe se 1 nëqoftëse frekuenca Brenda domainit është më e madhe se frekuenca krahasuese.

Për të matur termhood DT përdoret si një peshë mbi DP. Kjo njësi e kombinuar është pasha diskriminuese (DW) e një termi kandidat a (Wong et al, 2007).

$$DW(a) = DP(a)DT(a)$$

Në mënyrë që faktorët kontekstualë të ndikojnë njësitë DW modifikohet me një njësi që quhet Kontributi i Rregulluar Kontekstual (angl. Adjusted Contextual Contribution, ACC). Kjo njësi i referohet një bashkësie fjalësh të kontekstit, C_a, dhe përdor një matës ngjashmërie për fjalët në C_a dhe kandidatit a, bazuar në Distancën Google të Normalizuar (angl. Normalized Google Distance, NGD). Ngjashmëria është sim(a, c) = 1 − wNGD(a, c) ku w është një faktor shkallëzimi i distancës. Njësia matëse e termhood (TH) është:

$$TH(a) = DW(a) + ACC(a)$$

Kontributi i faktorëve kontekstualë trajtohet në mënyrë të ngjashme me modelin C-value/NC-value. Impakti i fjalëve të kontekstit është më i vogël për TH se sa për NC-value.

Duke krahasuar sjelljen e tre njësive të sipërpërmendura, NC-value, Peshave të Krahasuara dhe TH në një bashkësi prej 5156 termash kandidatë të gjeneruara nga disa dokumenta mjekësorë, (Wong et al, 2007), u evidentuan ndryshimet e mëposhtme:

- NC-value nuk përdor një korpus krahasues prandaj i rendit kandidatët pa marrë parasysh tendencën e domainit. Renditja ndikohet nga fjalët e një konteksti dhe nga frekuenca e domainit duke renditur në majë termat e thjeshtë dhe termat e shkurtër kompleksë.
- Peshat e krahasuara kanë tendencën t'i rendisin në majë kandidatët me frekuenca të ulëta. Shpesh këto janë terma kandidatë kompleksë që e arrijnë renditjen lart për të kokës dhe të faktit që nuk kanë tendencë të shfaqen në korpusin krahasues. Termat me të njëjtën kokë tentojnë të grupohen bashkë pavarësisht modifikatorit/veprimit.
- TH ul në pozicion kandidatët me tendencë domaini të ulët dhe i jep frekuencë të lartë kandidatëve në majë të renditjes. Kjo është në akordancë me faktin që frekuencat e larta janë tregues i termhood. Fjalët e kontekstit merren në konsideratë megjithatë ky fakt ka më pak rëndësi se NC-value.

NC-value dhe peshat krahasuese kanë standard më të lartë devijimi në raport me TH.

3.5.1.4. Probabiliteti i Termhood

Kjo njësi matëse propozohet si alternative për TH dhe synon të gjejë njësi probabilitare të termhood bazuar në parimet matematike bazë (Wong et al, 2007a). Probabiliteti i termhood (angl. Odds of Termhood) përdor konceptin e rëndësisë së një kandidati në mënyrë analoge me Gjetjen e Informacionit (angl. information retrieval). Në rastin e Nxjerrjes së Termave rëndësinë e termit nuk e përcaktojmë në raport me një query – si në Gjetjen e Informacionit – por në raport me një evidencë të karakteristikave të termit. Evidenca paraqitet sipas një vektori V =< E_1,E_2, ...,E_m >. Ngjarja që një term kandidat *a* ka rëndësi për një domain *d* e quajmë R_1 ndërsa ngjarjen që ky term është me rëndësi për domainin krahasues e quajmë R_2. Ngjarja që një kandidat *a* të

suportohet nga një vektor evidence do ta quajmë A. Në këto kushte, kërkohet të gjendet probabiliteti që A implikon R_1, fakt që shprehet nëpërmjet rregullave Bayesiane si më poshtë:

$$P(R_1|A) = \frac{P(A|R_1)P(R_1)}{P(A)}$$

Probabiliteti që një kandidat a ka rëndësi për d nëqoftëse jepet evidenca e tij është:

$$O = \frac{P(R_1|A)}{1-P(R_1|A)} = \frac{P(A|R_1)P(R_1)}{P(A)(1-P(R_1|A))}$$

Nëqoftëse korpusi krahasues merret si përfaqësues i koleksionit të dokumentave nga domaine ku *a* nuk është me rëndësi, mund të zëvendësohet (1 d P(R1|A)) me P(R2|A) e cila jep:

$$O = \frac{P(R_1|A)}{P(R_2|A)} = \frac{P(A|R_1)P(R_1)}{P(R_2|A)P(A)} = \frac{P(A|R_1)P(R_1)}{P(A|R_2)P(R_2)}$$

Kjo është një njësi matëse e cila mund të përdoret për të renditur termat kandidatë sepse shpreh probabilitetin e rëndësisë. Duke logaritmuar të dy anët marrim:

$$log\frac{P(A|R_1)}{P(A|R_2)} = log\frac{P(R_1|A)}{P(R_2|A)} - log\frac{P(R_1)}{P(R_2)}$$

Në rastin kur nuk kemi evidence, mundësia që *a* të jetë me rëndësi për *d* mund të merret e barabartë me mundësinë që të jetë pa rëndësi. Në këtë rast $P(R_1)/P(R_2)$ merret e barabartë me 1 dhe logaritmi i fundit në shprehjen e mësipërme është zero. Kjo bën probabilitetin e termhood (OT) të jetë:

$$OT(a) = log\frac{P(A|R_1)}{P(A|R_2)}$$

Për të realizuar llogaritjen e OT(a) supozohet pavarësia e elementëve në vektorin e evidences dhe për k = 1, 2 kemi:

$$P(A|R_k) = \prod_i P(E_i|R_k)$$

të cilën ë zëvendësojmë në OT(a) dhe marrim:

$$OT(a) = \sum_i \log\frac{P(E_i|R_1)}{P(E_i|R_2)}$$

Duke patur si të dhëna trajnuese korpusin e një domaini, një korpus krahasues dhe bashkësinë e evidencave/karakteristikave, OT(a) parashikohet për çdo term kandidat. Çdo evidencë individuale E_i përdoret për çdo rezultat S_i, ku:

$$S_i = \frac{P(E_i|R_1)}{P(E_i|R_2)}$$

kështu që

$$OT(a) = \sum_i \log S_i$$

Kjo njësi matëse ka disa avantazhe krahasuar me njësitë e tjera. Së pari, derivohet nga një framework probabilitar me supozime eksplicite. Së dyti, edhe evidencat individuale mund gjithashtu të formulohen në një framework probabilitar të ngjashëm dhe së treti, renditja empirike ka korespondecë të mirë me njësi matëse ad-hoc siç është TH. Për më tepër ajo ka devijim standard më të vogël se TH, rreth një e dyta e mesatares, ndërsa TH ka devijim standard në intervalin 1.5 – 3 herë mesataren (Wong et al, 2007a).

3.5.2. Nxjerrja e Fjalëve Kyçe nga Dokumenta Multigjuhësore

Nxjerrja multigjuhësore e fjalëve kyçe është nxjerrja e termave nga dy ose më shumë gjuhë me qëllim krijimin apo zgjerimin e një burimi të përbashkët psh. bankë dygjuhësore termash apo ontologji e domainit. Termi "nxjerrje multigjuhësore termash" përdoret edhe për metoda pa shumë njohuri të aplikuara në disa gjuhe njëkohësisht pa patur për qëllim të krijojmë një burim të përbashkët (Hippisley et al, 2005). Nxjerrja e termave në rastin multigjuhësor ka qëllime të ndryshme të cilat përcaktojnë kriteret për termhood. Një nga qëllimet mund të jetë suporti i nxjerrjes monogjuhësore të termave. Më pas, aftësia për të realizuar përkthimin përdoret si evidencë për termhood në gjuhën në shqyrtim. Ndryshimi kryesor është fakti që nuk duhet vetëm të nxjirren termat por, edhe të evidentohen termat korespondues midis gjuhëve. Nxjerrja multigjuhësore e termave kërkon që dokumentat të jenë të shkruar në të gjitha gjuhët në shqyrtim. Në rastin më të keq këto dokumenta nuk do të jenë të lidhur me njëri-tjetrin. Nëqoftëse dokumentat kanë versionin origjinal dhe përkthimin ose janë nga përkthime të të njëjtit burim themi që ato janë *paralelë*. Nëqoftëse ato mund të lidhen nëpërmjet temave themi që ato janë të *krahasueshëm*.

Termat kandidatë që kanë për çdo term korespondencën përkatëse në një gjuhë të dytë, kanë indeks të lartë përkthyeshmërie prandaj mund të shoqërohen me njësi termi të një gjuhe tjetër. Kjo karakteristike është shumë e rëndësishme nëqoftëse flasim për mundësinë e ndërtimit të një aplikacioni të pavarur nga gjuha (Spark Jones, 1972).

3.6. Sistemet për Gjetjen e Fjalëve Kyçe

Në fillimet e GjIK lindi nevoja e nxjerrjes së fjalëve kyçe. Që nga ajo kohë janë propozuar shumë sisteme që kryejnë këtë funksion. Disa prej tyre po i përmendim shkurtimisht sipas kronologjisë.

TFIDF 1972: ishte një hap përpara në krahasim më frekuencat e fjalëve. Për të rritur performancën, frekuenca inverse e dokumentit që llogaritet duke përdorur një korpus peshohet me frekuencën e fjalëve (Sparck Jones, 1972).

TERMINO 1991: është një model i bazuar në gramatikë/liguistikë (David and Plante, 1991).

Xtract 1993: kombinimi i një filtri linguistik dhe i një njësie statistikore (Smadja, 1993).

ANA 1995: mjet i pavarur nga gjuha për nxjerrjen automatike të termave (Enguehard and Pantera, 1995).

LEXTER 1996: funksionon si një korpus i etiketuar me anë të modeleve leksiko-sintaksore bazuar në elementë leksikorë dhe kategori sintaksore (Bourigault, 1996).

KeyGraph 1998: model i cili i përket kategorisë pa supervizim dhe bazohet në informacionin statistikor të termave (Ohsawa et al., 1998).

GenEx 1999: Zgjidhje e bazuar në algoritmin gjenetik për të caktuar parametrat për gjetjen e fjalëve kyçe (Turney, 1999).

KEA 1999: një algoritëm për nxjerrjen e fjalëve kyçe i bazuar në skemën Naïve Bayesian të të mësuarit. Performanca u krahasua me GenEx (Witten et al., 1999).

GlossEx 2002: nxjerr në mënyrë automatike elementët e Indeksit nga një bashkësi e pastrukturuar dokumentash (Park et al., 2002).

TextRank 2004: Bazohet në algoritmin PageRank të Google. Është një algoritëm i pasupervizuar për nxjerrjen e fjalëve kyçe që përdor grafin e ko-okurrencave të fjalëve (Mihalcea and Tarau, 2004).

Hulth 2004: përdor Proçesimin e Gjuhëve Natyrale (NLP) dhe Machine Learning për nxjerrjen e fjalëve kyçe (Hulth, 2004).

CollabRank 2008: model i nxjerrjes së shprehjeve kyçe nga një dokument i vetëm duke përdorur ndikimin reciprok të shumë dokumentave brenda kontekstit të një cluster-i. (Wan and Xiao, 2008).

RAKE 2010: metodë për nxjerrjen e fjalëve kyçe nga dokumenta individuale pa supervizim, e pavaruar nga domaini dhe gjuha (Berry and Kogan, 2010).

Në sesionet në vijin po përshkruajmë sistemet kryesore të nxjerrjes së fjalëve kyçe. Duke përcaktuar të dhënat mbi të cilat punon sistemi, autorin dhe kohën e publikimit, një përshkrim të algoritmit që aplikon sistemi dhe një vlerësim të aspekteve kryesore të tij.

3.6.1. TERMINO

Termino ishte një projekt mbi terminologjinë e bazuar në korpus, të zhvilluar dhe financuar nga Office de la Langue Française. Programi TERMINO përbëhet nga disa module që lehtësojnë Nxjerrjen e Termave në Frengjisht. Ekzistojnë disa versione të TERMINO të cilat përmirësojnë versionet e mëparshme. TERMINO është një model i bazuar në gramatikë i cili bazohet në hipotezën e ekzistencës së elementëve leksikorë dhe sintaksorë që mund të përdoren për të indentifikuar fjalët kyçe në një tekst. Teksti input i sistemit mjafton të jetë i koduar në formën ASCII. TERMINO bazohet në përshkrimin e një shprehjeje emërore dhe rrenjë(s/ve) së saj. Shprehjet emërore që nxirren nëpërmjet kësaj gramatike lokale, filtrohen nga një modul i cili krahason dhe vlerëson termat. Gjithashtu, TERMINO është i pajisur me një ambjent interaktiv për krijimin dhe menaxhimin e termave dhe bazohet kryesisht në njohuri linguistike dhe përmbledh tre nënsisteme:

1. *Paraproçesuesi*: filtrimi i tekstit për heqjen e elementëve/karaktereve të formatimit. Veçohen nga teksti fjalët dhe fjalitë dhe identifikohen emrat.
2. *Parsimi dhe Nxjerrja e Termave*: realizohet analiza morfologjike, parsimi i shprehjeve emërore dhe gjenerimi i termave.
3. *Ndërtimi dhe menaxhimi iteraktiv i termave*: një mjet shtesë për të siguruar një ndërfaqe përdoruesi për ndërtimin e bashkësisë së termave.

Nënsistemi më i rëndësishëm për t'u veçuar është Parseri Morfosintaksor i cili përbëhet nga tre module: parseri morfologjik; parseri sintaksor dhe detektori i sinapseve (angl. synapsy detector). Parseri morfologjik ka dy funksionalitete: a) kategorizimin automatik; b) identifikimin e lemave dhe etiketave. Sipas autorëve të TERMINO, 30% e fjalëve në gjuhën Frënge mund të kategorizohen në më shumë se një kategori. Kjo çoi në etiketimin e të gjitha kategorive të mundshme për secilën fjalë. Rrjedhimisht, ka tepricë fjalësh me etiketa të ndryshme. Kategorizimi dhe lematizimi merren duke aplikuar programin LCML, i cili është një parser morfologjik i formave leksikore që mund të lematizojë dhe etiketojë forma të reja leksikore.

Parseri sintaksor është përgjegjës për të eleminuar mospërputhjet (angl.ambiguities) të gjeneruara në fazat e mëparshme. Ai menaxhohet duke ndërtuar strukturën sintaksore të çdo fjalie. Së fundi, detektori i sinapseve (MRSF) zgjedh midis njësive sintaksore nga parseri, ato njësi emërore leksikore që kanë më shumë mundësi të jenë terma. MRSF u krijua posaçërisht për TERMINO dhe bazohet në parimin e ndërtimit të grupeve emërore (David and Plante, 1991). Kuptimi që i jepet sinapsit këtu është ai i një njësie polileksikore që është bërdhamë e një shpjehjeje emërore (NP). Pra, sinapset janë grupe NP, disa nga të cilat do të kthehen në terma, e disa jo. Për më tepër disa prej tyre mund të shërbejnë si "tema" që shërbejnë për të njohur konceptet e ndryshme të tekstit ose për të patur një përmbledhje të temës. Moduli MRSF përfshin

pesë nënmodule: 1) moduli head hunter; 2) moduli i njohjes së zgjerimit; 3) moduli i kategorizimit; 4) gjeneratori i sinapseve dhe 5) moduli i paraqitjes dhe vlerësimit.

Megjithatë, TERMINO ka një bashkësi mjetesh software më të gjerë se kaq dhe përfshin module të ndryshme që lejojnë manipulimin e të dhënave terminologjike. Këto mjete ndihmojnë në përcaktimin nëqoftëse në sinaps është apo jo një term dhe në krijimin e bazës së të dhënave terminologjike. TERMINO njeh midis 70% dhe 74% të termave kompleksë. Disa arsye pse TERMINO nuk njeh rreth 30% të termave janë: ndërprerja e në segmenteve të fjalëve, akronimet dhe emrat e zakonshëm me shkronja kapitale. Për më tepër ka 28% zhurmë, nga e cila 47% vjen për shkak të mark-up të gabuar dhe 53% për shkak të sinapsseve që i përkasin një fjalori të përgjithshëm (Cabré Castellví et al., 2001).

3.6.2. LEXTER

Meqënëse LEXTER është zhvilluar në një ambient që kishte në fokus aplikimin, programi vjen i shoqëruar me një ndërfaqe të pasur grafike për validimin dhe organizimin e termave kandidatë që merren nga korpusi. Dallimi kryesor midis LEXTER dhe TERMINO është teknologjia e përdorur për NLP. Ndërkohë që TERMINO bazohet në modelet gramatikore të bazuara në rregulla të tipit gjenerativ, LEXTER funksionon si një korpus i etiketuar me anë të modeleve leksiko-sintaksore bazuar në elementë leksikorë dhe kategori sintaksore (Jacquemin and Bourigault, 2003).

Përveç validimit dhe strukturimit të një termi, nxjerrja e termave kandidatë nga LEXTER, në vija të përgjthshme, përbëhet nga dy hapa: së pari, nëpërmjet identifikimit të kufinjve të shprehjeve, nxirren shprehjet emërore maksimale dhe së dyti, këto shprehje ndahen në pjesë përbërëse të cilat mund të shërbejnë si terma kandidatë. Shprehjet output quhen jo pa qëllim kandidatë. Në këtë mënyrë theksohet roli njerëzor në validimin pas proçesimit të nxjerrjes automatike.

Shprehjet emërore maksimale. Gramatika e LEXTER përbëhet nga disa shndërrues që quhen rregulla ndarës të cilat tregojnë kufinjtë më të mundshëm të një shprehjeje emërore. Meqënëse disa rregulla janë shumë të buta/tolerante, ato plotesohen me disa kufizime që bazohen në lista përjashtuese. Për shembull, një ndajfolje që pason pjesoren e shkuar (angl. past participle) shërben si kufi, “Valvulat e vendosura në tubat ushqyes” ndahet në “valvulat” dhe “tubat ushqyes”. Lista e përjashtimeve merret automatikisht nga proçedurat e brendshme të të mësuarit të cilat qartësojnë situatat problematike nëpërmjet okurrencave të sakta në korpus.

Shpërbërja në kandidatë. Shprehjet emërore maksimale të gjeneruara në fazën e mëparshme ndahen në shprehje binare të formuar nga koka dhe një zgjerim. Këto copëza nominale së bashku me shprehjet emërore maksimale përbëjnë bashkësinë e termave kandidatë të sistemit. Për shembull, në shprehjen “voltazh i tensionit të lartë”, gjenerohet kandidati “tension i lartë”. Njëlloj

si në rastin e parë, përdoret një proçedure të mësuari duke studiuar korpusin, me anë të së cilës zgjidhet midis disa pozicioneve potenciale për ndarje.

Termat kandidatë organizohen automatikisht në një rrjet të bazuar në elementë leksikorë në pozicione të ngjashme sintaksore. Kjo bashkësi e strukturuar termash kandidatë vendoset në një bazë të dhënash në mënyrë që të validohet dhe modifikohet nga një ekspert (Jacquemin and Bourigault, 2003).

3.6.3. Xtract

Në Xtract, (Smadja, 1993) përdoret kombinimi i një filtri linguistik dhe i një njësie statistikore. Fillimisht në korpus aplikohen mjete statistikore në mënyrë që të merren terma binare të bashkëvendosur. Më pas, këto bashkëvendosje i kalohen për parsim në Cass (Abney, 1990), dhe caktohen etiketat për bashkëvendosjet. Etiketat përcaktojnë natyrën e lidhjes sintaksore midis bashkëvendosjeve.

Fuqia e lidhjeve midis dy fjalëve në Xtract, bazohet në z-score. Kjo njësi jepet nga ekuacioni i mëposhtëm në të cilin *f(W, W')* është frekuenca e ko-okorrencave te *W* dhe *W'* në një dritare me 10 fjalë, ku $\overline{f} = \sum_w f(W, w)$ është frekuenca mesatare, dhe σ është devijimi standard i *f*.

$$z(W, W') = \frac{f(W, W') - \overline{f}}{\sigma}$$

Për çdo fjalë *W*, fjalët *W'* me fuqi të lidhjeve më të vogël se një prag i caktuar eleminohen. Më pas, frekuenca e ko-okorrencave analizohet në raport me pozicionet relative te *W* dhe *W'*. Në fund Xtract gjeneron çiftet *WW'* të tilla që frekuencat e larta arrihen për disa vlera në distancën nga *W* në *W'*.

Xtract nuk fokusohet në terma, por në bashkëvendosjen e përseritur të tyre. Disa nga bashkëvendosjet që merren nga Xtract konsiderone terma kandidatë. Strukturat e tjera quhen lidhje gramatikore dhe i korespondojnë strukturave kompozicionale pa patur status terminologjik, si për shembull "*priten investitorë*", për arsye sintaksore (Jacquemin and Bourigault, 2003).

3.6.4. ANA

ANA, (Enguehard and Pantera, 1995) është një mjet i pavarur nga gjuha për nxjerrjen automatike të termave. Ajo kombinon modulet për gjetjen e stringjeve të përafërta me modulin për gjetjen e fjalëve që ndeshen shpesh bashkë. Ky është një model inkremental në kuptimin që termat e gjetur në një hap të proçesit, shërbejnë si bazë për nxjerrjen e termave në hapin pasardhës. Proçedura inkrementale quhet *moduli i zbulimit* dhe është përgjegjës për gjetjen e termave kandidatë. Kjo pasohet nga *moduli i familjarizimit* në të cilin përcaktohen tre bashkësi të dhënash.

Moduli i familjarizimit përdor tre hapat e mëposhtëm për identifikimin e tre bashkësive të fjalëve.

a) Një lishtë lidhëzash që mban fjalët me frekuencë më të lartë (lidhësat janë ato që ndeshen shumëhpesh në çdo dokument e megjithatë nuk mbartin kuptim semantik).

b) Një bashkësi termash burim (*angl. seed terms*), te cilat vendosen manualisht dhe përbëjnë konceptet themelore në domainin e korpusit.

c) Një bashkësi fjalësh skematike që gjenden brenda ko-okorrencave të termave burim dhe janë kryesisht parafjalë dhe përcaktorë.

Moduli i zbulimit përbëhet nga bashkësia e termave burim nga moduli i familjarizimit dhe shërben si iniciator për proçesin inkremental te zbulimit të termave. Ai bazohet në tre skema të cilat përdorin termat e identifikuar në hapin paraardhës për të nxjerrë terma të rinj.

3.6.5. KeyGraph

Ky është një model i cili i përket kategorisë pa supervizim dhe bazohet në informacionin statistikor të termave (Ohsawa et al., 1998). KeyGraph nuk përdor as etiketat e pjesëve të ligjëratës, as korpuse të mëdha, por bazohet në ko-okorrencat e termave, segmentimin e grafeve dhe cluster-ave. Idea bazë qëndron në gjetjen e cluster-ave kryesorë të një dokumenti dhe supozimin që çdo cluster përmban fjalë kyçe. Outputi i KeyGraph mund të kategorizohet në *ngjarje, ngjarje clusterash, ngjarje vizuale* dhe *mundësi*. Në figurën 3.3 paraqitet autputi i KeyGraph.

Ngjarjet janë elementë që përshkruajnë diçka me rëndësi në domainin e problemit. Shpesh ngjarjet paraqiten si të dhëna tekst që përshkruajnë domainin e problemit, ku fjalët dhe grupet e fjalëve konsiderohen si ngjarje individuale. Ngjarjet në këto clustera ndeshen shpesh të shoqëruara. Për të evidentuar shkallën e shoqërimit të tyre, ngjarjet lidhen me vije të zezë të pandërprerë.

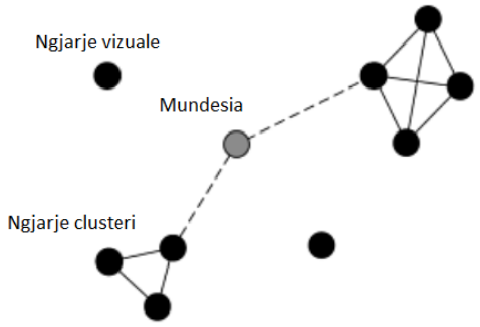

Figura 3.3. Paraqitja e të dhënave nga KeyGraph (Ohsawa et al., 1998).

Ngjarjet vizuale janë ato ngjarje që ndeshen shpesh në domainin e problemit por nuk kanë lidhje të forta me ngjarje të tjera. Për këtë arsye këto ngjarje nuk lidhen me asnjë ngjarje tjetër. *Mundësitë* janë ngjarje që ndeshen rrallë or, që kanë një ko-okorrencë të lartë me *ngjarje* apo ngjarje *clusterash*. Në këtë mëyrë ato kanë potencial për impakt të lartë në sistem, sidomos në qoftë se janë të lidhura me disa clustera. KeyGraph i identifikon *mundësitë* duke i vizatuar me vija të ndërprera.

3.6.6. GenEx

GenEx është një algoritëm gjenetik hibrid për nxjerrjen e fjalëve kyçe, i cili konsiston në dy komponente: Genitor dhe Extractor (Turney, 1999). Extractor aplikohet mbi dokumenta tekst me qëllim që të përcaktohet një bashkësi e peshuar fjalësh kyçe. Shprehjet kyçe kandidate janë të gjitha ato shprehje që përbëhen nga të paktën tre fjalë kyçe të njëpasnjëshme që nuk janë lidhëza. Kandidatëve i gjendet rrënja e fjalës duke i prerë ato në pesë karaktere. Kjo mënyrë është e shpejtë por ka disa disavantazhe. Fjalë me kuptime të ndryshme marrin të njëjtën rrënjë (psh. pjesët dhe pjesëtim), alomorfët nuk merren në konsideratë (psh. i moderuar dhe modest marrin rrënjë të ndryshme), dhe fjalëve të shkurtra nuk u caktohet rrënjë (psh. pikë dhe pikat). Pas kësaj proçedure, çdo kandidat peshohet sipas frekuencës së shumëzuar me pozicionin e saj në tekst. Peshat e shprehjeve kandidate rriten edhe më shumë. Pasi zgjidhet forma e plotë me e shpeshtë për çdo shprehje me rrënjë fjalësh, Extractor nxjerr shprehjet me peshë më të lartë duke u bazuar në 12 parametra si faktori i rritjes për shprehjet e gjata dhe madhësia e shprehjes kyçe përfundimtare. Genitor është një algoritëm gjenetik përdor të dhëna trajnuese për të përcaktuar vlerat më të mira të parametrave. GenEx u vlerësua për 360 artikuj nga domaine të ndryshme nga të cilat 80% u vlerësuan si shprehje kyçe të pranueshme nga vlerësues njerëzore dhe 60% e tyre u vlerësuan si shprehje kyçe të mira për dokumentin.

3.6.7. KEA

Algoritmi i Nxjerrjes së Shprehjeve Kyçe (angl. Keyphrase Extraction Algorithm - KEA), është i ngjashëm me GenEx, përveç teknikave të ndryshme të të mësuarit. Algoritmi KEA përbëhet nga dy faza:

- *Trajnimi*: krijimi i një modeli për identifikimin e shprehjeve kyçe duke përdorur dokumenta trajnues në të cilët njihen fjalët kyçe të autorëve.
- *Nxjerrja*: zgjedhja e shprehjeve kyçe nga një dokument i ri duke përdorur modelin e mësipërm.

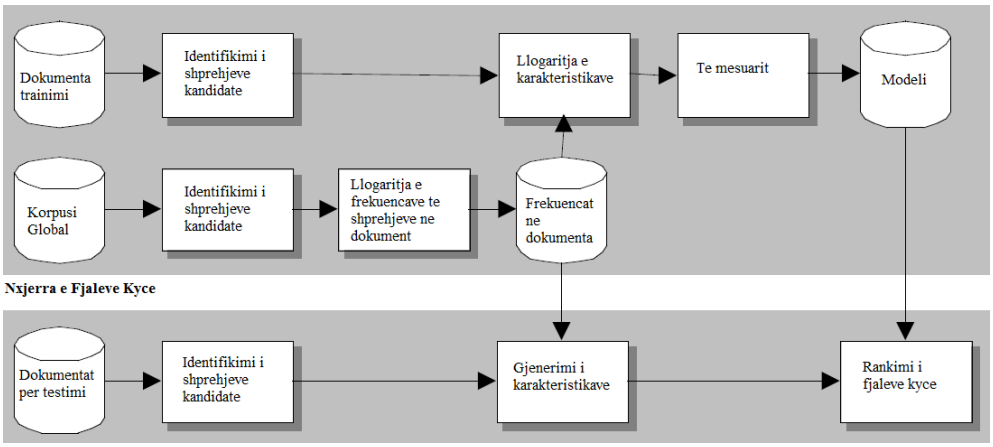

Figura 3.4. Proçesi i trajnimit dhe nxjerrjes së Shprehjeve Kyçe (Witten et al., 1999).

Në fazën e identifikimit të kandidatëve, KEA pëcakton sekuencat tekst sipas kufinjve ortografikë si shenjat e pikësimit, numrat dhe rreshtat e rinj të cilat më pas ndahen në tokens. Pas kësaj, KEA nxjerr shprehjet kandidate të përbëra nga një ose më shumë fjalë, që nuk fillojnë dhe mbarojnë me shenja pikësimi ose lidhëza. Gjatësia minimale dhe maksimale e një shprehje kyçe vendoset nga përdoruesi. Secilit kandidat i gjendet rrënja e fjalës duke përdorur stemerin (Lovins, 1968) dhe versioni me frekuencë më të lartë ruhet për output. Në fazën e filtrimit llogaritet vlera e dy karakteristikave për secilin kandidat: vlera e TFIDF dhe pozicioni i parë ku ndeshet kandidati. Për të analizuar të dhënat trajnuese përdoret një klasifikator Naïve Bayes i cili krijon dy bashkësi peshash: njëra për kandidatët që përputhen me shprehjet kyçe të vendosura manualisht dhe tjetra për të gjithë pjesën tjetër të kandidatëve. Në fazën e filtrimit , probabiliteti i përgjithshëm që një kandidat të jetë shprehje kyçe, llogaritet bazuar në këto pesha. Kandidatët renditen sipas probabiliteteve, dhe shprehjet me probabilitet më të lartë përfshihen në bashkësinë rezultat të shprehjeve kyçe. Pas trajnimit mbi 100 dokumenta dhe testimit me 500, KEA nxjerr 0.9 shprehje kyçe të sakta midis 5 të parëve në renditje. Autorët e KEA nuk raportojnë vlerat e precision dhe recall (Witten et al., 1999).

3.6.8. GlossEx

Elementët indeks të një dokumenti (angl. glossary) janë fjalë ose grupe fjalësh që përfaqësojnë konceptet e domainit në një bashkësi të dhënë dokumentash. Glossex ka për qëllim të nxjerrë në mënyrë automatike elementët e Indeksit nga një bashkësi e pastrukturuar dokumentash. Komponentët përbërës të këtij sistemi janë: identifikimi i elementëve të indeksit, marrja e elementëve të indeksit, filtrimi, agregimi/grumbullimi dhe renditja siç paraqitet në figurën më poshtë:

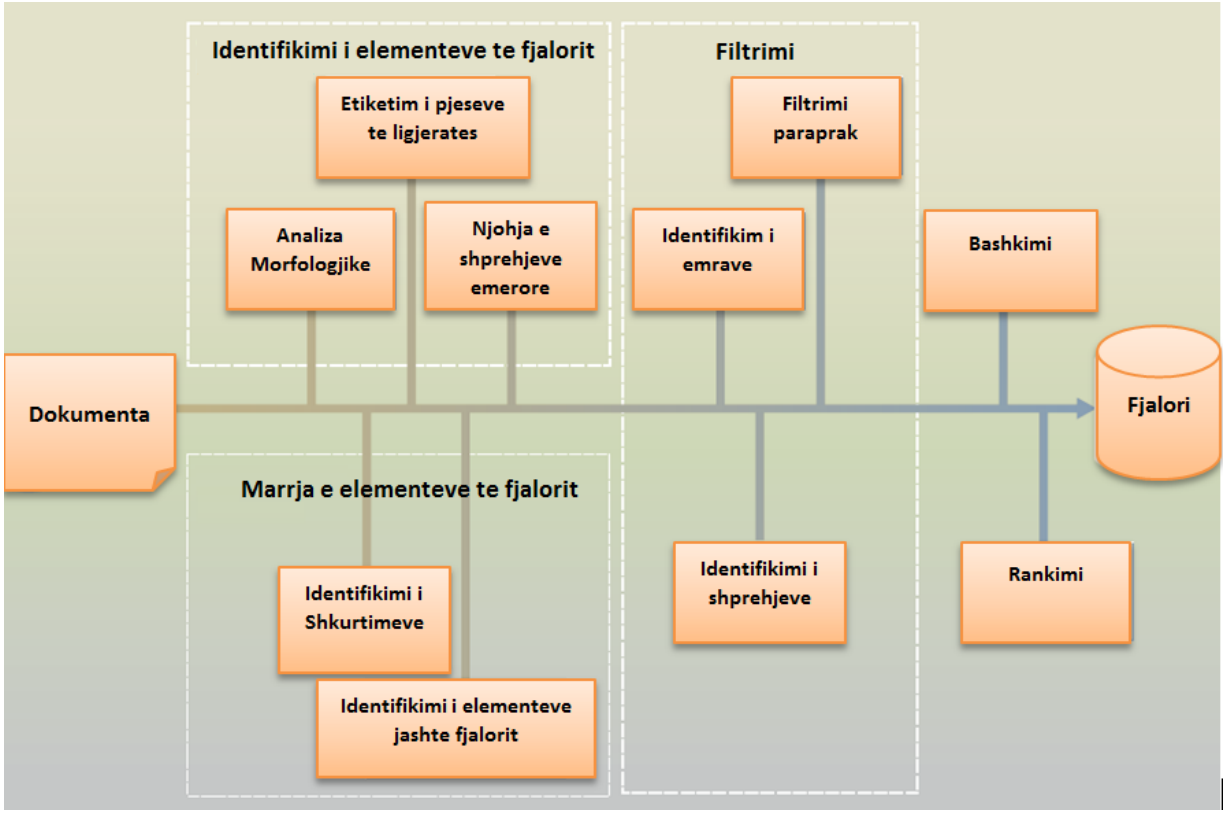

Figura 3.5. Komponentët e Glossex (Park et al., 2002).

Identifikimi i elementëve të indeksit: Ky modul përfshin disa hapa: kërkimin leksikor, analizën morfologjike, etiketimin e pjesëve të ligjëratës (angl. POS tagging) dhe njohjen e shprehjeve emërore. Meqënëse shumica e fjalëve indeks janë shprehje emërore ose folje, kërkohet sipas strukturës së mëposhtme të bazuar në një studim të Justeson dhe Katz dhe në analizën e dokumentave nga ekspertët e domainit (Park et al., 2002).

Noun = *NN*|*NP*|*NNS*|*NPS*

AdjectivalModifier = *JJ* · (*CC* · *JJ*)*

Glossary item = (*DT* · (*VBG*|*VBN*)+ · (*Noun*|*AdjectivalModifier*)* · *Noun*

DT (nga angl. Determiner) shënohen lidhëzat me të cilat zakonisht fillon një grup emëror – psh. the, my, your, his, her, its; our, their, whose, this, that, these, those, which.

VBG dhe VBN përdoren për të treguar format e pashtjelluara të foljeve.

JJ përdoret për ndajfoljen.

CC për lidhëzat.

NN dhe NNS përfaqësojnë emrat e pëgjithshëm në njëjës dhe shumës.

NP dhe NPS përfaqësojnë emrat e përvetshëm në njëjës dhe në shumës.

"|" tregon OR logjik.

"." sekuencë.

"*" përsëritje zero ose më shumë herë.

"+" përsëritje një ose më shumë herë.

Marrja e elementëve të indeksit: Dokumentat dhe në veçanti dokumentat teknikë përmbajnë shumë forma të shkurtuara dhe fjalë që nuk gjenden në një fjalor të dhënë. Për shembull në gjuhën angleze fjalët: *Remanufacturability*, *oxidizability*, dhe *airbreathing* janë shembuj fjalësh që nuk gjenden në fjalor. Shembuj shkurtimesh mund të kemi: *4H (four-wheel drive high), DOHC (double overhead cam)* dhe *NVH (noise, vibration, and harshness).* Identifikimi i shkurtimeve dhe i fjalëve që nuk ndeshen në fjalor është shumë i rëndësishëm për të kuptuar dokumentat teknikë dhe për të nxjerrë informacionin e nevojshëm.

Filtrimi: modulet e mëparshmë gjenerojnë një numër të madh termash kandidatë, shumica e të cilëve nuk duhen marrë në konsideratë. Glossex përdor dy lloje filtrimi të termave: heqja e termave pa rëndësi nga indeksi dhe heqja e rrethanorëve (angl. modifiers) nga shprehjet emërore të cilët nuk janë brenda domainit. Për shembull, rrethanori *remote* është brenda domainit në *remote server*, ndërsa rrethanori *new* është jashtë domainit në *new server*.

Agregimi dhe rankimi: bazohen në dy heuristika. E para është një njësi e cila vlerëson shkallën e përputhshmërisë me domainin – shënohet TD. Njësia e dytë lidhet me kohezionin e termave. Nëqoftëse $|t| = n$ është numri i fjalëve që formojnë termin t, atëhere kohezioni i termit merret nga formula më poshtë:

Ku w_j është fjala e *j-të* në termin *t*.

$$TC_{D_i}(t) = \frac{n.tf_{t,D_i}.\log tf_{t,D_i}}{\sum_{j=0}^{n} tf_{w_j,D_i}}$$

Këto dy njësi kombinohen duke përdorur dy koefiçentë të caktuar nga përdoruesi, α dhe β.

$$GlossEx(t) = \alpha.TD(t) + \beta.TC(t)$$

3.6.9. TextRank

TextRank (Mihalcea and Tarau, 2004) është një model i bazuar në një model renditje sipas grafeve i cili përdoret në aplikacionet e proçesimit të gjuhëve natyrore. Idea e këtij algoritmi lindi nga PageRank që përdoret nga Google për indeksimin e faqeve, analizën e strukturës së linkeve

dhe rrjeteve sociale. Ky model krijon një rrjet ku nyjet janë termat e dokumentit dhe brinjët janë lidhjet midis termave që ndeshen bashkë. Një çift termash quhet që ndeshen bashkë nëqoftëse ato janë 2 deri në 10 terma larg njëra tjetrës në dokument. Lidhjet kanë një peshë që vendoset duke përdorur algoritmin PageRank. Lidhjet janë të padrejtuara dhe simetrike. Fjalët kyçe nxirren duke renditur nyjet dhe duke marrë n të parat prej tyre.

3.6.10. Hulth

Algoritmi i dizenjuar dhe zhvilluar nga Hulth në 2004 është një model i supervizuar i Machine Learning, i cili ka për qëllim ndërtimin e modeleve probabilitare nga dokumenta të cilëve u janë caktuar paraprakisht fjalët kyçe. Përveç kësaj, të dhënat mbi të cilat do të aplikohet ky model duhet më parë të paraproçesohen dhe paraqiten në një formë të përshtatshme për algoritmin. Në rastin e nxjerrjes automatike të fjalëve kyçe paraproçesimi konsiston në dy hapa. Hapi i parë, është nxjerrja e termave kandidatë nga dokumentat dhe hapi i dytë është llogaritja e vlerave të karakteristikave për termat kandidatë. Qëllimi i modelit të të mësuarit është gjetja e atyre karakteristikave që dallojnë termat kandidatë që janë fjalë kyçe nga ato që nuk janë të tilla. Trajnimi i modeleve të parashikimit, përdor vlerat e karakteristikave të fjalëve kyçe të paracaktuara për dokumentat e bashkësisë së trajnimit. Vlerësimi i performancës së modelit të ndërtuar realizohet duke e aplikuar modelin në dokumenta që nuk janë përdorur në gjetjen e modelit. Hulth përdor si sistem referimi të saktësisë fjalët kyçe që janë caktuar manualisht nga indeksuesit profesionistë. Për vlerësimin e saktësisë përdoret metoda precision – sa nga fjalët kyçe të gjetura automatikisht janë fjalë kyçe të vendosura manualisht, dhe recall – sa nga fjalët kyçe të vendosura manualisht janë gjetur nga modeli automatik. Për të gjeneruar termat kandidatë, Hulth krahason për përputhshmëri, n-gram-et e gjetur me anë të parsimit dhe sekuencave të pjesëve të ligjëratës, për të veçuar shprehjet emërore të vlefshme. Rezultatet më të sakta merren me anë të parsimit dhe ato më pak të sakta me anë të nxjerrjes së n-gram-eve. Për filtrim, Hulth veçon nga TFIDF karakteristikën e frekuencës së termave dhe frekuencës inverse të dokumentit, përshtat për to karakteristikën e okurrencës së parë nga KEA, dhe shton një karakteristikë të re që ruan modelin e pjesës së ligjëratës së kandidatit. Disa modele kanë më shumë mundësi të përcaktojnë shprehje kyçe. Eksperimentet që përdorin Naïve Bayes, pemët e vendimmarrjes dhe grupe të tjerë klasifikatorësh, tregojnë që kombinimi i disa modeleve të parashikimit jep rezultate më të mira: F-measure tregon vlera mbi 45%, ndër më të lartat midis metodave të nxjerrjes së shprehjeve kyçe. Megjithatë, precision dhe recall janë llogaritur jo duke u bazuar mbi numrin total të shprehjeve kyçe të caktuara, por në ato që ndodhen në dokumenta. Rrjedhimisht, rezultatet e Hulth-it nuk janë të krahasueshme drejtpërdrejtë me të tjerët (Hulth, 2004).

3.6.11. CollabRank

CollabRank është një model i nxjerrjes së shprehjeve kyçe nga një dokument i vetëm duke përdorur ndikimin reciprok të shumë dokumentave brenda kontekstit të një cluster-i. Ky algoritëm implementohet dukë aplikuar në fillim një algoritëm për clustering dhe më pas duke përdorur një

algoritëm rankimi të bazuar në grafe për nxjerrjen e shprehjeve kyçe të dokumentave të një cluster-i. Zgjidhja i referohet artikujve të gazetave, meqënëse ato janë dokumentat më të përhapur në web dhe në shumicën e rasteve nuk kanë fjalë kyçe të paracaktuara. Algoritmi i këtij modeli jepet më poshtë:

1. **Krijimi i Clusters**: grupimi i dokumentave të bashkësisë D në disa clustera dukë përdorur algoritmat përkatës të ndërtimit të clusterave.

2. **Nxjerrja Bashkëvepruese e Shprehjeve Kyçe**: për çdo cluster C kryhen hapat e mëposhtëm për të nxjerrë shprehje kyçe të një dokumenti që i përket një clusteri në forme batch.

    1) **Vlerësimi i fjalëve në nivel cluster:** ndërtohet një graf afërsie global G bazur në fjalët kandidate të zgjedhura sipas filtrave sintaksorë në dokumentat e një cluster-i të dhënë C, dhe më pas aplikohet algoritmi i rankimit i bazuar në grafe për të llogaritur pikët e afërsisë në nivel cluster-i për secilën fjalë.

    2) **Nxjerrja e Shprehjeve Kyçe në Nivel Cluster-i:** për çdo dokument d në cluster, vlerësohen shprehjet kandidate të dokumentit bazuar në pikët e fjalëve përbërëse të shprehjes. Në fund zgjidhen disa shprehje me pikët më të larta si shprehjet kyçe të dokumentit.

Eksperimentet e kryera në një dataset me 308 artikuj gazetash të plotësuar me shprehje kyçe të caktuara paraprakisht në mënyrë manuale pasqyrojnë efiçensën e modelit CollabRank. Nga ana tjetër vihet re që performanca e nxjerrjes së shprehjeve kyçe është e lidhur ngushtë me cilësinë e kontekstit në cluster. Algoritme të ndryshme clustering gjenerojnë kontekste të ndryshme për nxjerrjen bashkëvepruese të shprehjeve kyçe nga një dokument. Nga eksperimentet e kryera rezulton që vlerat më të mira për F-measure = 0.312 arrihen kur përdoret Kmeans Clustering (Wan and Xiao, 2008).

3.6.12. RAKE

RAKE është shkurtim për Rapid Automatic Keyword Extraction. Kjo është një metodë e pasupervizuar dhe e pavarur nga domaini dhe nga gjuha për nxjerrjen e fjalëve kyçe nga dokumenta të veçuar. Parametrat input të RAKE janë një listë lidhëzash (angl. stoplist), një bashkësi shprehjesh kufizuese dhe një bashkësi fjalësh kufizuese. Këto përdoren për të copëzuar dokumentin në fjalë kyçe kandidate të cilat janë sekuenca të renditura fjalësh. Shoqërimi me njëri-tjetrin i fjalëve kyçe kandidate jep një kuptim të matshëm në identifikimin e shprehjeve kyçe kandidate.

RAKE e fillon nxjerrjen e fjalëve kyçe duke parsuar tekstin në një bashkësi fjalësh kyçe kandidate të vendosura në një listë fjalësh (Figura 3.6) të ndara nga një karakter i dhënë, i cili më

pas konvertohet në një listë me sekuenca fjalësh të njëpasnjëshme duke përdorur shprehjet apo fjalët kufizuese. Fjalët që i përkasin të njëjtës sekuencë u vendoset i njëjti pozicion në tekst dhe së bashku konsiderohen si një fjalë kyçe e vetme.

Compatibility – systems – linear constraints – set – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – minimal set – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – minimal supporting set – solving – systems – systems

Figura 3.6. Fjalë kyçe kanditate pas parsimit

Pasi është identifikuar çdo kandidat dhe është ndërtuar matrica e fqinjësisë (Figura 3.7), llogaritet një peshë për secilën fjalë kyçe kandidate e cila merret si shumë e peshave të fjalëve përbërëse të saj.

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
algorithms	2						1																		
bounds		1																							1
compatibility			2																						
components				1																					
constraints					1								1												
constructing						1																			
corresponding	1						1																		
criteria								2																	
diophantine									1	1			1												
equations									1	1			1												
generating											1			1					1						
inequations												2				1					1				
linear					1				1	1			2												
minimal											1			3				2	1			1			
natural															1		1								
nonstrict												1				1									
numbers															1		1								
set														2				3				1			
sets											1			1					1						
solving																				1					
strict												1									1				
supporting														1				1				1			
system																							1		
systems																								4	
upper		1																							1

Figura 3.7. Matrica e fqinjësisë për fjalët kyçe kandidate (Berry and Kogan, 2010).

Metrikat që përdoren për peshat e fjalëve bazohen në shkallën dhe frekuencën e nyjeve të fjalëve në matricë: (1) frekuenca e fjalëve (*freq(w)*), (2) shkalla e fjalëve (*deg(w)*), dhe (3) raporti i shkallës me frekuencën (*deg(w)/freq(w)*). Peshat e metrikave për fjalët në shembullin e Figurës 3.7 jepen në Figurën 3.8.

Si përfundim, *deg(w)* favorizon fjalët që ndeshen më shpesh në fjalë kyçe të përbëra nga sekuenca të gjata fjalësh; *deg(minimal)* ka peshë më të lartë se *deg(system).* Fjalët që ndeshen shpesh pavarësisht numrit të fjalëve shoqëruese të tyre favorizohen me anë të *freq(w);*

freq(system) ka peshë më të lartë se *freq(minimal)*. Fjalët në të cilat dominon ndodhja në fjalë kyçe me sekuenca të gjata favorizohen nga *deg(w)/freq(w)*;

	algorithms	bounds	compatibility	components	constraints	constructing	corresponding	criteria	diophantine	equations	generating	inequations	linear	minimal	natural	nonstrict	numbers	set	sets	solving	strict	supporting	system	systems	upper
deg(w)	3	2	2	1	2	1	2	2	3	3	3	4	5	8	2	2	2	6	3	1	2	3	1	4	2
freq(w)	2	1	2	1	1	1	1	2	1	1	1	2	2	3	1	1	1	3	1	1	1	1	1	4	1
deg(w) / freq(w)	1.5	2	1	1	2	1	2	1	3	3	3	2	2.5	2.7	2	2	2	2	3	1	2	3	1	1	2

Figura 3.8. Peshat e fjalëve llogaritur nga matrica e fqinjësisë (Berry and Kogan, 2010).

Pasi është realizuar renditja e kandidatëve sipas peshave, konsiderohen si fjalë kyçe T kandidatët e parë, ku T në këtë model merret sa 1/3 e totalit të fjalëve në matricën e fqinjësisë. Në rastin konkret kemi 28 fjalë, prandaj T=9 fjalë kyçe (Berry and Kogan, 2010).

3.7. Metodat e Vlerësimit të Sistemeve të Gjetjes së Fjalëve Kyçe

Performanca e algoritmeve të nxjerrjes së shprehjeve apo fjalëve kyçe matet duke krahasuar output-in e algoritmit me shprehjet kyçe të vendosura manualisht. Kjo njësi bazohet në shkallën e përputhshmërisë midis shprehjeve të gjeneruara automatikisht nga algoritmi dhe shprehjeve të vendosura manualisht nga autori i dokumentit. Në sesionet më poshtë do të përkufizojmë përputhshmërinë në këtë kontekst dhe do të përshkruajmë si llogaritet performanca nëpërmjet përputhshmërisë.

3.7.1. Kriteret për Përputhshmërinë e Shprehjeve

Nëqoftëse autori ka vendosur manualisht shprehjen "*rrjetat neurale*" dhe një algoritëm që nxjerr automatikisht fjalët kyçe gjen "*rrjetë neurale*", ne duam ta konsiderojmë këtë si përputhmëri pavarësisht se njëra shprehje është në numrin shumës dhe tjetra në numrin njëjës. Nga ana tjetër, nëqoftëse autori ka vendosur manualisht shprehjen "*rrjetë neurale*" dhe algoritmi gjeneron termin "*rrjetë*", ne nuk do ta konsiderojmë këtë si përputhmëri meqënëse ka lloje të ndryshme rrjetash në varësi të kontekstit. Do të themi që një shprehje e caktuar manulisht, përputhet me një shprehje të gjeneruar automatikisht nga algoritmi nëqoftëse ato i korespondojnë të njëjtës sekuencë të rrënjëve të fjalëve. Rrënja e fjalës është ajo që mbetet pas heqjes se parashtesës dhe prapashtesës nga një fjalë. Sipas këtij përkufizimi shprehja "rrjetë neurale" përputhet me shprehjen "*rrjetat neurale*" por jo me termin "*rrjetë*". Renditja e fjalëve në sekuencë është shumë e rëndësishme, kështu shprehjet: "*avion shpëtimi*" dhe "*shpëtim avioni*", nuk përputhen. Për të saktësuar kriterin e përputhmërisë duhet të përcaktojmë rregulla shumë strikte dhe formale në lidhje me konvertimin e një fjale në rrënjën e saj (Turney, 1999).

3.7.2. Përcaktimi i Rrënjëve të Fjalëve (angl. stemming)

Dy algoritmet më të përdorur për gjetjen e rrënjëve të fjalëve janë Porter dhe Lovins, të cilët janë implementuar për rastin e fjalëve në gjuhën Angleze. Të dy algoritmet përdorin rregulla heuristike për të hequr ose transformuar prapashtesat e fjalëve (Turney, 1999). Një model tjetër për gjetjen e rrënjëve është përdorimi i fjalorit i cili në mënyrë eksplicite liston rrënjët e çdo fjalë që ndeshet në një tekst të dhënë. Për shkak të kompleksitetit dhe burimeve hardware të nevojshme për ndërtimin e fjalorit, preferohet të përdoren heuristikat në raport me modelin e bazuar në fjalor. Stemeri Lovins është më agresiv në krahasim me stemerin Porter, pra stemeri Lovins mund të shoqërojë dy fjalë me të njëjtën rrënjë, por me probabilitet më të madh gabimi (Krovetz, 1993). Për shembull, stemeri Lovins shoqëron termat "*psikologji*" dhe "*psikolog*" me të njëjtën rrënjë fjale. Nga ana tjetër, stemeri Porter shoqëron termat "*policë*" dhe "*policia*" me dy rrënjë të ndryshme fjalësh "*polic*" dhe "*polici*", ndërsa stemeri Lovins do t'i shoqëronte në mënyrë të pasaktë me "*polic*". Me këtë arsyetim, gjetja agresive e rrënjëve të fjalëve është më efektive se ajo konservative për nxjerrjen e fjalëve kyçe. Një mënyrë më efektive e gjetjes së rrënjëve të fjalëve është aplikimi i iteracioneve mbi stemerin Lovins, pra përsëritet algoritmi i stemerit Lovins derisa fjala nuk ndryshon më. Për shembull, nëqoftëse jepet si input fjala "*incredible*", stemeri Lovins gjeneron si output "*incred*". Më pas jepet si input fjala "*incred*" dhe gjenerohet output "*incr*", i cili kalohet përsëri si output, por mbetet i pandryshuar. Iteracionet në këtë mënyrë rrisin agresivitetin e stemerit (Turney, 1999).

3.7.3. Precision dhe Recall

Problemin e gjetjes së fjalëve kyçe mund ta shohim si një problem klasifikimi, nëqoftëse e shohim dokumentin si një bashkësi fjalësh dhe shprehjesh dhe secilën e klasifikojmë ose jo si pjesë të bashkësisë së fjalëve kyçe. Vlerësimi i një algoritmi që nxjerr fjalët kyçe nga një dokument kërkon të përcaktojë në mënyrë të matshme shkallën në të cilën ky klasifikim i përgjigjet klasifikimit manual njerëzor.

Njësia kryesore matëse për algoritmet e Gjetjes së Informacionit të Kërkuar – Information Retrieval, është Precision dhe Recall.

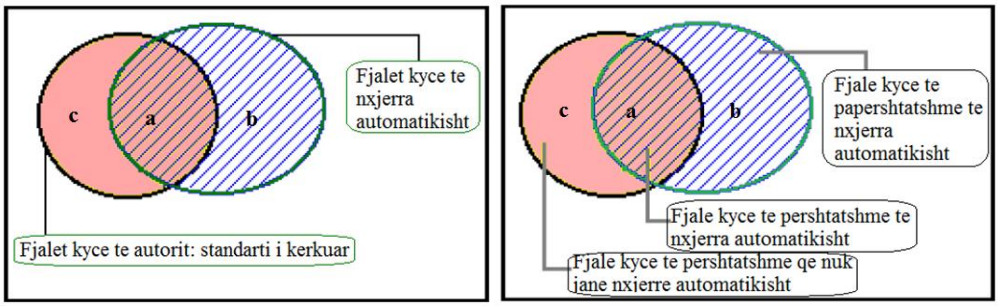

Figura 3.9. Vlerësimi i algoritmit të nxjerrjes së fjalëve kyçe

Siç tregohet në Figurën 3.9. është e rëndësishme të krahasohet rezultati i gjetur me rezultatin e konsideruar të saktë dhe të llogaritet përputhshmëria midis tyre. Rezultati i aplikimit të një algoritmi për nxjerrjen e fjalëve kyçe nga një corpus mund të përmblidhet me anë të figurës së mëposhtme.

	Klasifikohet manualisht si fjale kyce	Klasifikohet manualisht JO si fjale kyce
Klasifikohet si fjale kyce nga algoritmi	*a*	*b*
Klasifikohet JO si fjale kyce nga algoritmi	*c*	*d*

Figura 3.10. Matrica e klasifikimit të fjalëve kyçe

Variabla *a* përfaqëson numrin e herëve që një fjalë kyçe e gjeneruar manualisht nga autori përputhet me një fjalë kyçe të gjeneruar automatikisht nga algoritmi. Variabli *d* përfaqëson numrin e herëve që autori dhe algoritmi pohojnë që një term nuk është fjalë kyçe. Variablat *b* dhe *c* përfaqësojnë numrin e herëve që klasifikimi manual dhe automatik nuk japin të njëjtin rezultat për klasifikimin e termit.

Nëqoftëse konsiderojmë edhe njëherë shprehjet "*rrjetë neurale*" dhe "rrjetat neurale", të cilat përputhen me shprehjen "*rrjet neural*", atëherë themi që proçesi i klasifikimit realizohet për rrënjën e fjalës. Meqënëse kërkohet një numër/vlerë e vetme që të tregojë përformancën e një algoritmi të nxjerrjes së fjalëve kyçe, përdoret një funksion i *a, b, c* dhe *d*. Një funksion i tillë jep saktësinë si më poshtë:

$$accuracy = \frac{a+d}{a+b+c+d}$$

Një problem që lind nga përdorimi i këtij funksioni është se për shkak të shpërndarjes së pabarabartë të shembujve – shumica e termave nuk janë fjalë kyçe, marrim vlera të larta vetëm nga gjetja e klasës së shembujve negativë. Në këtë rast, nëqoftëse një algoritëm i gjetjes së fjalëve kyçe do të gjeneronte bashkësi boshe të fjalëve kyçe, atëherë saktësia do të rezultonte me vlere mbi 99% (Turney, 1999).

Kërkuesit shkencorë në fushën e GjIK (angl. Information Retrieval) e përdorin precision dhe recall për të vlerësuar performancën e një motorri kërkimi. Nëqoftëse problemin e nxjerrjes së fjalëve kyçe nga një dokument e shohim si problem të kërkimit të disa fjalëve në bashkësinë e fjalëve të dokumentit, atëhere jemi në të njëjtat kushte për përdorimin e Precision dhe Recall.

$$\text{PRECISION} = \frac{a}{a+b} * 100\% \qquad \text{dhe} \qquad \text{RECALL} = \frac{a}{a+c} * 100\%$$

PRECISION është raporti i numrit të fjalëve kyçe të përshtatshme që janë nxjerrë nga algoritmi automatik, me numrin total të fjalëve kyçe të nxjerra nga algoritmi automatik (të përshtatshme dhe jo të përshtatshme).

RECALL është raporti i numrit të fjalëve kyçe të përshtatshme që janë nxjerrë nga algoritmi automatik, me numrin total të fjalëve kyçe të cilësuara manualisht si të përshtatshme. Zakonisht kjo vlerë paraqitet si përqindje.

Secili nga këto dy vlera, precision dhe recall lidhen me njëra tjetrën sipas një raporti të zhdrejtë. Për shembull, nëqoftëse ne themi që e gjithë bashkësia e dokumentave është e dobishme, atëhere vlera e recall do të jetë 100%, por vlera e precision do të jetë 0%. Nga ana tjetër, nëqoftëse zgjedhim një dokument për të cilin jemi të sigurtë që është i dobishëm dhe themi që vetëm ai është dokument i dobishëm, atëherë precision do të jetë 100%, por vlera e recall do të jetë afërsisht 0%. Ajo që na intereson është një njësi e performancës e cila jep rezultat të lartë vetëm kur precision dhe recall janë në vlera të balancuara (Turney, 1999).

Në eksperimentin në vijim do të përdorim dy metoda për të matur performancën. Të dyja këto metoda përdoren gjerësisht në kërkimin shkencor në fushën e GjIK. (1) Analizojmë vlerën e precision me disa rregullime për numrin e fjalëve kyçe të gjeneruar automatikisht. Kjo metodë tregon si varion performanca kur përdoruesi ndryshon numrin e dëshiruar të fjalëve kyçe. (2) Përdorim F-measure (van Rijsbergen, 1979), (Lewis, 1995):

$$F\text{-}measure = \frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}} = \frac{2a}{2a+b+c}$$

Kur vlerat e precision dhe recall janë afërsisht të barabarta, F-measure është afërsisht sa mesatarja e precision dhe recall. Kur precision dhe recall nuk janë të balancuara, F-measure është më e vogël se mesatarja e tyre.

4. Gramatika Formale e Shqipes

Ky kapitull përmban një vështrim të përgjithshëm mbi gjuhën Shqipe. Gjuha Shqipe është një degë unike e gjuhëve Indo-Europiane, ku bëjnë pjesë disa gjuhë të tjera siç është gjuha Gjermane, Latine, Greke dhe Sllave. Gramatika e gjuhës Shqipe ndahet në katër pjesë: fonetika, leksikologjia, sintaksa dhe morfologjia.

Gjatë zhvillimit të këtij kapitulli do të merremi me sintaksën dhe morfologjinë e gjuhës Shqipe në mënyrë që të kuptojmë më mirë strukturën e saj dhe të jemi në gjendje të identifikojmë modelet që do na shërbejnë për nxjerrjen fjalëve kyçe. Burimi kryesor i këtyre njohurive është fjalori i gjuhës Shqipe dhe gramatika e saj (Agalliu et al., 2002).

4.1. Morfologjia e Gjuhës Shqipe

Morfologjia është pjesa e gramatikës që studion format e ndryshme të fjalëve, kuptimin e këtyre formave, klasifikimin e fjalëve në pjesë ligjërate, si dhe formimin e fjalëve. Pjesët e ndryshme të ligjëratës janë: emrat, mbiemrat, përemrat, foljet, ndajfoljet, numëruesit, lidhëzat, parafjalët dhe pjesëzat/pjesoret, pasthirrmat. Disa prej tyre janë të ndryshueshme, si: emri, mbiemri, përemri, folja, numërori, sepse e ndryshojnë formën e tyre gjatë përdorimit në fjali; kurse të tjerat janë të pandryshueshme, si: ndajfolja, lidhëza, pjesëza, parafjala, pasthirrma, sepse nuk ndryshojnë formën e tyre gjatë përdorimit në fjali.

4.1.1. Emrat

Emri është pjesë e ndryshueshme e ligjëratës, që emërton diçka.

4.1.1.1. Kategoritë Gramatikore të Emrit

Kategoritë kryesore gramatikore janë: gjinia, trajta, numri, rasa dhe lakimi. Emrat ndahen në tre gjini: mashkullore, femërore dhe asnjëanëse. Ekzistojnë dy trajta të emrit që janë: e shquar dhe e pashquar. Emrat në formën e shquar përdoren për të kryer funksionin e kryefjalës në një fjali. Emrat në formën e pashquar marrin përpara tyre një nga fjalët ‘një’, ‘ca-disa’, bazuar kjo në faktin nëse emri është në numrin njëjës apo shumës (Agalliu et al., 2002).

Rasat që një emër mund të marrë janë pesë dhe janë si vijon:

Rasa emërore, rasa gjinore, rasa dhanore, rasa kallëzore dhe rasa rrjedhore. Emrat në rasën emërore kryesisht kryejnë detyrën e subjektit. Emrat në rasën gjinore përdoren për të treguar lidhje ose përkatësi, dhe emrat në këtë rasë shoqërohen gjithmonë me parashtesat *‘i’, ‘e’, ‘të’*, “*së*” që vendosen përpara tyre. Emrat në rasën dhanore shërbejnë si objekt indirekt në një fjali, në shumicën e rasteve ata shoqërohen me një parashtesë. Rasa kallëzore përdoret kryesisht për të formuar objektin në një fjali. Emri në këtë rast është objekti mbi të cilin bie veprimi i shprehur

nga folja. Rasa rrjedhore përdoret për emrat që shprehin menyrë, cilësi ose prejardhje. Emrat në këtë rasë shoqërohen gjithmonë me një parashtesë si: '*prej'*, '*pas'*, '*para'*.

Për sa i përket lakimit, emrat ndahen në tre grupe. Në grupin e parë futen emrat që mbarojnë me '*i*', në grupin e dytë futen emrat që mbarojnë me '*u*' dhe në grupin e tretë emrat që mbarojnë me '*a-ja'*. Lakimi i parë dhe i dytë përdoren për emrat në gjininë mashkullore, kurse lakimi i tretë përdoret për emrat në gjinine femërore.

4.1.1.2. Llojet e Emrave

Gramatika dallon llojet e ndryshme të emrave të cilët ndahen si më poshtë vijon: të përgjithshëm dhe të përveçëm, frymorë dhe jo frymorë, konkretë dhe abstraktë, të numërueshëm dhe të panumërueshëm, përmbledhës dhe të lëndës.

Emra të përgjithshëm janë emrat që tregojnë qënie ose sende ne menyrë të përgjithshme si: *djalë, vajzë, mal, kukull* etj. Emra të përveçëm janë emrat që tregojnë qënie ose sende të dalluar nga të tjerët që i përkasin të njëjtës klasë si për shembull: *Korabi (mal), Erzeni (lumë),* etj. Emrat e përveçëm shkruhen gjithmonë me shkronjë të madhe.

Emrat frymorë janë emrat që tregojnë qënie të gjalla si: *vajzë, djalë, mace, luan*, etj. Emrat jofrymorë janë emrat që tregojnë sende jo të gjalla: *bankë, laps, libër, tabelë,* etj.

Emrat konkretë janë emrat që tregojnë qenie ose sende që mund të perceptohen nëpërmjet të parit, të dëgjuarit, të prekurit, të nuhaturit ose të shijuarit si: *djalë, lule, zhurmë, parfum,* etj. Emrat abstraktë janë emrat që tregojnë ide ose mendime që nuk mund të perceptohen me anë të shqisave si: *besim, dashuri, zemërim*, etj.

Emrat e numërueshëm janë emrat që tregojnë sende ose qënie që mund të numërohen si: *pulë, fshatar, mace,* etj. Emrat e panumërueshëm janë emrat që tregojnë sende që paraqiten si një masë e padallueshmë, si një lëndë e pandashme si: *oksigjen, hidrogjen, magnez*, etj.

Emrat përmbledhës janë emrat që shënojnë nje grup personash ose sende të një lloji si: *bota, vegjëlia, rinia, pleqëria, njerëzimi,* etj. Emrat e lëndës janë emrat që emërtojnë një lëndë homogjene që nuk mund të numërohet. Ata emërtojnë prodhime ushqimore, kultura bujqësore, prodhime kimike, etj si: *gjalpë, qumësht, kos, grurë*, etj.

4.1.2. Mbiemri

Mbiemri është një pjesë e ndryshueshme e ligjëratës që tregon një cilësi apo karakteristikë të emrit. Mbiemri përshtatet në numër, në gjini dhe në rasë me emrin me të cilin ai lidhet. Mbiemrat ndahen në dy grupe: mbiemra të nyjshëm dhe të panyjshëm. Nyjet tregojnë se mbimrat janë në rasa të ndryshme. Mbiemrat kanë gjininë mashkullore dhe femërore dhe numrin njëjës dhe atë

shumës. Mbiemri në gjuhën Shqipe del me formën e njëjësit dhe të shumësit, në varësi të numrit të emrit që ai përcakton: *djalë i vogël* ose *djem të vegjël*.

Shumësi i mbiemrave të nyjshëm të gjinisë mashkullore formohet duke ndryshuar nyjën '*i*' në '*të*': *i lartë, të lartë*. Shumësi i mbiemrave të nyjshëm të gjinisë femërore formohet: duke ndryshuar nyjen '*e*' në '*të*' ose duke ndryshuar nyjen '*e*' në '*të*' dhe me mbaresën '*a*'.

Mbiemri ka tre shkallë:

a) Shkalla pohore (ku mbiemri tregon thjesht cilësinë e sendit): *Liqeni i Ohrit është i thellë*.
b) Shkalla krahasuese, ku mbiemri tregon cilësinë e një sendi duke e krahasuar me cilësinë e një sendi tjetër. Ajo ndahet në: krahasore të barazisë (*po aq i thellë sa…*), krahasore të sipërisë (*më i thellë se…*) dhe krahasore të ultësisë (*më pak i thellë se…).*
c) Shkalla sipërore: *liqeni i Ohrit është shumë i thellë*.

4.1.3. Përemri

Përemri është një pjesë e ndryshueshme e ligjëratës, që zëvendëson emrin. Ekzistojnë disa lloje përemrash si vijon:

Përemrat vetorë - *unë, ti, ai, ajo* (në njëjës) dhe *ne, ju, ata, ato* (në shumës): tregon veta të ndryshme në njëjës dhe shumës, vetën e parë (vetën që flet), vetën e dytë (vetën me të cilën flasim) dhe vetën e trete (vetën për të cilën flasim).

Përemrat vetvetorë - *vetja, vetvetja*: tregon se i njëjti person vepron dhe i njëjti pëson: Flas me vete. Përemri vetvetor përdoret kryesisht për të treguar persona dhe jo sende. Ky përemër ka kuptim përemëror dhe strukture emërore, ai përdoret me dhe pa parafjalë. Ai përdoret në të njëjtën formë për të dyja gjinitë dhe numrat. Përemri dëftor tregon qënie ose sende ose tipare të tyre që ndodhen afër ose larg folësit. Ky përemër përdoret më vete ose i shoqëruar me një emër.

Përemrat dëftorë - *ky, kjo, këta, këto, i tillë, i këtillë*:

Përemrat pronorë - *imi, yti, asaj, jonë, juaj*: tregon se sendi i dhënë i përket nje vete të caktuar. Ai vendoset pas emrit që përcakton: shkolla jonë. Por ndonjëherë edhe përpara: im atë. Përemri pronor i vetës së parë tregon se sendi i përket folësit ose një grupi ku bën pjesë edhe folësi*: libri im, libri ynë*. Përemri pronor i vetës së dytë tregon se sendi i përket bashkëbiseduesit ose një grupi ku bën pjesë edhe bashkëbiseduesi: *topi yt, topi juaj*. Përemri pronor i vetës së tretë tregon se sendi i përket dikujt që nuk është i pranishëm në bisedë: *makina e tij, makina e tyre*. Për vetën e tretë përdoret dhe përemri pronor i vet, që tregon se sendi i përket kryefjalës së fjalisë: *makina e vet*. Përemri pronor ka forma të veçanta për të dyja gjinitë në të tri vetat. Përemrat pronorë përdoren në njëjës dhe në shumës: *libri im, librat e mi, libri ynë, librat tanë*.

Përemrat pyetës - *kush, cili, cila, çfarë*: shërbejnë për të pyetur për qenie.

Përemrat lidhorë - *që, i cili, e cila, të cilët*: janë përemrat me të cilët i referohemi një fjale a një fjalie të tërë. Ata zënë vendin e një grupi emëror ose të një përemri dhe shërbejnë për të lidhur dy pjesë të një fjalie të përbërë. Për shembull, "*Poshtë urës dukeshin varkat që notonin*".

Përemrat e pacaktuar - *dikush, askush, të gjithë*: tregon njerëz ose sende në mënyrë të pacaktuar.

4.1.4. Foljet

Folja është pjesë e ndryshueshme e ligjëratës, e cila tregon një veprim. Ajo zë vendin qëndror në fjali. Folja zgjedhohet dhe ka kategoritë e saj gramatikore.

4.1.4.1. Kategoritë Gramatikore të Foljeve

Kategoritë gramatikore të foljes janë: zgjedhimi, mënyra, koha, forma, veta dhe numri.

Folja ka tri zgjedhime:

- Zgjedhimi i parë, i cili mbaron me (j).
- Zgjedhimi i dytë, i cili mbaron me bashkëtingëllore.
- Zgjedhimi i tretë, i cili mbaron me zanore.

Folja mund të zgjedhohet në formën veprore (që tregon veprimin që e kryen ti) dhe në formën joveprore (që tregon veprimin që kryen dikush tjetër mbi ty). Vetat klasifikohen në: vetën e parë, vetën e dytë dhe vetën e tretë në numrin njëjës (unë, ti, ai/ajo) dhe shumës (*ne, ju, ata/ato*).

Mënyrat e foljes janë si vijon:

Dallojmë gjashtë lloje të mënyrave të foljes:

- Mënyra dëftore;
- Mënyra lidhore;
- Mënyra kushtore;
- Mënyra habitore;
- Mënyra dëshirore;
- Mënyra urdhërore.

Kohët kryesore të foljes dhe ato të mënyrës dëftore janë:

- Koha e tashme;
- Koha e pakryer;
- Koha e kryer e thjeshtë;
- Koha e kryer;
- Koha më se e kryer;
- Koha e kryer e tejshkuar;

- Koha e ardhme;
- Koha e ardhme e përparme.

Mënyra lidhore ka katër kohë: Kohën e tashme; Kohën e pakryer; Kohën e kryer; Kohën më se të kryer. Mënyra kushtore ka dy kohë: Kohën e tashme dhe Kohën e kryer. Mënyra habitore ka katër kohë: Kohën e tashme; Kohën e pakryer; Kohën e kryer dhe Kohën më se të kryer. Mënyra dëshirore ka dy kohë: Kohën e tashme dhe Kohën e kryer. Mënyra urdhërore ka vetëm vetën e dytë njëjës dhe shumës.

Kohët e foljes ndahen në:

Të thjeshta: Koha e tashme; Koha e pakryer; Koha e kryer e thjeshtë dhe Koha e ardhme.
Të përbëra: Koha e kryer; Koha më se e kryer; Koha e kryer e tejshkuar dhe Koha e ardhme e Përparme.

4.1.4.2. Tipet e Ndryshme të Foljeve

Dallojmë lloje të ndryshme të foljeve:

Foljet ndihmëse (*kam dhe jam*) dhe gjysmëndihmëse (*mund, duhet*…);

Foljet kalimtare dhe jokalimtare. Kalimtare janë ato që kanë një kundrinor, ndërsa jokalimtare janë ato që nuk kanë kundrinor. Foljet kalimtare ndahen në: kalimtare të drejta, ato që kanë një kundrinor të drejtë dhe në kalimtare të zhdrejta, ato që kanë një kundrinor të zhdrejtë.

Foljet pavetore (ato që tregojnë një dukuri atmosferike p.sh.*vetëtimë, bubullimë,* etj.) dhe njëvetore (ato që kanë vetëm vetën e tretë p.sh.*bie shi - ai bie*).

Format e pashtjelluara të foljes janë:

Pjesorja: *punuar*;
Paskajorja: *për të punuar*;
Përcjellorja: *duke punuar*;
Mohorja: *pa punuar*;

Foljet e parregullta, temën e së kryerës së thjeshtë dhe të pjesores e kanë krejt të ndryshme nga tema e trajtës përfaqësuese. Foljet e rregullta quhen kështu ngaqë kalojnë rregullisht nga një kohë në tjetrën (*shikoj – shikoja – shikova – shikuar*). Folja këpujore shërben për të vënë lidhje pjesën kryesore të kryefjalës me kallëzuesorin e kryefjalës, pra shërben si këpujë. Folja këpujore bazë ka foljen *jam* (*është*).

4.1.5. Ndajfoljet

Ndajfolja është pjesë e pandryshueshme e ligjëratës, që plotëson para së gjithash foljen, duke treguar mënyrë, vend, kohë, sasi. Ndajfoljet mund të plotësojnë edhe një mbiemër ose ndajfolje tjetër. Llojet e ndajfoljes janë katër: ndajfolje mënyre; ndajfolje vendi; ndajfolje kohe; ndajfolje sasie. Ndajfolja e mënyrës luan rolin e grupeve parafjalore rrethanorë mënyre. Ndajfoljet e mënyrës tregojnë cilësinë e veprimit, tregojnë ecurinë e veprimit dhe përmasën e veprimit. Ndajfolja e vendit luan rolin e grupeve parafjalore të plotësit të vendit. Ndajfoljet kryesore të vendit janë: *afër, larg, pranë, rreth, djathtas, majtas, aty, atje, diku, gjëkund, gjithkund, anembanë, gjithandej, kudo, ngado*, etj, që tregojnë vendin ku kryhet veprimi. Ndajfolja e kohës luan rolin e grupeve parafjalore rrethanore kohe. Ndajfoljet kryesore të kohës janë: *dje, pardje, nesër, pasnesër, paradite, pasdite, sot, sonte, tani, vjet*, etj, që tregojnë kohën e saktë të kryerjes së veprimit; *njëherë, gjithmonë, kurdo, së shpejti, së lashti*, etj, që tregojnë një kohë të papërcaktuar. Ndajfolja e sasisë luan rolin e grupeve emërore rrethanore sasie. Ndajfoljet kryesore të sasisë janë: *shumë, pak, tepër, fort, njëherë, dy herë*, etj, që tregojnë në ç'masë realizohet veprimi; *fort, tepër, së tepërmi, njëfish, dyfish, trefish, aq, kaq, mjaft, pak*, etj, që shprehin intensitetin ose shkallën e një veprimi a të një cilësie. Ndajfolja përdoret në shkallën pohore, krahasore dhe sipërore. Ato formohen njësoj si shkallët e mbiemrit.

4.1.6. Numëruesi

Numëruesi është pjesë e ndryshueshme e ligjëratës, që tregon numër dhe sasi të caktuar qeniesh ose sendesh: *një, dy, tre, : dy fëmijë, tri gra, katër burra, 20 metra*, etj. Fjalët që tregojnë radhë si: *i parë, i dytë, i dhjetë*, të cilat kanë kategoritë gramatikore të mbiemrave, nuk trajtohen si numërorë, por si mbiemra.

Numëruesit përdoren për të treguar:

- Një sasi të caktuar: *Në punë kishin ardhur vetëm pesë veta*.
- Një sasi të pacaktuar: *Në punë kishin ardhur katër a pesë veta*.
- Datën, vitin: *Me sa jemi sot? – Më 28 Nëntor*.
- Një periudhë të caktuar kohe: - *Kur ke qenë në Amerikë? – Nga 1990 deri më 1997*.
- Përmasa ose përqindje: *0,5 m, 50%*.

Numërorët janë: Fjalë të parme: *zero, një, dy, tre, katër*, etj; Fjalë të përngjitura: *njëmbëdhjetë, dymbëdhjetë*, etj; Fjalë të përbëra: *tridhjetë, pesëdhjetë*, etj; Lokucione: *tridhjetë e një, pesëdhjetë e tre*, etj.

Numërorët në përgjithësi nuk e kanë kategorinë gramatikore të gjinisë: *katër djem, katër vajza*. Bën përjashtim numri tre, i cili e ka kategorinë e gjinisë: *tre djem, tri vajza*.

4.1.7. Lidhëza

Lidhëza është pjesë e pandryshueshme e ligjëratës, që shërben për të lidhur fjalë, grupe fjalësh ose fjali midis tyre. Sipas funksionit lidhëzat ndahen në lidhëza bashkërenditëse dhe lidhëza nënrenditëse. Lidhëzat bashkërenditëse lidhin dy elemente të të njëjtës natyrë ose të të njëjtit funksion. Ato janë: Shtuese; Veçuese; Kundërshtore; Përmbyllëse. Lidhëzat nënrenditëse lidhin dy elemente që kanë marrëdhënie varësie nga njëri-tjetri. Ato janë: Ftilluese; Vendore; Kohore; Shkakore; Qëllimore; Krahasore; Mënyrore; Kushtore; Rrjedhimore; Lejore.

4.1.8. Parafjala

Parafjala është pjesë e pandryshueshme e ligjëratës, që qëndron para emrave, përemrave, numërorëve dhe ndajfoljeve duke vendosur marrëdhënie varësie ndërmjet fjalëve: U rreshtuan për tre.

Parafjalët, sipas strukturës morfologjike janë:
Të thjeshta: *me, në, nga, afër, larg, para, etj*.;
Të përngjitura: *nëpër, përmbi, përveç, sipas, etj*.;
Me shprehje: *ballë për ballë, rreth e qark, etj*.

Parafjalët sipas lidhjeve sintaksore klasifikohen në:
Parafjalë të rasës emërore: *nga, te (tek)*: *U rrëzua nga lodhja*.
Parafjalë të rasës gjinore: *me anë (anën), me rastin, me përjashtim, në saje, në vend* etj. *Me anë të një miku arrita t'i kapërcej vështirësitë*.
Parafjalë të rasës kallëzore: *në, me, pa, për, më, mbi, nën, ndër, nëpër* etj. *Erdhi vonë në shtëpi*.
Parafjalë të rasës rrjedhore: *prej, ndaj, buzë* etj. *Ishte nga letre*.

4.1.9. Pasthirrma

Pasthirrma është pjesë e pandryshueshme e ligjëratës, që shërben për të treguar ndjenjat e folësit ose përdoret për të tërhequr vëmendjen e dëgjuesve*: O, sa mirë me qënë Shqiptar*!

Pasthirrmat janë karakteristike për gjuhën e folur dhe kuptimi i tyre shpesh qartësohet nga intonacioni. Pasthirrmat grupohen në pasthirrma emocionale (që shprehin ndjenja) dhe në pasthirrma urdhërore (që shprehin dëshirën ose vullnetin e folësit). Pasthirrmat vendosen zakonisht në fillim të fjalisë, por shpesh përdoren edhe në mes, edhe në fund të saj. Pas tyre vihet presje. Kur shqiptohen me fuqi të veçantë, pas tyre vihet pikëçuditje: *Ua! Paske ardhur!*

4.1.10. Onomatopetë

Onomatopetë janë fjalë që shërbejnë për të riprodhuar tingujt, zërat, zhurmat, britmat. Në dallim nga pasthirrmat, ato nuk shërbejnë për të shprehur ndjenja: *Mjau, mjau – bënte macja*. Ato mund të përfshihen edhe te pasthirrmat, por gjithsesi dallohen prej tyre.

4.1.11. Pjesëza

Pjesëza është pjesë e pandryshueshme e ligjëratës, që plotëson kuptimin e një fjale ose të një fjalie. Sipas kuptimit, pjesëzat grupohen në: Dëftuese; Përcaktuese – saktësuese; Përforcuese; Përafruese – veçuese ose kufizuese; Përafruese; Pohuese; Mohuese; Pyetëse; Dyshuese; Zbutëse; Emocionale shprehëse.

4.2. Sintaksa e Gjuhës Shqipe

Sintaksa është pjesa e gramatikës që studion fjalët gjatë ligjërimit, rregullat e bashkimit të tyre në grupe fjalësh e në fjali si dhe rolin që luajnë ato. Përbërësit kryesorë të fjalisë në gjuhën Shqipe ndahen në dy grupe të mëdha: grupi emëror (kryefjala) dhe grupi foljor (kallëzuesi). Grupi emëror tregon për personat për të cilët flitet, kurse grupi foljor tregon çfarë thuhet për ta. Si shëmbull mund të përmëndim fjalinë "*Të gjithë fëmijët rrinin të qetë.*" ku "*të gjithë fëmijët*" është grupi emëror, ndësa "*rrinin të qetë*" është grupi foljor. Fjalia është një grup fjalësh të lidhura bashkë sipas kuptimit dhe rregullave gramatikore të gjuhës. Në gjuhën e folur fjalitë karakterizohen nga një intonacion ose vijë melodike e veçantë (zbritëse, ngjitëse etj.). Në gjuhën e shkruar ato përcaktohen nga shenjat e pikësimit (pikë, pikëpyetje, pikëpresje, pikëçuditje).

4.2.1. Tipet e Fjalive

Një fjali mund të jetë e thjeshtë ose e përbërë. Fjalia është e thjeshtë kur ka vetëm një folje. Fjalia është e përbërë kur ka më shumë se një folje. Fjalitë ndahen në katër tipe kryesore: dëftore, pyetëse, nxitëse dhe dëshirore.

Fjalia dëftore shërben për të treguar diçka, për të dhënë një informacion rreth një fakti, rreth një vëzhgimi, një gjykimi, një opinioni. Fjalia dëftore, kur është e mëvetësishme, shqiptohet me një intonacion tregues. Toni zbret në fund të fjalisë. Ai ngrihet në fjalën që mban theksin logjik, kur është i dukshëm. Fjalia dëftore ka foljen zakonisht në mënyrën dëftore, por mund ta ketë edhe në mënyrën lidhore, habitore e kushtore.

Fjalia pyetëse përdoret për të pyetur për diçka që folësi nuk e di ose e di pjesërisht dhe pret nga bashkëbiseduesi një përgjigje. Këto fjali përdoren dendur në dialogje, sepse bashkëbiseduesit për çdo pyetje që bëjnë, presin një përgjigje. Dallohen tri lloje fjalish pyetëse: fjali pyetëse tërësore; fjali pyetëse të pjesshme dhe fjali pyetëse retorike.

Fjalia është **pyetëse tërësore**, kur bashkëbiseduesi i përgjigjet pyetjes me po, jo ose me foljen e fjalisë pyetëse: - *Ishe sot te mjeku? Po. (Isha.)*. Fjalia është pyetëse e pjesshme, kur folësi ka njohuri për atë që pyet, por ai do të mësojë më shumë. Këtyre pyetjeve nuk u jepet përgjigje me

po ose jo, por me një ose disa fjalë: - *Ç'muaj është tani? – Janar*. Fjalia është pyetëse retorike, kur folësi pyet pa pritur që të marrë përgjigje: - Ç'të presësh nga ti?

Fjalia nxitëse shërben për të nxitur dikë që të zbatojë një urdhër, një këshillë, një sugjerim, një ftesë. Në fjalitë nxitëse foljet janë zakonisht në mënyrën urdhërore, lidhore e dëftore.

Fjalia dëshirore shërben për të shprehur një dëshirë, një urim, një mallkim. Fjalia dëshirore mund të shprehë edhe dëshirën që të mos ndodhë diçka. Fjalitë dëshirore e kanë foljen në mënyrën dëshirore dhe në mënyrën lidhore. Në fund të fjalive dëshirore vihet pikëçuditje.

4.2.2. Format e Fjalive

Fjalitë dëftore, pyetëse, nxitëse dhe dëshirore mund të paraqiten në shumë forma: veprore ose pësore, pohore ose mohore dhe thirrmore. Fjalia është në formën veprore, kur folja e saj është e formës veprore dhe është në formën pësore, kur folja e saj është e formës joveprore (ose pësore). Në formën mohore të fjalisë mohohet një pohim. Pra, fjalia e formës pohore pëson një shndërrim mohor. Fjalitë bëhen thirrmore kur shoqërohen me një ndjenjë të fortë kënaqësie, habie, frike, zemërimi etj.

4.2.3. Pjesët Përbërëse të Fjalisë

Pjesët kryesore të fjalisë janë kryefjala, kallëzuesi dhe plotësuesit e fjalisë.

Kryefjala është gjymtyrë kryesore e fjalisë. Ajo gjendet me pyetjet: cili?, cila?, cilët? dhe kush?. Ajo tregon një a më shumë frymore apo sende. Kryefjala zakonisht qëndron para foljes. Kryefjala shprehet me:

- emër në rasën emërore, të shquar e të pashquar, në numrin njëjës e shumës: *Dritani u lag nga shiu. (kush?)*
- përemër (dëftor, të pacaktuar, lidhor, vetor të çdo vete a numri): *Ajo më kërkoi një trëndafil të kuq.*
- numëror: *Të dy filluam të ecnim duke çaluar në korridor.*
- grup emëror: *Yt vëlla të la vetëm.*

Kryefjala mund të marrë një kallëzuesor: "*Shqiponja është simboli ynë kombëtar*."

Kallëzuesori i kryefjalës tregon një cilësi të personit apo sendit që shënohet nga kryefjala. Kallëzuesi është gjymtyra kryesore e fjalisë, që tregon një veprim apo gjendje të kryefjalës. Ai gjendet me pyetjet: ç'bën?, ç' është?. Kryefjala dhe kallëzuesi janë të lidhur ngushtë me njëri-tjetrin. Kështu, po të ndryshojë kryefjala, ndryshon edhe kallëzuesi. Kallëzuesi përshtatet me kryefjalën në numër dhe në vetë. Kallëzuesi zakonisht qëndron pas kryefjalës.

Llojet e kallëzuesit janë:

- Kallëzuesi foljor i thjeshtë (një folje): *Ti më munde*.

- Kallëzuesi foljor i përbërë (dy e më shumë folje): *Po djali sa zinte të recitonte, harronte vjershën.*
- Kallëzuesi emëror (jam+emër, mbiemër, përemër, numëror): *Dy prej luanëve ishin meshkuj*.

Gjymtyrët e dyta (apo plotësit e fjalisë) janë: rrethanori; kundrinori; përcaktori.

Rrethanori është gjymtyrë e dytë e fjalisë. Ai tregon rrethanat e kryerjes së veprimit, si: kohë, vend, mënyrë, shkak, sasi, qëllim. Ai në fjali gjendet me pyetje të ndryshme: kur?, ku?, si?, pse?, sa?, përse?. Rrethanori gati gjithmonë qëndron pas foljes ose kallëzuesit, por ai shpesh del edhe para foljes ose kallëzuesit.

Rrethanori i kohës tregon kohën e kryerjes së veprimit. Ai lëviz lirshëm në fjali dhe gjendet me pyetjen: kur?: "*Mbrëmë shkuam në kinema*."

Rrethanori i kohës shprehet me:

- emër me e pa parafjalë: "*Ndeshja finale do të zhvillohet të mërkurën*."
- grup emëror: "*Ndeshja finale u zhvillua para një jave*."
- ndajfolje ose shprehje ndajfoljore: "*Ndeshja finale u zhvillua dje."*
- formë të pashtjelluar: "*Me të zbardhur fëmijët u nisën në ekskursion*."

Rrethanori i vendit tregon vendin e kryerjes së veprimit. Ai lëviz lirshëm në fjali. Rrethanori i vendit gjendet me pyetjen: ku?: "*Ata shkuan në shkollë."*

Rrethanori i vendit shprehet me:

- emër me parafjalë: "*Librat i mbaj në bibliotekë*. (ku?)"
- grup emëror: "*Jam nisur për në Galerinë e Arteve*. (për ku?)"
- ndajfolje: "*Zhurma dëgjohej sipër*. (ku?)"

Rrethanori i mënyrës tregon mënyrën e kryerjes së veprimit. Ai lëviz lirshëm në fjali. Rrethanori i mënyrës gjendet me pyetjen: si?: "*Ajo këndon bukur."*

Rrethanori i mënyrës shprehet me:

- një ndajfolje ose shprehje ndajfoljore (me emër në rasën kallëzore): "Dëbora mbulonte shpejt *e shpejt (si?) çdo gjë."*
- një grup emëror: "*Një flutur po thithte me ëndje të madhe (si?) nektarin e luleve*."
- një formë të pashtjelluar: "*Djemtë ecnin me nxitim pa parë (si?) rrugën."*

Rrethanori i shkakut tregon një shkak që sjell një pasojë të caktuar. Ai varet nga kallëzuesi, por ndonjëherë edhe me emër ose mbiemër. Rrethanori i shkakut zë vend pas gjymtyrës nga varet. Ai gjendet me pyetjen: pse?: "*Ajo shkoi te mjeku, sepse ishte e sëmurë*."

Rrethanori i shkakut shprehet me:

- ndajfolje: "*Pse ke bërë gjithë këto gabime në diktim, Dritan?"*
- emër me parafjalë: *"- I kam bërë nga nxitimi, mësuese*."
- grup emëror: "*Rezultatet i ke të ulëta për shkak të mungesave."*
- përcjellore: "*Duke mos u mjekuar me kohë, sëmundja iu rëndua."*

Rrethanori i sasisë tregon sasi. Ai gjendet me pyetjen: sa?: "*Ata morën pak libra*."

Rrethanori i sasisë shprehet me:

- një ndajfolje ose shprehje ndajfoljore: "*Mbrëmë ajo hëngri pak."*
- një grup emëror ku bërthama tregon një njësi matjeje dhe ka përpara zakonisht një numëror: "*Nëna bleu 5 kg mollë*."

Rrethanori i qëllimit tregon qëllimin e kryerjes së veprimit. Ai varet nga kallëzuesi. Rrethanori i qëllimit zë vend pas gjymtyrës nga varet. Ai gjendet me pyetjen: përse?: "*Ai shkoi për të luajtur."*

Rrethanori i qëllimit shprehet me:

- ndajfolje: "*Përse ke ardhur deri këtu?"*
- emër me parafjalë: "*Djemtë kanë ardhur këtu për punë."*
- grup emëror: "*Prindërit përpiqen për të mirën e fëmijëve të tyre."*
- paskajore: *"Ajo ka shkuar në Amerikë për t'u specializuar*."

Kundrinori është gjymtyrë e dytë e fjalisë. Ai zakonisht varet nga folja dhe tregon mbi kë (cilin) bie veprimi i shprehur prej saj. Ai vihet në atë rasë që kërkon gjymtyra prej së cilës varet. Sipas kuptimit dhe formës së shprehjes, kundrinori është i drejtë dhe i zhdrejtë.

Llojet e kundrinorit janë: I drejtë; I zhdrejtë: me parafjalë ose pa parafjalë.

Kundrinori i drejtë është gjymtyrë e dytë e fjalisë që shënon objektin tek i cili kalon veprimi i shprehur nga folja. Objekti mund të jetë frymor ose jofrymor. Kundrinori i drejtë u përgjigjet pyetjeve: kë?, cilin?, cilën?, çfarë?. Kundrinori i drejtë është një grup emëror (i shprehur me një emër, një përemër ose një emër dhe një përcaktor) në rasën kallëzore pa parafjalë dhe që lidhet me një folje të formës veprore: "*Era e marrë tundte rrapin e moçëm të Jubinës*."

Në qoftë se fjalinë veprore me kundrinor të drejtë e shndërrojmë në fjali pësore, kundrinori i drejtë bëhet kryefjalë e foljes në formën pësore, kurse kryefjala, kundrinor me parafjalë: "*Era e malit e shpërndante këngën gjithandej. Kënga shpërndahej nga era e malit gjithandej*."

Kundrinori i drejtë shprehet:

- me emër (zakonisht të shquar) në rasën kallëzore: "*E ndoqe ndeshjen mbrëmë?"*
- me përemër: "*E shoh atë, të shoh edhe ty."*

- me grup emëror: "*Vizitova liqenin e Prespës vitin e kaluar.*"
- me trajtë të shkurtër të përemrit vetor: *"Do ta njoftoj vetë Arbrin për orarin e ndeshjes."*

Kundrinori i zhdrejtë pa parafjalë ka lidhje të drejtpërdrejtë me foljen dhe përdoret gjithmonë i shoqëruar me trajtë të shkurtër të përemrit vetor. Mund të marrin kundrinor të zhdrejtë foljet kalimtare dhe jokalimtare, veprore dhe joveprore. Kundrinori i zhdrejtë pa parafjalë u përgjigjet pyetjeve: kujt?, cilit?, cilës?, cilëve?, cilave?. Kundrinori i zhdrejtë pa parafjalë është një grup emëror (i shprehur me një emër, një përemër ose me një emër dhe një përcaktor) në rasën dhanore, që shoqërohet gjithnjë nga një trajtë e shkurtër e përemrit vetor në rasën dhanore. Ai ndërtohet gjithmonë pa parafjalë dhe tregon sende ose qenie mbi të cilat bie veprimi i shprehur nga folja. Kundrinori i zhdrejtë pa parafjalë luan rol të rëndësishëm, është plotës kryesor. Ai nuk është i lidhur me foljen aq sa kundrinori i drejtë. Vendi i kundrinorit të zhdrejtë pa parafjalë është zakonisht pas foljes. Por ai mund të zërë lirisht pozicione të ndryshme në fjali.

Kundrinori i zhdrejtë me parafjalë ka një përdorim më të gjerë e të larmishëm sesa kundrinori i zhdrejtë pa parafjalë. Ai varet kryesisht nga folja, por edhe nga pjesë të ndryshme të ligjëratës. Kundrinori i zhdrejtë me parafjalë i përgjigjet pyetjeve: për kë?, me se?, nga kush?, prej kujt?, nga se?. Kundrinori i zhdrejtë me parafjalë është një grup emëror (i shprehur me një emër, një përemër ose me një emër dhe një përcaktor) në rasën emërore, kallëzore ose rrjedhore, i paraprirë gjithmonë nga një parafjalë. Kundrinori i zhdrejtë me parafjalë plotëson më mirë kuptimin e foljes dhe të fjalisë, por nuk është plotës kryesor, sepse folja në fjali mund të përdoret edhe pa të. Parafjalët kryesore me të cilat ndërtohet kundrinori i zhdrejtë me parafjalë, janë: *nga, te, me, për, pa, mbi, nën, kundër*; shprehjet parafjalore – *bashkë me, me anë të, në lidhje me*. Kundrinori i zhdrejtë me parafjalë, i ndërtuar nga një emër ose një emër dhe një përcaktor, mund të zëvendësohet nga një përemër. Kundrinori i zhdrejtë me parafjalë dallohet nga dy kundrinorët e tjerë, sepse ndërtohet gjithmonë me parafjalë.

Përcaktori. Fjalët që saktësojnë, sqarojnë, përcaktojnë emrin bërthamë me të cilin lidhen, quhen përcaktorë. Emri në rasën gjinore të shquar është gjithmonë në funksionin e përcaktorit. Ai shprehet me: emër, mbiemër dhe përemër. Llojet e përcaktorit janë:

- Me përshtatje, i cili shërben për të përcaktuar cilësinë e emrit bërthamë;
- Me drejtim, i cili shërben për të treguar një tipar të emrit që përcakton;
- Me bashkim;
- Me ndajshtim, i cili emërton emrin bërthamë.

4.2.4. Formimi i Fjalëve në Gjuhën Shqipe

Sipas llojit të ndajshtesës fjalëformuese dhe sipas mënyrës së bashkimit të saj me temën, dallohen këto lloje të prejardhjesh (Kadiu, 2012).

a) Prejardhja parashtesore: Një numër i madh fjalësh janë formuar me anë të parashtesimit, p.sh. *përshpejtoj, përgëzoj, përuroj etj.*

b) Prejardhja prapashtesore: Është mënyra e formimit të fjalëve të reja me anë prapashtesash të cilat bashkohen me tema fjalësh. Gjatë prapashtesimit tema fjalëformuese mund të pësojë ndryshime fonetike: zanorja fundore e temës fjalëformuese bie, kur prapashtesa nis me zanore: Me anë prapashtesash formohen emra, mbiemra, folje, ndajfolje prej të njëjtës pjesë të ligjëratës.

Prapashtesat në përgjithësi kanë funksion klasifikues d.m.th. ato bashkë me kuptimin e ri që i japin temës së cilës i ngjiten, përcaktojnë edhe pjesën e ligjëratës ku hyn fjala e re. Por ka edhe prapashtesa që shërbejnë për të formuar dy pjesë të ndryshme ligjërate p.sh. emra dhe mbiemra *(ë)s, -ar, -tar, -or, -tor* etj. Në gjuhën shqipe ka edhe prapashtesa homonimike. Ato përputhen në shqiptim e në shkrim, por kanë kuptime krejt ta palidhura ndërmjet tyre. Në këto raste kemi të bëjmë me dy prapashtesa të ndryshme: *-as elbasanas* dhe *-as barkas, -çe nipçe* dhe *–çe partizançe -je marrje* dhe *-je këmbje*.

Po ashtu gjuha shqipe ka edhe prapashtesa sinonimike, por zakonisht ato ose kanë sfera veprimi pak e shumë të veçanta, ose njëra prej tyre është prodhimtare, kurse tjetra joprodhimtare. Prapashtesa sinonimike *-i dhe -(ë)si*, që përdoret për të formuar emra femërorë, e para bashkohet kryesisht me tema më *-r a -s dituri, zejtari, befasi* etj. Kurse në rastet e tjera mbizotërojnë prapashtesat *-(ë)si, - (ë)ri*, por edhe këto kanë sferat e tyre të veçanta të veprimit.

c) Prejardhja parashteso-prapashtesore: Përveç prejardhjes me parashtesa ose me prapashtesa dallohet edhe një nënndarje tjetër e kësaj mënyre fjalëformimi, prejardhja parashteso-prapashtesore: një teme i ngjiten njëkohësisht një parashtesë e një prapashtesë dhe të dyja këto morfema luajnë rolin e një formanti të vetëm fjalëformues, p.sh. *zbukuroj (z.... o).*

d) Prejardhja pa ndajshtesë: Ka një numër fjalësh, të cilat nga pikëpamja kuptimore janë të motivueshme, të prejardhura, megjithëse nga pikëpamja strukturore paraqiten pa ndonjë ndajshtesë fjalëformuese. Të tilla janë emrat *hap-i, mund-i* të formuara përkatësisht prej foljeve *hap, mund* po ashtu foljet *huaj, krip* etj. të formuara përkatësisht prej emrave dhe mbiemrave *hua, kripë* etj.

Emrat e përgjithshëm të vetave formohen me anë të këtyre prapashtesave:

a) me prapashtesën *-ës* formohen emra njerëzish të karakterizuar sipas veprimit që kryejnë ose profesionit a punës që bëjnë. Kjo prapashtesë u shtohet kryesisht foljeve në temën e pjesores. P.sh. *mbledhës, shitës, vjelës.* Tek emrat e tipit *dëgjues, krijues, shkrues, ndërtues* etj., prapashtesa bashkohet me temën e pjesores së gegërishtes.

b) me prapashtesat *-ar, -tar, -atar, -or, -tor, -ator, -ak, -as, -an -jan -it, -iot* formohen emra të përgjithshëm njerëzish të karakterizuar sipas veprimit që kryejnë, sipas një tipari të tyre, sipas mjeshtërisë që ushtrojnë, po ashtu dhe sipas vendit ku banojnë ose kanë lindur.

c) Për formimin e emrave të vetave përdoren edhe disa prapashtesa të reja të burimit të huaj si: *-ist, -ik, -ant, -ent, -ier* dhe *–er*. si p.sh. *artist, futbollist, diplomant (diplomë), emigrant (emigroj), kursant (kurs), karrocë (karrocier), portë (portier), bojaxhi, qiraxhi, shakaxhi* etj.

Për formimin e emrave të sendeve përdoren prapashtesa të ndryshme, disa nga të cilat shërbejnë për të formuar edhe emra vetash si p.sh. *(ë)s, -or(e), -ar*. Për të formuar emra sendesh janë përdorur edhe prapashtesat e emrave abstraktë: *-esë* dhe *-je*. Prapashtesat e tjera janë *-ishte, -(ë)sirë(a), -urin(a), -ac, -açe, -e*. Me anë të prapashtesës *-(ë)s(e),* formohen nga foljet edhe emra mashkullorë e femërorë që tregojnë një vegël, një orendi a një pjesë veshjeje. P.sh. *çelës, matës, ndezës*, etj.

Për formimin e disa emrave të përgjithshëm të vendeve përdoren prapashtesa *-ishte, -ishtë, -ore, -tore, -inë, -ajë*. P.sh. *fidanishte, lulishte, mëngjesore, bulmetore, ëmbëltore, pjesore, dëftore, luginë, çmendinë, shkretëtirë, zbrazëtirë*, etj.

Për formimin e emrave abstraktë përdoret një numër i madh prapashtesash.

a) Me anë të prapashtesave *-i, -(ë)si, -(ë)ri* formohen kryesisht prej mbiemrash dhe emrash dhe më pak prej foljesh, ndajfoljesh e përemrash, emra femërorë që tregojnë një veçori ose cilësi abstrakte, p.sh. *bukuri, dituri, lumturi, vjetërsi, fisnikëri, besnikëri, pleqëri, ligjshmëri,* etj.

b) Emrat astraktë që emërtojnë veprime ose gjendje formohen nga temat foljore me anë të prapashtesave *-im, -je, -esë, -atë, -imë, -më.* Prapashtesat *-im* dhe *-je* janë prodhimtare, sidomos në gjuhën letrare. P.sh. *besim, dorëzim,këmbim, kërcim, veshje, lindje, thyerje, humbje, mbetje, jetesë, harresë, martesë, dhuratë, kafshatë, britmë, krismë, psherëtimë, vetëtimë,* etj.

c) Për formimin e emrave abstraktë në gjuhën shqipe përdoren edhe disa prapashtesa të huaja si *-llëk, -izëm, -azh, -urë*. P.sh. *budallallëk, fodullëk, optimizëm, montazh, atletikë,* etj.

Disa prapashtesa u japin emrave ngjyrime të ndryshme ndjenjësore dhe shprehëse. Prapashtesat më të përdorshme me kuptim përkëdhelie e zvogëlimi janë: *-th, -z(ë).* P.sh. *zogth, lulez, shtëpizë*, etj. Rrallë përdoren edhe disa prapashtesa të tjera si *-ush*. P.sh. *babush, lepurush, vogëlush,* etj. Me *-ushkë* kemi shembujt *fletushkë, lambushkë* etj. Me prapashtesën –icë kemi shembullin *rrugicë*. Me shtrirje dialektore të ngushtë përdoren edhe prapashtesat -kë dhe -çkë. P.sh. *çupkë, guriçkë, deriçkë, koriçkë* etj. Prapashtesat e përkëdhelisë *-ko* dhe *-çe* u shtohen disa emrave mashkullorë. P.sh. *dajko, vëllako, birçe, nipçe,* etj. Prapashtesat që shërbejnë për të formuar emra me një kuptim keqësues janë *-ac, -acak, -acuk, -aq, -alaq, -aluq, -aluq, -alec, -arash, -avec* e disa

të tjera më të rralla. Këto përdoren edhe si mbiemra të panyjshëm me po atë kuptim keqësues a përbuzës. Me prapashtesën *-ac* formohen emra prej emrash. P.sh. *dorac, qyrrac, tullac,* ose prej një foljeje si p.sh *përtac*.

Me prapashtesën *-acak* formohen emra prej emrash, p.sh. *burracak, frikacak, morracak, rrenacak* ose prej foljesh, p.sh. *ngordhacak, rrjepacak, vjedhacak* etj. Ndonjëherë ajo këmbehet me prapashtesën *-acuk*, p.sh. *rrjedhacuk, vardhacuk* ,etj.

Prapashtesat e tjera përdoren më pak: *-alaq: ngordhalaq, -aluq: shtrembaluq, trashaluq, -alec: barkalec, -aman: çalaman, frikaman, qaraman: -arash: mëngjarash: -avec: burravec, grindavec, qullavec, qurravec* (Kadiu, 2012).

5. Analiza dhe Projektimi i Algoritmit

Qëllimi i këtij kapitulli është ndërtimi i një algoritmi për nxjerrjen e automatizuar të fjalëve kyçe – një bashkësi fjalësh të dokumentit që përgjithësojnë përmbajtjen e tij. Shpesh here gjetja e fjalëve kyçe është përdorur për të zgjidhur problemin e krijimit të përmbledhjes së një dokumenti, të ndërtimit të indeksuesve të dokumentave, për të klasifikuar tekste, për të identifikuar dokumentat e kërkuar nëpërmjet një query të përdoruesit, etj. Të gjithë këto probleme janë të varuara nga gjuha në të cilën është shkruar dokumenti.

Në këtë punim, ne kemi pëdorur teknikat aktuale të vlerësimit statistikor të peshave të fjalëve në një dokument duke i shtuar elementë linguistikë specifikë për gjuhën Shqipe dhe duke i kombinuar me algoritma klasifikimi të Machine Learning për të rritur cilësinë e fjalëve kyçe të gjetura.

Në këtë kapitull do të paraqesim në mënyrë të detajuar komponentët përbërës të algoritmit të gjetjes së fjalëve kyçe të një dokumenti. Ky algoritëm prezantohet si nënproblem i atij të kërkimit të informacionit, të aplikuar në korpuse në gjuhën Shqipe të kategorizuara në klasa. Algoritmi është i organizuar në disa module logjike. Moduli i parë lidhet me elementët linguistikë në gjuhën Shqipe; moduli i dytë lidhet me elementet statistikorë të nxjerrë nga trajnimi i korpusit dhe moduli i tretë shton elementë të Machine Learning për të optimizuar rezultatet e arritura.

5.1. Komponenti Linguistik

Komponenti linguistik është një pjesë e rëndësishme e algoritmit të propozuar për gjetjen e fjalëve kyçe i cili i takon fazës së paraproçesimit të dokumentave. Punimet shkencore në fushën e Text Mining në gjuhë të tjera, (shumica e punimeve i takojnë gjuhës Angleze), theksojnë faktin që paraproçesmi linguistik rrit ndjeshëm performancën e algoritmave Machine Learning të aplikuara për Text Mining. Komponenti linguistik që ne po përdorim përmban të gjitha rregullat gramatikore të gjuhës Shqipe dhe është i bazuar në rregullat e gramatikës së gjuhës Shqipe për fjalëformimin (Agalliu et al., 2002). Komponenti linguistik ka dy funksione kryesore:

5.1.1. Eleminimi i Lidhësave/Konektorët e Tekstit

Konektorët e tekstit janë fjalët e pandryshueshme që shërbejnë për të lidhur fjalë, grupe fjalësh ose fjali midis tyre. Për shkak se konektorët e tekstit nuk mbartin kuptim në përcaktimin e fjalëve kyçe do t'i heqim ato nga dokumenti në fazën e paraproçesimit. Për këtë është implementuar një listë konektorësh të gjuhës Shqipe e cila përdoret pët të identifikuar konektorët nga dokumentat tekst. Listën e plote të lidhësave të marra në konsideratë e gjeni në Aneksin A.

5.1.2. **Gjetja e Rrënjëve të Fjalëve të Thjeshta**

Problemi i gjetjes së rrënjëve të fjalëve ka për qëllim segmentimin automatik të fjalëve në njësitë më të vogla me kuptim, në morfema, p.sh. fjala *përpunoj* mund të ndahet në morfemat përbërëse: *për* (parashtesa) , *pun* (rrënja), *o* (prapashtesa) dhe *j* (mbaresa). Për shkak të numrit të madh të formave të ndryshme të fjalëve gjetja e rrënjëve të fjalëve që ndodhen në një tekst kthehet në një detyrë të nevojshme. Moduli i kthimit të fjalëve në rrënjët e tyre synon largimin automatik të pjesëve të ndryshme të një fjale duke i krahasuar me listat paraprake të ndajshtesave. Nëse këto segmente fjalësh ndodhen në listat përkatëse të parashtesave, prapashtesave apo të rrënjëve, programi i heq duke arritur në zbulimin e rrënjës bazë e cila është e pacopëzueshme më tej në njësi me kuptim. Këto algoritma përdoren në tipe të ndryshme problemesh të proçesimit të gjuhëve natyrore; në nxjerrjen e informacionit bazë nga një tekst i dhënë si dhe në sistemet e kërkimit në bazat e të dhënave. Nëse nuk do të shfrytëzoheshin këta segmentues, dikush që do të kërkonte informacion për një fjalë të caktuar, p.sh. *program*, do të merrte një informacion të kufizuar vetëm për këtë fjalë pa e zgjeruar më tej me format e ndryshme të kësaj fjale, si *programim*, *programues*, *programe*, etj. Metodat për gjetjen e rrënjëve të fjalëve paraqiten në skemën e mëposhtme:

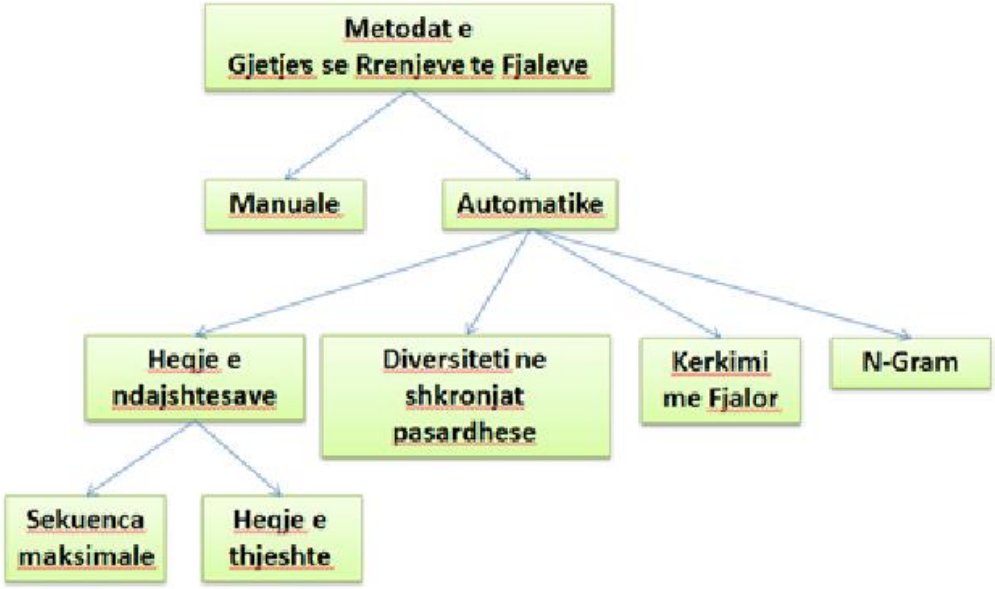

Figura 5.1. Metodat e gjetjes së rrënjëve të fjalëve

Metoda më e thjeshtë për gjetjen e rrënjëve të fjalëve është ajo e gjetjes dhe fshirjes së ndajshtesave për secilën fjalë, e cila bazohet në rregullat e derivuara nga analiza morfologjike e tekstit. Kjo metodë është zgjedhur të përdoret në fazën e paraproçesimit të dokumentave në algoritmin tonë. Gjetja e rrënjëve të fjalëve ne fazën e paraproçesimit në përgjithësi ul me 50% madhësinë e dokumentave që do të indeksohen, duke rritur performancën e zgjidhjes. Për sa i

përket rritjes së efiçensës është provuar që në rastin më të keq, në problemet e proçesimit të gjuhëve natyrore, gjetja e rrënjëve të fjalëve nuk ndikon në efiçensë ndërsa në përgjithësi e përmirëson ndjeshëm atë (Frakes and Baeza-Yates , 1992).

Nga ky komponent do të marrim output një dokument të ri i cili përmban vetëm rrënjët e fjalëve dhe nuk përmban konektorët e tekstit.

Algoritmi i përdorur nga ne trajton një bashkësi rregullash duke zgjedhur rregullin përkatës për secilën fjalë. Në këtë mënyrë janë eleminuar parashtesat apo prapashtesat më të gjata të cilat përban fjala. Ky rregull nuk zbatohet nëqoftëse rrënja a fjalës që formohet pas heqjes së tyre ka vetëm një shkronjë. Në gjuhën Shqipe nuk ekzistojnë rrënjë fjalësh me vetëm një shkronjë, randaj në një rast të tillë kërkojmë rrënjë fjalësh me të paktën dy shkronja. Një rregull formohet nëpërmjet disa kushteve të cilat përdoren sipas rastit. Në disa raste ekzistojnë fjalë të cilat kanë dy zanore apo prapashtesa njëra pas tjetrës. Në një rast të tillë, rregullat e heqjes së prapashtesave apo zanoreve do të aplikohen dy herë (Sadiku, 2011). Në Figurën 5.2 jepet shembulli i një rregulli të thjeshtë. Algoritmi i përdorur për nxjerrjen e rrënjëve të fjalëve përmban gjithësej 134 rregulla për fjalët e thjeshta.

```
 if ((word.endsWith("ësh") || word.endsWith("esh")) && word.length() > 5) {
if (word.contains("shpesh")) {
    return word;
} else {
    word = word.substring(0, lastPos - 3);
}
}
```

Figura 5.2. Shembull i një rregulli për nxjerrjen e rrënjës së fjalës

Ndryshe nga gjuhët e tjera, gjuha Shqipe nuk ka një rregull të përgjithshëm për të formuar shumësin prandaj nuk është përfshirë në bashkësinë e rregullave një rregull i tillë. E njëjeta gjë mund të thuhet edhe për gjininë, femërore, mashkullore apo neutrale. Rast i veçantë i algoritmit është rasti i një prapashtese që përfundon me zanore. Fjalët të cilat përfundojnë me papashtesa të tilla duhet më parë të proçesohen para se të aplikohet heqja e zanores. Në Figurën 5.3 jepen hapat që ekzekutohen nga algoritmi (Sadiku, 2011). Çdo hap ka rregullat e tij të derivuara nga analiza morfologjike e gjuhës Shqipe.

1. Heqja e parashtesave: mënyra me e shpeshtë e formimit të pjesëve të ligjëratës janë parashtesat dhe prapashtesat. Hapi i parë i algoritmit trajton parashtesat. Rregullat e përdorura përfshijnë parashtesat më të rëndësishme të gjuhës Shqipe.

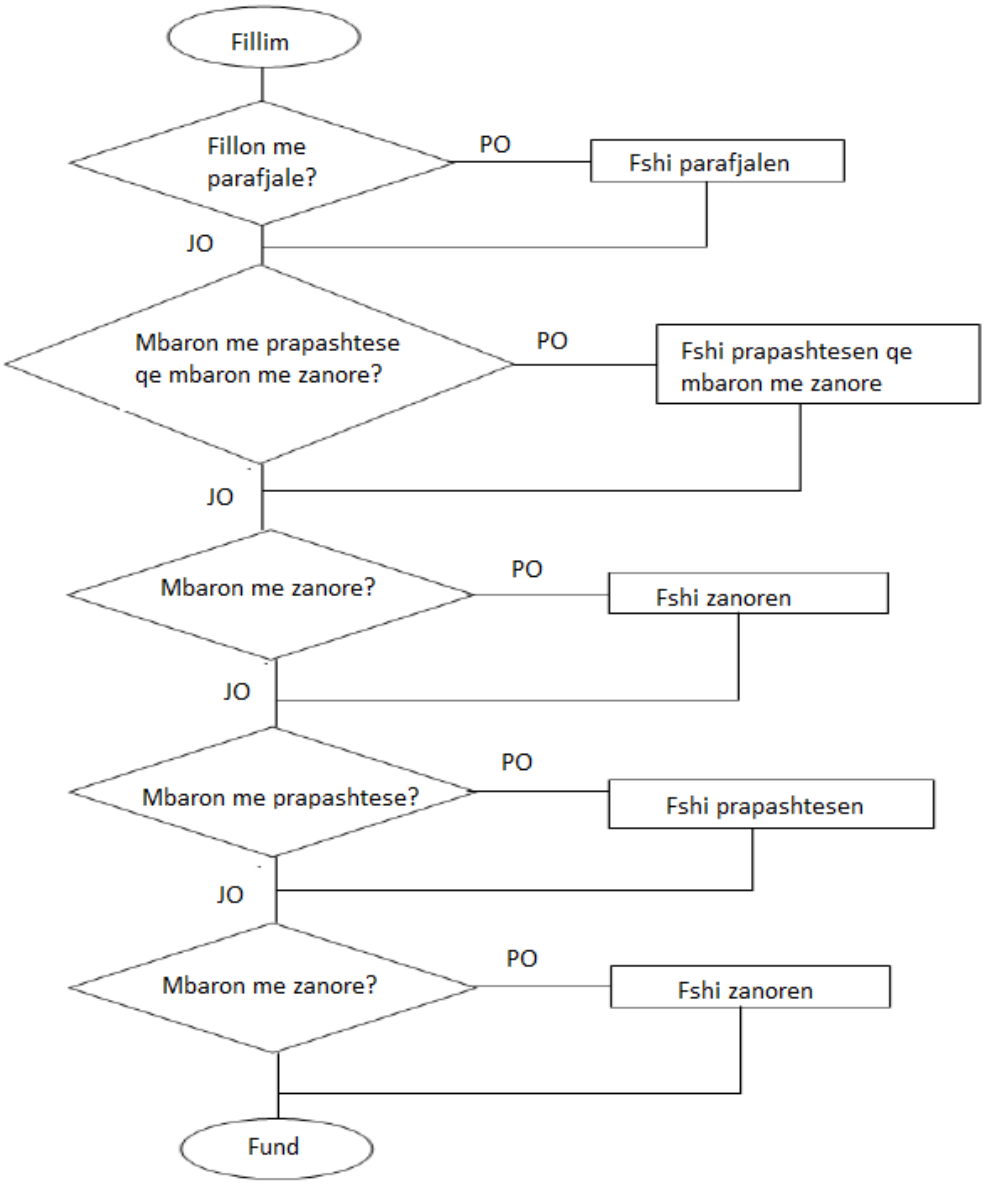

Figura 5.3. Algoritmi i gjetjes së rrënjës së fjalës (Sadiku, 2011).

2. Prapashtesat që mbarojnë me zanore: disa nga prapashtesat në gjuhën Shqipe përfundojnë me zanore, si psh *–je* (*veshje*, etj.) ose *–shmëri* (*ngjashmëri*, *gadishmëri*, etj.) të cilat proçesohen në një hap të veçantë të algoritmit. Është e rëndësishme që ky hap të proçesohet para atij të eleminimit të zanores.

3. Eleminimi i zanores: shumë nga fjalët, të cilat preken ose jo nga rregulli i parë, japin një formë të fjalës e cila mund të përmirësohet duke eleminuar zanoren fundore *si –a, -e, -ë, -i, -o, -u*, psh, fjala *mbivlerë* modifikohet nga rregulli i parë duke marrë formën vlerë, megjithatë mund të modifikohet edhe më tej duke eleminuar zanoren fundore. Fjalë si *dera*,

hëna, punë, mësuese, djalë, vajzë, libri, teli, djalo, biro, etj, pavarësisht se nuk preken nga rregulli i parë mund të modifikohen duke hequr zanoren fundore.

4. Heqja e prapashtesave: Një bashkësi e plotë e rregullave të përdorura jepet në Aneksin A. Eleminimi i prapashtesave është i ndarë në dy hapa, meqënëse kemi parapashtesa që përfundojnë me zanore. Këto janë trajtuar në hapin 2.

5. Eleminimi i zanores: pas hapit 4, mbeten disa fjalë që përfundojnë me zanore. Si shembull marrim fjalën *gadishmëri* nga hapi 3 e cila kthehet në formën *gati*. Kjo fjalë mund të proçesohet me tej duke fshirë zanore *–a* dhe duke marrë rrënjën *gat*.

5.1.3. Probleme në Gjetjen e Rrënjës së Fjalëve në Gjuhën Shqipe

Gjuha Shqipe bën pjesë në grupin e gjuhëve Indo-Europiane dhe ka rregulla gramatikore dhe strukturore të veçanta. Rrjedhimisht, duhen më shumë përpjekje per të parashikuar një metodë për nxjerrjen e rrënjëve të fjalëve që të përfshijë edhe rastet e veçanta. Këto rastë të veçanta në shumicën e rasteve lidhen me fjalët e përbëra kryesisht nga emrat dhe mbiemrat.

- **Rastet e veçanta për emrat**

1. Fjalë që përbëhen nga dy emra

   a) Emër + emër ndajfoljor i formuar me prapashtesën –ës. Psh. *bukë+pjek-ës, gur+skalit-ës, rroba+qep-ës, etj*.

   b) Emër + emër ndajfoljor i formuar me prapashtesat –je ose –im. Psh. *gjak+derdh-je, besë+lidh-je, letër+këmb-im, dëm+shërbl-im, etj*.

   c) Emër + emër. Psh. *ditë+lindje, mes+dit, vend+banim, vaj+guri, qymyr+guri, pikë+pamje, pikë+pjetje, rrugë+dalje, etj*.

2. Fjalë që përbëhen nga një ndajfolje dhe një emër

Shembull: *bashkë+bisedim, bashkë+luftëtar, drejt+shkrim, keq+kuptim, etj*.

3. Fjalë që përbëhen nga një përemër dhe në emër

Shembull: *vetë+vendosje, vetë+shërbim, dy+luftim, tre+mujor, etj*.

4. Fjalë që përbëhen nga një emër dhe një ndajfolje

Shembull: *gushë+kuq, lulë+kuqe, gojë+mbërthyeri, etj*.

5. Fjalë që formohen nga një folje dhe një emër

Shembull: *thith+lopë, frymë+marrje, etj*.

Rastet 2 dhe 3 mësipër zgjidhen nëpërmjet heqjes së parashtesave. Për shembull, heqja e parashtesës vetë dhe bashkë zgjidh formën e përbërë të emrit.

Rastet 1, 4 dhe 5 përdorin një bashkësi rregullash që përmbajnë emrat me produktive për fjalëformimin e përbërë në gjuhën Shqipe, si edhe pjesët e përbëra më të shpeshta si *hedhës, pjekës, bërës,* etj (Gjati, 2013). Për shkak të formimit të tyre është e vështirë që të krijohen rregulla për çdo emër në gjuhën Shqipe, pavarësisht se në këtë moduk janë trajtuar shumica e emrave më produktive dhe të përdorshëm.

- **Rastet e veçanta për mbiemrat**

1. Mbiemër i përbëra me dy rrënjë emërore

Shembull: *hundë+shkabë, zemër+lepur, zemër+gur, kokë+kungull, etj*.

2. Mbiemër i përbërë me një rrënjë emërore dhe një rrënjë mbiemërore

Shembull: *bark+gjerë, bel+hollë, kokë+prerë, etj*.

3. Mbiemër i përbërë më dy rrënjë mbiemërore

Shembull: *elektro+magnetik, elektro+mekanik, gjermano+lindor, etj*.

4. Mbiemër i përbërë me një numërues dhe një mbiemër

Shembull: *një+mujor, treqind+vjeçar, shumë+vjeçar, disa+ditor, etj*.

5. Mbiemër i përbërë me një rrënjë të pjesshme dhe një rrënjë ndajfoljore

Shembull: *dasha+mir, dasha+keq , etj*.

Njëlloj si në rastin e rregullave për emrat, edhe këtu është përdorur një bashkësi rregullash të bazuara në pjesën e parë dhe të dytë të fjalës së përbërë.

- **Krijimi i ndajfoljeve**

Shumica e ndajfoljeve formohen duke bashkuar dy ose më shumë pjesë të ligjëratës. Bashkësia më e rëndësishme e ndajfoljeve formohet nga përemrat.

1. Ndajfoljet e përbëra ky elementi i parë është ndajfolja, psh *gjithë*, dhe elementi i dytë mund të jetë një ndajfolje përemërore, një emër ose një pjesë tjetër e ligjëratës, psh. *gjith+ashtu, gjith+herë, gjith+një, etj*. Ky rast trajtohet duke aplikuar heqjen e parashtesave te fjalës, psh heqja e parashtesës *gjithë* dhe *gjith* redukton ndajfoljet e përbëra.

2. Ndajfoljet e përbëra përemërore ku elementi i parë i bashkimit është një nga përemrat pyetës, *kur, ku, nga, si, qysh* dhe elementi i dytë është folja *do*, psh. *ku+do, kur+do, si+do, nga+do, sa+do*, etj.

3. Ndajfoljet e përbëra ku elementi i dytë është ndajfolje përemërore *ku, kund/kundi* dhe elementi i parë mund të jetë një përemër, ndajfolje ose pjesore, psh. *asgjë+kundi, as+kund, tjetër+kund*, etj.

4. Ndajfoljet e përbëra ku elementi i dytë është emri *herës* dhe elementi i parë mund të jetë një përemër, ndajfolje ose pjesore. Psh. *asnjë+herë, atë+herë, nga+herë, për+herë*, etj. Këto rregulla nuk janë trajtuar mëqënëse në përgjithësi nuk janë shumë produktive në krijimin e fjalëve të reja. Për më tepër nuk ka një rregull të përgjithshëm me të cilin formohen këto fjalë.

5. Ndajfoljet e formuara nga një parafjalë dhe një emër. Psh. *për+ballë, për +bri, për +krah, për +ditë*, etj. Ky rast trajtohet me anë të heqjes së parafjalëve. Psh. heqja e parashtesës për redukton ndajfoljet e përbëra.

6. Ndajfoljet e formuara nga dy emra të lidhur ose jo me anë të një zanoreje lidhëse. Psh. *anekënd, fytafyt, kryekëput, shpeshherë*, etj. Edhe ky është një rast për të cilin nuk mund të përdoren rregulla të fjalëformimit.

5.1.4. Gjetja e Rrënjëve të Fjalëve të Përbëra

Për të trajtuar problemin e gjetjes së rrënjëve të fjalëve të përbëra është zgjedhur të përdoret përsëri një model i bazuar në rregulla i cili i bashknagjitet rregullave për gjetjen e rrënjëve të fjalëve të thjeshta.

Për këtë janë përdorur rregullat për tre grupe të fjalëve të përbëra si më poshtë: (Gjati, 2013).

Në grupin e parë përfshihen fjalët të cilat përdorin një vijë ndarëse për të identifikuar përbërjen e dy fjalëve. Psh. *ekonomiko-shoqëror, politiko-ekonomik*, etj. Këto fjalë janë shumë të zakonshme në gjuhën Shqipe. Sipas "-" fjala ndahet në dy fjalë përbërëse të cilat trajtohen si fjalë të thjeshta për të nxjerrë rrënjët e tyre.

Grupi i dytë i fjalëve të përbëra, i cili trajtohet nga algoritmi janë ato fjalë të cilat formohen me anë të parafjalëve. Rastet e marra parasysh janë:

"shumë", "poshtë", "bashk", "bashkë", "jashtë", "kundër", "gjysmë", "gjysm", "gjithë", "gjith", "sipër", "prapa", "drejt", "mirë", "para", "anti", "sapo", "vetë", "krye","tej", "keq", "akt", "arki", "plot", "brenda", "pas", "nën", "pas","pa", "mbi", "mos", "ndër", "porsa", "disa", "hidro", "aero", "antropo", "biblio", "bio", "deci", "kilo", "hekto", "pseudo", "neo", "poli", "gjeo", "filo", "mono", "mikro", "mili", "astro", "foto", "auto", "avio", "centi", "kino", "makro", "mega", "termo", "psiko", "super", "ultra", "radio", "elektro".

Gjithashtu, siç shihet më sipër, në të njëjtin grup fjalësh të përbëra, janë përfshirë edhe fjalët e përbëra të formuara nga parafjalë ndërkombëtare si: hidro, bio, aero, biblio, deci, kilo, etj. Edhe këto janë shumë të shpeshta në fjalëformimin në gjuhën Shqipe. Pasi fjala e përbërë ndahet në dy pjesë, parafjala hiqet duke mbajtur vetëm rrënjën e fjalës. Për këtë grup janë përdorur rreth 90 rregulla të fjalëformimit.

Grupi i tretë trajton fjalët e përbëra nga një numërues. Janë marrë në konsideratë numëruesit më të zakonshëm në fjalëformim në Shqip. Algoritmi e ndan fjalën e përbërë në dy fjalë të thjeshta të cilat trajtohen në mënyre individuale.

Grupi i katërt merr në konsideratë fjalët e përbëra të cilat si pjesë të dytë kanë një nga fjalë të zakonshme si: *bukëpjekës, samarpunues, orëndreqës, gurskalitës* për të cilat nuk mund të gjenerohen rregulla fjalëformimi. Çdo fjalë e përbërë që përfundon me njërën prej këtyre fjalëve do të dekompozohet dhe do të përdoret për nxjerrjen e rrënjëve të fjalëve. Lista e fjalëformuesve më të shpeshtë të këtij tipi jepet më poshtë.

punues", "bërës", "dashës", "mbajtës", "ndreqës", "shitës", "vrasës", "ngjirur", "sjellës", "rritës", "hedhës", "pastrues", "skalitës", "gdhendës", "matës", "shpues", "vjeçar", "ngrënës", "çjerr", "qepës", "dashuri", "ngrohës", "ftohës", "nxjerrës", "madh", "dhënës", "mbartës", "dhënje", "artë", "fyer", "vajt", "plotë", "pak", "muajsh", "vogël", "hollë", "marrës", "mbytës", "lidhës", "vënës", "shkelës", "dashje", "vend", "det", "shtypës", "pagues", "ngritës", "larës", "thithës", "gjerë", "ngusht", "keq", "mbledh", "pjekës", "kuptim", "ëmbël" , "lartë", "blerës", "mirë", "shpërblim", "pritës", "prurës", "dhe", "javë", "gur", "jetë", "natë", "lind", "ngjyr", "verior", "gjatë", "kuq", "gjilpër", "kryq", "lugë", "gjysh", "trash", "verdh", "beton", "udhë", "tharë", "vrar", "akuzë", "jug", "shkurt", "hapur", "ngjitur", "shtrëmbër", "bardh", "zez", "humb", "prer", "shuar", "mal", "fryr", "zbraz", "pamës", "dalë", "grisur", "besim", "plasur", "hyrje", "mendje", "lopë", "rëndë", "zi", "qarkull", "kalim", "spec", "varur", "shkrim", "shkruaj", "komand", "shpejt", "lënduar", "bosh", "shpuar", "thatë", "sulm", "synim", "synues", "ngrir", "ngjesh", "pushim", "bukur", "derr", "mace", "majmun", "qen", "cjap", "butë", "fortë", "heqës", "dhjet", "adhurim", "asgjës", "drejt", "shënim", "lakuar", "mbrojtës", "thyer", "thinj", "shtrembër", "brejtës", "vrasje", "mekanik", "magnet", "teknik", "dredhur", "shqiptim", "tingëll", "ndezës", "mbjellës", "zbulues", "shumë", "kërkues", "lëshues", "edukim", "edukohem", "kundër", "perëndim".

Më poshtë jepet një paraqitje grafike e rregullave të fjalëve të përbëra për secilin grup. Rendi i trajtimit të tyre në algoritëm është ai i ilustruar në Figurën 5.4.

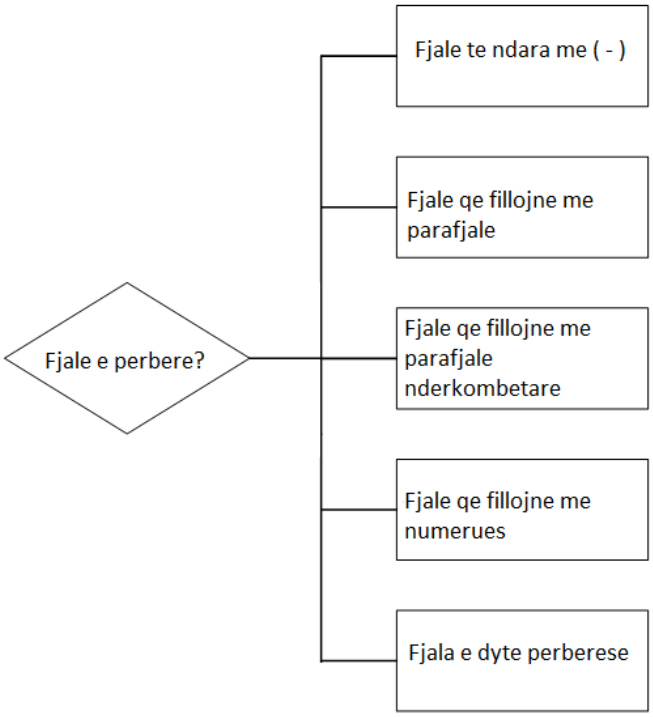

Figura 5.4. Grupet e fjalëve të përbëra (Gjati, 2013).

5.2. Komponenti Statistikor

Komponenti statistikor që do të përdorim synon të përdorë një metodë statistikore të peshave të fjalëve që përshkruajnë dokumentin. Funksioni i peshave që do të përdorim e vë fokusin në metodën TFIDF – Term Frequency Inverse Document Frequency, e cila është e lidhur ngushtë me shumë metoda të machine learning. Bazat matematike dhe analizën e kësaj metode e gjejmë më të detajuar në çështjen 3.3.3 të kësaj teze. Llogaritja e vlerës TFIDF për çdo fjalë të dokumentit tekst kalon në dy faza: Së pari, llogaritet TF, frekuenca e çdo fjale në dokument duke patur parasysh që sa më e lartë të jetë frekuenca e një fjale, aq më shume peshë ka ajo fjalë në atë dokument. Së dyti, llogaritet njësia IDF - inverse document frequency, e cila mat frekuencën e fjalës së dokumentit në një korpus të dhënë. Në këtë mënyrë nëqoftëse, një fjalë në dokument ka frekuencë të lartë në korpus, atëherë ajo nuk ka peshë diskriminuese për atë dokument.

Si input në këtë komponent do të shërbejë dokumenti që përmban vetëm rrënjët e fjalëve në mënyrë që frekuencat e llogaritura të përfshijnë të gjitha format e shfaqjes së një fjale në dokument, duke i rritur kuptimin vlerës që merr TFIDF-ja.

5.3. Komponenti i Machine Learning

Komponeti i Machine Learning do të përdoret për të zgjidhur problemin e klasifikimit të një dokumenti në një klasë në bazë të tematikës së dokumentit. Kjo kërkon një korpus fillestar të

organizuar në klasa, i cili të perdoret për trajnimin paraprak të modelit. Për t'i dhënë zgjidhje problemit të klasifikimit, komponenti ynë i Machine Learning do të bazohet në klasifikatorin Naive Bayes i cili ben pjesë në grupin e klasifikatorëve probabilitarë. Pavarësisht së efiçensa e tij shpesh herë tejkalohet nga teknika të tjera si Boosted Trees, Max Entropy, Support Vector Machines, etj, Naive Bayes është shumë efiçent nga ana komputacionale në terma të CPU-së dhe memories dhe kërkon një bashkësi të vogël të dhënash për trajnim. Koha e nevojshme për trajnim në Naive Bayes është ndjeshëm më e vogël krahasuar me metoda të tjera alternative. Në problemin e gjetjes së fjalëve kyçe është shumë e rëndësishme përformanca në llogaritje sepse kërkimi i informacionit nëpërmjet fjalëve kyçe duhet të ekzekutohet në kohë reale. Nga ana tjetër, na intereson që saktësia e klasifikatorit të arrijë të paktën një efiçensë të pranueshme si modul i ndermjetëm që ka për qëllim përmirësimin e saktësisë së zgjidhjes. Ja vlen të theksojmë që në shumë raste efiçensa e Naive Bayes është shumë afër me atë të teknikave më të komplikuara dhe më të ngadalta (Huang, J. 2003).

Ekzistojnë disa variante të Naive Bayes, nga të cilat ne kemi zgjedhur Naive Bayes Multinomial sepse ai përdor frekuencavet e fjalëve në një dokument, element shumë i rëndësishëm ky në problemet e klasifikimit, si pasojë është shumë e zakonshme që ky lloj klasifikatori të përdoret për klasifikim sipas tematikës.

Në klasifikatorin që përdor Naive Bayes karakteristikat e përdorura supozohen të pavarura nga njëra tjetra. Pavarësisht se ky supozim zakonisht nuk është i vërtetë, studimet e bëra me klasifikimin Bayesian tregojnë që ka disa arsye teorike të efiçensës së këtij klasifikatori. Optimaliteti dhe efiçensa e klasifikimit Bayesian është provuar matematikisht duke përdorur raste dokumentash me karakteristika që nuk rezultojnë të pavarura por që eleminojnë njëra tjetrën duke mos cënuar veçoritë e klasifikatorit (Zhang, 2004). Pavarësisht se probabilitetet e llogaritura nga Naive Bayes kanë vlera të ulta saktësie, vendimet në lidhje me klasifikimin janë relativisht të mira. Naive Bayes i rrit artificialisht vlerat e probabilitetit të klasës, megjithatë nëqoftëse ne e përdorim klasifikatorin vetëm për të marrë një vendim dhe jo për të parashikuar probabilitete, atëherë vendimarrja është e saktë (Manning et al, 2009).

Në klasifikatorin e teksteve ne përdorim fjalët/rrënjët e fjalëve të dokumentit për ta klasifikuar atë në klasën e duhur. Duke përdorur rregullin "maximum a posteriori" arrijmë në klasifikatorin e mëposhtëm:

$$c_{map} = \arg\max_{c \in C} \left(P(c \mid d) \right) = \arg\max_{c \in C} \left(P(c) \prod_{1 \le k \le n_d} P(t_k \mid c) \right)$$

Ku t_k janë fjalët e dokumentit, C është bashkësia e klasave të përdorura në klasifikim, P(c|d) është probabiliteti me kusht i klasës c nëqoftëse jepet dokumenti d, P(c) është probabiliteti paraprak i klasës c dhe $P(t_k|c)$ është probabiliteti me kusht ë fjalës t_k nëqoftëse jepet klasa c.

Në mënyrë që të gjejmë se në cilën klasë do ta klasifikojmë një dokument të ri, duhet të llogarisim prodhimin e probabilitetit të secilës fjalë të dokumentit nëqoftëse jepet një klasë, shumëzuar me probabilitetin e mëparshëm të asaj klase. Pasi e bëjmë këtë për çdo klasë të bashkësisë C, zgjedhim atë me probabilitetin më të lartë.

Veprime të tilla do të gjeneronin numra kaq të vegjël, sa fizikisht do të ishte e pamundur të ruheshin në kujtesë , si rrjedhim do të merrnim një numër 0 duke e humbur dobinë e analizës sonë. Për të eleminuar këtë problem, në vend të maksimizimit të prodhimit të probabiliteteve maksimizohet shuma e logaritmave të tyre (Manning et al, 2009).

$$c_{map} = \arg\max_{c \in C} \left(\log P(c) + \sum_{1 \le k \le n_d} \log P(t_k \mid c) \right)$$

Në këtë mënyrë, në vend që të zgjedhim klasën me probabilitetin më të lartë zgjidhet ajo me vlerën e logaritmit më të lartë. Mëqënëse funksioni logaritmik është monoton, vendimi i MAP mbetet i njëjtë. Një problem tjetër që lind pas modelit të fundit, është rasti kur një fjalë e caktuar nuk ndeshet fare në një klasë të caktuar. Probabiliteti me kusht për të do të ishte 0, ndërsa sipas shumës së logatimave log(0) është një vlerë e papërcaktuar. Për të marrë parasysh këtë rast të veçantë, shtojmë vlerën 1 në çdo rast.

$$P(t \mid c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'}) + B'}$$

Ku B është numri i termave në fjalorin V.

Naive Bayes Multinomial vlerëson probabilitetin me kusht të një fjale kur jepet klasa duke llogaritur frekuencën relative të termave t në dokumentat që i përkasin klasës c:

$$P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

Në këtë mënyrë marrim parasysh numrin e herëve që një term *t* ndeshet në dokumentat e trajnimit të klasës c. Agoritmat e trajnimit dhe të testimit jepen në Figurën 5.5, (Manning et al, 2009).

```
TrainMultinomialNB(C, D)
V ← ExtractVocabulary(D)
N ← CountDocs(D)
for each c ∈ C
do N_c ← CountDocsInClass(D, c)
    prior[c] ← N_c / N
    text_c ← ConcatenateTextOfAllDocsInClass(D, c)
    for each t ∈ V
    do T_ct ← CountTokensOfTerm(text_c, t)
    for each t ∈ V
    do condprob[t][c] ← (T_ct + 1) / (Σ_t' (T_ct' + 1))
return V, prior, condprob

ApplyMultinomialNB(C, V, prior, condprob, d)
W ← ExtractTokensFromDoc(V, d)
for each c ∈ C
do score[c] ← log prior[c]
    for each t ∈ W
    do score[c] += log condprob[t][c]
return arg max_{c∈C} score[c]
```

Figura 5.5. Algoritmi Naive Bayes Multinomial (Manning et al, 2009).

5.4. Projektimi i Algoritmit

Struktura e algoritmit është e organizuar në dy faza kryesore sipas skemave në Figurën 5.6 dhe Figurën 5.7. Në fazën e parë përdoret si input një bashkësi me dokumenta të kategorizuar në klasa të paracaktura. Modeli gjenerohet vetëm një herë dhe përdoret për çdo test.

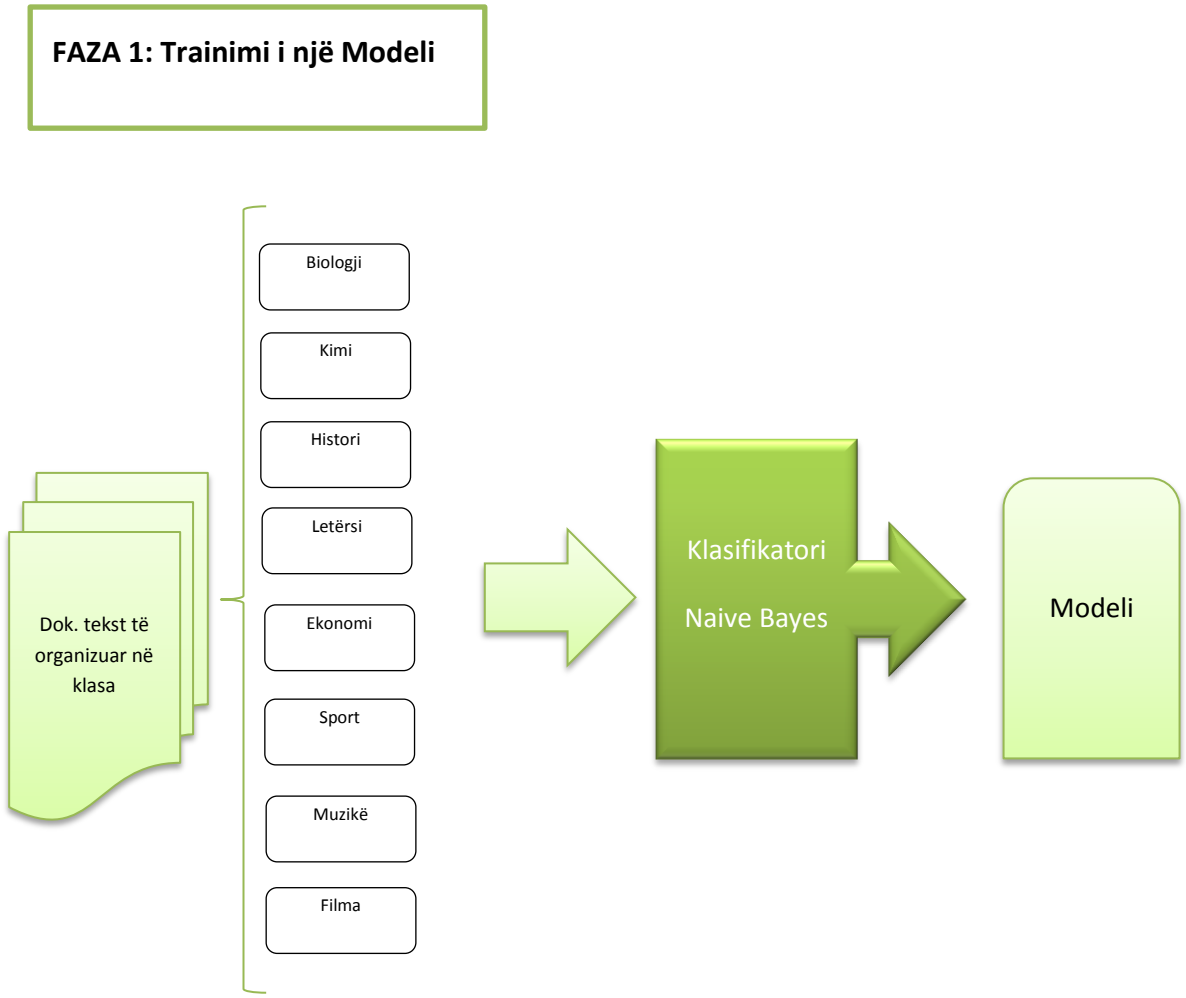

Figura 5.6. Trainimi i klasifikatorit për të gjeneruar modelin

Faza e dytë lidhet me gjetjen e fjalëve kyçe të një dokumenti i cili nuk ben pjesë në bashkësinë e dokumentave të përdorur për testim, por i përket njërës prej klasave me të cilat është trajnuar modeli. Komponenti Linguistik e paraproçeson dokumentin duke shrytëzuar elementë që varen nga gjuha si heqja e lidhësave dhe kthimi i termave në rrënjë të fjalës.

Nga komponenti statistikor dalin të gjithë termat të peshuar me anë të peshës statistikore TFIDF. Duke patur parasysh faktin që termat me vlerë të ulët të tfidf kanë probabilitet të vogël për të qënë fjalë kyçe të vlefshme, marrim në shqyrtim si terma kandidatë 30% të termave me peshë më të lartë statistikore. Roli i klasifikatorit është që të evidentojë termat kandidatë, mungesa e të cilëve në dokument e klasifikon atë në një klasë tjetër.

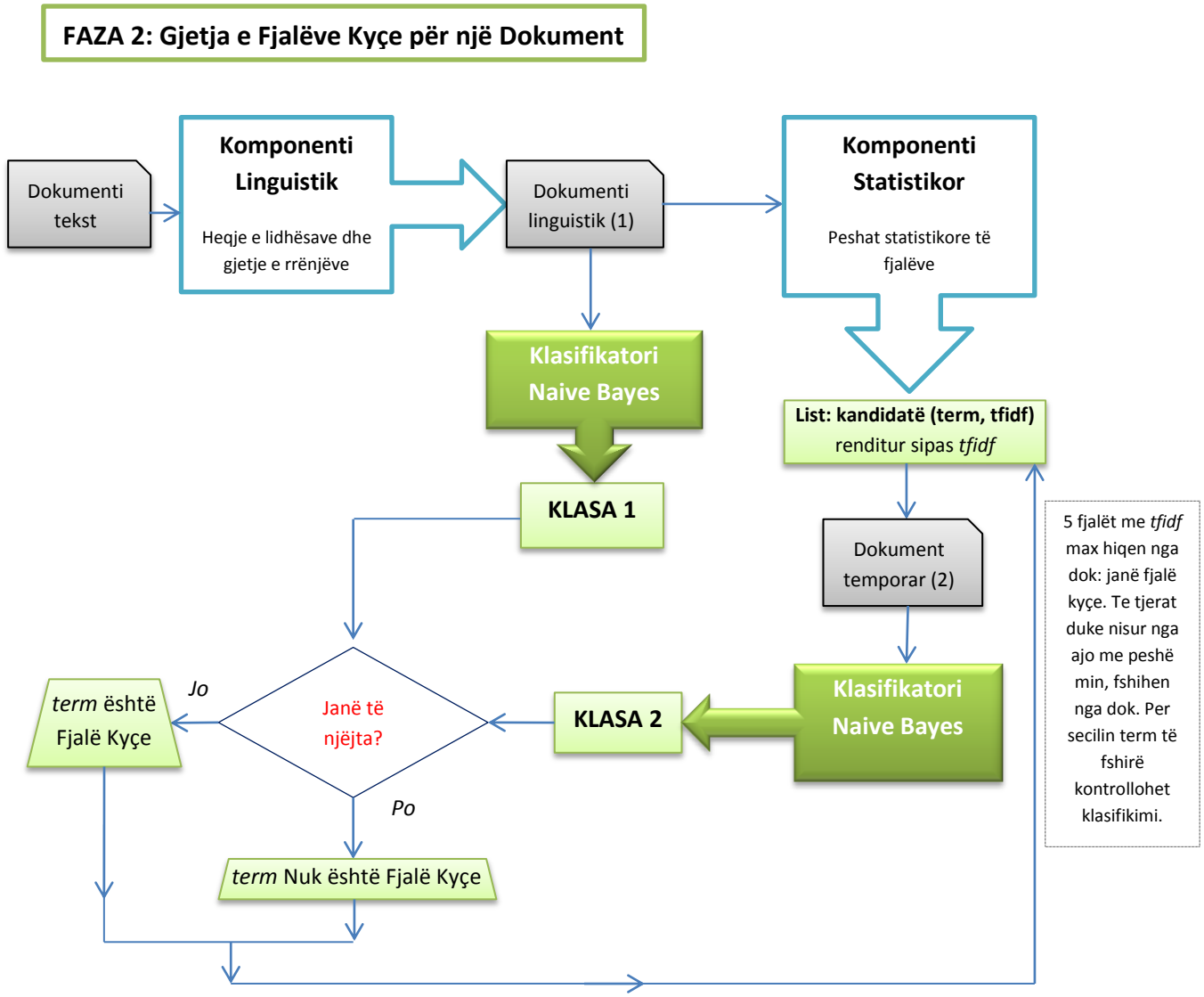

Figura 5.7. Skema e algoritmit të gjetjes së fjalëve kyçe të një dokumenti

5.5. Detaje të Implementimit

Për të implementuar një program sipas algoritmit të propozuar në sesionin 5.4 është përdorur *Java NetBeans IDE 7.4*. Meqënëse, në përgjithësi të dhënat tekst janë të veshtira për t'u përdorur si objekt i llogaritjeve, është përdor një bazë të dhënash në MySQL Server për të ruajtur për çdo dokument trajnimi, termat përbërës të tij së bashku me vlerat e peshave.

Në mënyrë që modeli i gjeneruar në fazën e parë, të jetë sa më optimal, është shtuar edhe në fazën e parë moduli i paraproçesimit të dokumentave me anë të komponentit linguistik para se dokumentat të përdoren për trajnim nga Klasifikatori Naive Bayes. Të dhënat linguistike dhe statistikore për secilin dokument janë ruajtur në një bazë të dhënash si në Figurën 5.8.

Komponenti Linguistik dhe Komponenti Machine Learning janë implementuar në Java, ndërsa Komponenti Statistikor është ndërtuar duke përdorur store procedure dhe disa funksione në MySQL të cilat manipulojnë tabelën e termave. Më poshtë po japim një paraqitje të thjeshtuar të skemës së bazës së të dhënave.

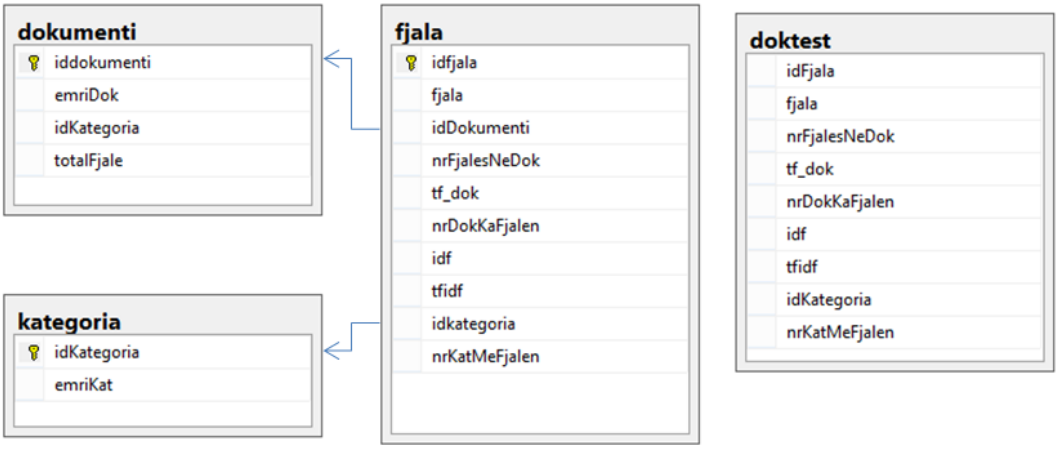

Figura 5.8. Baza e të dhënave MySQL për dokumentat e organizuar sipas peshave statistikore

6. Vlerësimi Eksperimental

Saktësia dhe efikasiteti i gjetjes së fjalëve kyçe të një dokumenti nuk varet nga fuqia llogaritëse, pavarësisht se ajo do të ndikonte në performancë të mirë në rastin e proçesimit të një sasie të madhe teskti.

6.1. Korpusi i Dokumentave

Dokumentat e përdorur në këtë studim janë dokumenta tekst të ndarë në dy grupe të mëdha: dokumentat që përdoren për trajnimin e modelit dhe dokumentat që përdoren për testimin e algoritmit të gjetjes së fjalëve kyçe. Çdo dokument i përdorur nga ne është zgjedhur që të përmbajë mesatarisht rreth 1000 fjalë. Theksojmë këtu që përmbajtja e dokumentave është tekst i pastër, pa titull dhe pa nënçështje të etiketuara. Në rast se teksti do të përmbante tituj dhe çështje të etiketuara, ato do të shërbenin si kandidatë shumë të mirë për fjalë kyçe, pa patur nevojë për vlera të larta të peshave statistikore apo aftësi diskriminuese për dokumentin.

Dokumentat që përdoren për trajnimin e modelit janë organizuar në 8 kategori/klasa të ndryshme, Biologji, Ekonomi, Filma, Histori, Kimi, Letërsi, Muzikë, Sport si paraqitet në figurën 6.1, ku çdo klasë përmban 50 dokumenta të ndryshëm.

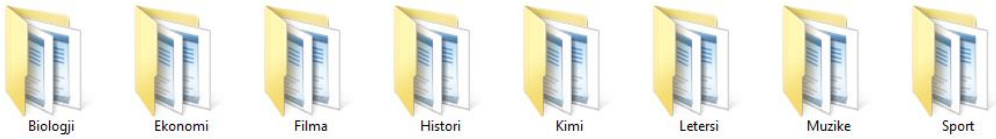

Figura 6.1. Dokumentat e trajnimit të organizuar në klasa

Dokumentat që përdoren për testimin e algoritmit të gjetjes së fjalëve kyçe janë dokumenta të rinj, të cilët i përkasin klasave të paracaktuara por nuk janë përdorur më parë për trajnimin e modelit. Janë përdorur gjithsej 160 dokumenta testues, 20 dokumenta për secilën klasë.

6.2. Metoda e Vlerësimit

Qëllimi kryesor i këtij punimi është prezantimi i një algoritmi të nxjerrjes së fjalëve kyçe në tekste në gjuhën Shqipe, i cili tejkalon performancën e metodave statistikore duke shtuar elementë të Machine Learning. Pavarësisht se llogaritja e performancës së sistemeve të tilla është një proçes i paformalizuar dhe i vështirë, ne do përpiqemi të përshtasim disa teste për qëllimin tonë duke pranuar limitet e këtyre vlerësimeve. Në rastin e aplikacioneve të nxjerrjes së fjalëve kyçe, saktësia, (angl.precision), zakonisht matet me raportin e fjalëve kyçe të gjetura nga sistemi që janë të përshtatshme për dokumentin, me numrin total të fjalëve kyçe të pranuara si të tilla. Fjalë kyçe të përshtatshme do të quhen ato fjalë kyçe që janë caktuar manualisht nga autorët e dokumentave

dhe që me marrëveshje do t'i pranojmë si të përshtatshme. Në këtë marrëveshje duhet të theksojmë që zgjedhja e fjalëve kyçe për një dokument varet nga konteksti i përdorimit të tij, prandaj mund të ketë mospërputhje edhe midis fjalëve kyçe që caktohen nga dy persona të ndryshëm si të përshtatshëm për të njëjtin dokument.

Në mënyrë që të arrijmë të evidentojmë një nivel tolerance të pranueshëm për përputhshmërinë e fjalëve kyçe të nxjerra automatikisht nga sistemi ynë, me fjalët kyçe të vendosura nga autori i dokumentit, kemi menduar të zhvillojmë një test i cili vlerëson nivelin e përputhshmërisë së dy bashkësive me fjalë kyçe të caktuara nga dy profesionistë të ndryshëm për të njëjtin dokument. Këtë test do ta quajmë "*Test i tolerancës ndaj fjalëve kyçe të vendosura nga autori*". Për secilën klasë/kategori janë zgjedhur dy profesionistë të ndryshëm të fushës, të pavarur nga njëri tjetri. Secilit i është kërkuar të caktojë rreth 10 fjalë kyçe për secilin dokument nga grupi i dokumentave për testim. Në këtë mënyrë, për çdo dokument testimi kemi dy bashkësi të ndryshme me fjalë kyçe. Meqënëse, cilësinë e algoritmit tonë do ta masim nëpërmjet një grupi fjalësh kyçe të caktuara si të përshtatshme nga një profesionist, testi i parë do të lidhet me vlerësimin sasior të subjektvitetit të fjalëve kyçe të vendosura nga një profesionist/ekspert duke matur variancën midis fjalëve kyçe të vendosura nga dy profesionistë të ndryshëm. Për këtë ne kemi llogaritur vlerën e koefiçientit kappa të Cohenit (Cohen 1960), (Carletta,1996)**.** Ky koefiçient përdoret shpesh për të matur shkallën e homogjenitetit/konsensusit midis ekspertëve që caktojnë manualisht fjalët kyçe. Ai merr parasysh edhe homogjenitetin e rastësishëm. Koeficineti kappa llogaritet si në vijim:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

ku *Pr(a)* është konsensusi relativ midis ekspertëve mbi një fjalë kyçe, dhe *Pr(e)* është probabiliteti hipotetik i konsensusit të rastësishëm. Nëqoftëse ekspertët janë në konkordancë të plotë me njëri tjetrin atëherë $\kappa = 1$ dhe nëqoftëse përveç homogjenitetit rastësor, nuk ka konkordancë tjetër midis ekspertëve, atëherë $\kappa = 0$. Detajet e implementimit dhe rezultatet e këtij testi do t'i gjeni në sesionin 6.3.1.

Rezultatet e testit të parë do të na ndihmojnë të vlerësojmë më objektivisht përputhmërinë e fjalëve kyçe të nxjerra automatikisht me fjalët kyçe të nxjerra manualisht si të përshtatshme. Për të matur këtë përputhshmëri kemi ndërtuar një test që po e quajmë "*Testi i cilësisë së fjalëve kyçe automatike*". Si standard për këtë vlerësim do të përdorim njërin nga grupet me fjalë kyçe të vendosura manualisht nga një profesionist i fushës. Matja sasiore e cilësisë së fjalëve kyçe automatike do të realizohet, ashtu si edhe në testin e parë, nëpërmjet *precision* dhe *recall.*

Një element i veçantë në algoritmin tonë të nxjerrjes së fjalëve kyçe është përdorimi i klasifikatorit Naive Bayes. Siç kemi shpjeguar në çeshtjet e mësipërme, klasifikatori ka nevojë për një korpus trajnimi. Testi i rradhës ka për qëllim të vlerësojë ndikimin e madhësisë së korpusit trajnues mbi përputhshmërinë e fjalëve kyçe të gjetura me standardin e zgjedhur (fjalët kyçe të

vendosura manualisht nga një profesionist). Këtë test do ta quajmë "*Testi i ndikimit të madhësisë së korpusit trajnues*".

Metoda e propozuar nga ne për nxjerrjen ë fjalëve kyçe kombinon modelet statistikore me modelet e Machine Learning. Për të llogaritur shkallën e ndikimit të përdorimit të metodave të Machine Learning ndaj modelit të thjeshtë statistikor kemi vendosur të realizojmë një test tjetër, "*Testi i vlerësimit të shkallës së përmirësimit ndaj modelit statistikor*" i cili ka për qëllim të masë përmirësimin që sjell përdorimi i klasifikatorit Naive Bayes. Edhe në këtë rast do të përdoret precision dhe recall për të matur cilësinë e arritur nga një model i thjeshtë statistikor, si *TFIDF*.

Meqënëse është e pritshme që një pjesë e fjalëve kyçe të llogaritura nga algoritmi ynë automatik, do të kenë peshë statistikore të lartë, është e arsyeshme të përcaktojmë sa përqind e fjalëve kyçe të nxjerra nga algoritmi ynë, nuk janë në grupin e fjalëve me vlerë të lartë të tfidf, por megjithatë konsiderohen si përcaktuese për dokumentin. Ky do të përbëjë testin që po e quajmë "*Testi i fjalëve kyçe me peshë statistikore të ulët*".

Meqënëse vlerësimi i efiçensës së algoritmave në GjIK është përgjithësisht një proçes i vështirë, do të kryejmë një test që po e quajmë "*Testi i vlerësimit manual të fjalëve kyçe*", me të cilin synojmë të përdorim ekspertë njerëzorë për të vlerësuar fjalët kyçe të gjeneruara nga algoritmi automatik si pozitive ose negative.

Në sesionin e mëposhtëm do të trajtojmë në detaje implementimin e testeve të sipërpërmendura duke analizuar dhe vlerësuar rezultatet.

6.3. Rezultatet

Algoritmi automatik i nxjerrjes së fjalëve kyçe u ekzekutua mbi një korpus me 160 dokumenta tekst të ndarë në 8 klasa. U nxorrën në total 1125 falë kyçe, pra mesatarisht 7.03 fjalë kyçe për çdo dokument. Çdo dokument është i shoqëruar me dy bashkësi me rreth 10 fjalë kyçe, të cilësuara manualisht si të përshtatshme nga dy ekspertët e zgjedhur të fushës. Në tabelën e mëposhtme paraqiten në mënyrë sasiore të përmbledhur të dhënat e korpusit të ndara në klasa, përkatësisht për grupin e dokumentave të testimit dhe të trajnimit. Dokumentat testues u përdorën si input për sistemin tonë të nxjerrjes automatike të fjalëve kyçe. Numri i fjalëve kyçe të gjeneruara për çdo kategori paraqitet në Tabelën 6.1.

Tabela 6.1. *Përmbledhje e të dhënave të korpusit*

Klasa	NrDok. Train	Fjalë/ dokTrain	NrDok. Test	Fjalë/ dokTest	FK.automatike/ Klase	Nr.FK Eksperti nr.1	Nr.FK Eksperti nr.2
Biologji	50	1870	20	2038	158	200	203
Ekonomi	50	859	20	941	123	198	201
Filma	50	615	20	818	107	200	200
Histori	50	1740	20	1944	203	219	211
Kimi	50	1564	20	1716	124	208	200
Letërsi	50	1361	20	2003	181	205	220
Muzikë	50	533	20	625	115	200	200
Sport	50	720	20	954	124	200	200
Gjithësej	**400** dok	**1157.75** fjalë/dokTrain	**160** dok	**1379.875** fjalë/dokTest	**7.0938** fk/dok	**10.1875** fk/dok	**10.21875** fk/dok

Në çështjet e mëposhtme po diskutojmë mbi rezultatet e secilit prej testeve të zhvilluara mbi këto të dhëna.

6.3.1. Test i Tolerancës ndaj Fjalëve Kyçe të Vendosura nga Autori

Qëllimi i këtij testi është përcaktimi i shkallës së përputhshmërisë se fjalëve kyçe të vendosura manualisht nga dy ekspertë të ndryshëm. Meqënëse fjalët kyçe të caktuara nga ekspertët pranohen si të përshtatshme për dokumentin dhe do t'i përdorim si standard për krahasimin me fjalët kyçe të nxjerra automatikisht nga algoritmi ynë, duam të vlerësojmë paraprakisht cila do të ishte një vlerë e pranueshme për përputhshmërinë e fjalëve kyçe automatike me fjalët kyçe të vendosura nga ekspertët.

Tabela 6.2. *Fjalët kyçe të nxjerra manualisht nga dy ekspertë të ndryshëm për të njëjtin dokument*

	Fjalët kyçe sipas ekspertëve	
Dokumenti	**Eksperti nr.1**	**Eksperti nr.2**
dok1.txt	*pas pavaresisë*	*letërsia pas pavarësisë*
	tregim	*realizmi kritik*
	letërsi realiste	*romani*
	proza	*F.Noli*
	roman	*S.Spase*
	S.Spase	*P.Marko*
	rryme demokratike	*Migjenit*
	Migjeni	*F.Konica*
	kritikë letrare	*M.Kuteli*
	publicistikë	*L.Poradeci*

Pjesë nga përmbajtja e *dok1.txt*

Me çlirimin e Shqipërisë dhe krijimin e shtetit shqiptar, letërsia do të hynte në një periudhë të re zhvillimi (1912-1939). Traditat e shquara në lëmin e poezisë do të lëvronin më tej nëpërmjet krijimtarisë së pasur të poetit të shquar Gjergj Fishta dhe të poetëve të tjerë si N. Mjeda, F. Noli, L. Poradeci, A. Asllani, V. Prenushi etj. Letërsia do të shënonte një etapë të re me lindjen e romanit shqiptar në dekadën e dytë të këtij shekulli me romanet e N. Nikajt, F. Postolit, Z. Arapit dhe më pas me ato të H. Stërmillit e S. Spasses.
Nga ana tjetër, në lëmin e prozës do të ndjehet një ngritje e nivelit të tregimit, e cila do të kualifikohet artistikisht nga M. Kuteli dhe E. Koliqi. Gjatë periudhës së Pavarësisë, tema patriotike nuk do të jetë më në qendër të letërsisë. Problemet e mprehta shoqërore do të trajtohen nga brezi i letrarëve realistë, ku përmendim poetin dhe prozatorin e talentuar Migjeni, i cili la gjurmë të thella dhe ndikoi në rrjedhat e letërsisë moderne shqiptare.
Në këtë drejtim mund të përmendim letrarët e rinj si P. Marko, N. Bulka, Dh. Shuteriqi, G. Pali, A. Çaçi etj. Në vitet e çlirimit kombëtar e shoqëror do të zhvillohet letërsia shqipe me tipare të reja, si një letërsi realiste. Edhe pse në kushtet e një shteti monist, letërsia shqipe gjatë afro pesë dekadave (1944-1990) njohu pjekurinë e saj, u konsolidua dhe u afirmua si një letërsi me vlera të nivelit evropian. Një moment i rëndësishëm në letërsinë shqiptare në këto vite janë në fushën e romanit, ku mund të përmendim D. Agolli, J. Xoxa, S. Spasse, A, Abdihoxha, F. Gjata etj.
Përfaqësues i shquar i rrymës demokratike është Fan Noli, figurë politike por edhe poet, publicist e përkthyes. Lirik i njohur i kohës është Lasgush Poradeci. Nga ana tjetër një figurë qendrore e letërsisë dhe e publicistikës ishte At Gjergj Fishta. Ndër prozatorët janë romancierët Ndoc Nikaj, Foqion Postoli dhe Haki Stërmilli.
Kulmin e realizmit kritik, me pasqyrimin e gjendjes së masave më të varfër dhe me theks të fortë të revoltës shoqërore e shënojnë veprat poetike dhe prozaike të Migjenit ("Vargjet e lira", 1935). Faik Konica ndërkohë ishte një kritik letrar me veprimtari shumë të zgjeruar publicistike që ndikoi jo pak në këtë letërsi.

Figura 6.2. Një pjesë nga dokumentit që i është dhënë secilit prej ekspertëve

Nga Tabela 6.2. duket që midis dy ekspertëve, të cilët i kanë nxjerrë fjalët kyçe në mënyrë të pavarur nga njëri tjetri, ka diferenca të dukshme në vendosjen e fjalëve kyçe. Të pyetur për këto ndryshime, ekspertët shpjeguan që ato janë rezultat i kontekstit në të cilin e ka parë dokumentin secili prej ekspertëve. Ato shprehën tolerancë ndaj fjalëve kyçe të vendosura nga eksperti tjetër, duke i quajtur ato gjithashtu të pranueshme për dokumentin. Të njëjtën tolerancë ne sugjerojmë ta përdorim edhe për grupin e fjalëve kyçe që ka gjeneruar sistemi ynë. Më poshtë kemi matur në mënyrë sasiorë shkallën e përputhshmërisë se dy grupeve më fjalë kyçe të vendosura për çdo dokument testimi nga ekspertët duke përdorur koeficientin *k* (Carletta,1996).

Nëqoftëse D është dokumenti të cilit dy ekspertët do t'i nxjerrin fjalët kyçe, do të shënojmë F_D numrin e fjalëve të ndryshme kandidate për të qënë fjalë kyçe në dokument. Vlerat me interes për ne janë:

- F1: numri i fjalëve kyçe kandidate të zgjedhura njëkohësisht nga të dy ekspertët.
- F2: numri i fjalëve kyçe kandidate të zgjedhura nga eksperti i parë, por jo nga eksperti i dytë.
- F3: numrin e fjalëve kyçe kandidate të zgjedhura nga eksperti i parë, por jo nga i dyti.
- F4: numri i fjalëve kyçe kandidate që nuk janë zgjedhur nga asnjëri prej ekspertëve.

Fjalë kyçe kandidate, paraqitur në Tabelën 6.3 në kolonën *Fk.Kandidat*, do të quajmë të gjithë termat unikë në dokument që merren pasi papraproçesohet dokumentin. Këto terma janë rreth 60% e termave fillestarë të dokumentit të paproçesuar nga ana linguistike.

Tabela 6.3. *Vlerësimi i koeficientit K për konkordancën midis autorëve*

Klasa	Fk. Kandidat	Eksp.1	Eksp.2	F1	F2	F3	F4	P(a)	P(e)	Kappa
Biologji	1200	200	203	102	98	101	899	0.8342	0.6575	0.5158
Ekonomi	498	198	201	88	110	113	187	0.5522	0.4345	0.2081
Filma	449	200	200	121	79	79	170	0.6481	0.4393	0.3724

Histori	1196	219	211	97	122	114	863	0.8027	0.6363	0.4575
Kimi	1008	208	200	86	122	114	686	0.7659	0.6002	0.4145
Letërsi	1285	205	220	73	132	147	933	0.7829	0.6408	0.3955
Muzikë	496	200	200	81	119	119	177	0.5202	0.4331	0.1536
Sport	581	200	200	65	135	135	246	0.5353	0.4501	0.1548
Totali	**6713**	**10.1875**	**10.2188**	**713**	**917**	**922**	**4161**	**0.7261**	**0.5468**	**0.3956**

Sipas Tabelës 6.3. dhe sipas çështjeve të trajtuara në sesionin 6.2, P(a), P(e) dhe Kappa janë llogaritur si:

$$P(a) = \frac{F1+F4}{Fk.Kandidat} \qquad P(E) = \frac{(F1+F2)(F1+F3)+(F3+F4)(F2+F4)}{total^2} \qquad \kappa = \frac{\Pr(a)-\Pr(e)}{1-\Pr(e)}$$

Sipas koeficientit Kappa, konkordanca midis fjalëve kyçe të dy ekspertëve është rreth 40%.

6.3.2. Testi i Cilësisë së Fjalëve Kyçe Automatike

Precision dhe recall të sistemit automatik të nxjerrjes së fjalëve kyçe llogariten duke përdorur si standard grupin e fjalëve kyçe të caktuara manualisht nga eksperti.

Për çdo dokument d_i, marrim $FkMan = fkm_1 \ldots fkm_n$ si fjalët kyçe të caktuara manualisht nga eksperti. Shënojmë N numrin e të gjitha fjalëve kyçe të caktuara nga eksperti. Për çdo i, j, nëqoftëse $i \neq j$ atëherë $kfm_i \neq fkm_j$, pra fjalët kyçe nuk përsëriten brenda të njëjtit grup. Po shënojmë $FkAut = fka_1 \ldots fka_m$ fjalët kyçe që nxirren automatikisht nga sistemi ynë. Shënojmë M numrin e të gjitha fjalëve kyçe të nxjerra nga sistemi automatik. Edhe në ketë rast kemi për çdo $i \neq j$ atëherë $kfa_i \neq fka_j$.

Precision dhe recall llogariten si në vijim:

Për çdo $0 < i < m$ kontrollojmë nëqoftëse fka_i gjendet në ndonjë $fkm_i \in FkMan$. Në këtë mënyrë mblidhen të gjitha përputhjet e fjalëve kyçe të dokumentave. Këtë njësi po e shënojmë si NrP,(numri i përputhjeve).

$$Recall = \frac{NrP}{N} \quad \text{dhe} \quad Precision = \frac{NrP}{M}$$

Në mënyrë që të balancojmë kontributin e këtyre dy njësive do të marrim:

$$F\text{-}measure = \frac{2*Precision*Recall}{(Precision+Recall)}$$

Rezultatet e përmbledhura për çdo kategori po i paraqesim në Tabelën 6.4.

Tabela 6.4. *Precision dhe recall për sistemin automatik të nxjerrjes së fjalëve kyçe*

Klasa	M	N	NrP	Precision	Recall	F-Measure
Biologji	158	200	88	0.4400	0.5570	0.4916
Ekonomi	123	198	61	0.3081	0.4959	0.3801
Filma	107	200	49	0.2450	0.4579	0.3192
Histori	203	219	72	0.3288	0.3547	0.3412
Kimi	124	208	79	0.3798	0.6371	0.4759
Letersi	181	205	63	0.3073	0.3481	0.3264
Muzike	115	200	57	0.2850	0.4957	0.3619
Sport	124	200	38	0.1900	0.3065	0.2346
Totali	**1135**	**1630**	**507**	**0.3110**	**0.4467**	**0.3667**

Vlerat e precision dhe recall varen nga dy faktorë kryesorë: cilësia e vlerësimit njerëzor nga eksperti dhe cilësia e paraproçesimit linguistik që i është bërë korpusit. Për sa i përket çështjes së parë, ne jemi përpjekur të marrim ekspertë të fushës për të caktuar fjalët kyçe, megjithatë këto ekspertë nuk janë të trajnuar për nxjerrje fjalësh kyçe dhe përveç njohurive të thelluara në funshën përkatëse të dokumentit, nuk kanë njohuri mbi kriteret që duhet të ndjekin për të caktuar fjalët kyçe të duhura. Për sa i përket çështjes së dytë, për gjuhën Shqipe ka akoma punë në përmirësimin e proçesimit automatic linguistik.

6.3.3. Testi i Ndikimit të Madhësisë së Korpusit Trajnues

Klasifikatori Naive Bayes ka një rol të rëndësishëm në përmirësimin e performancës së modelit të thjeshtë statistikor. Naive Bayes dhe në përgjithësi shumë nga modelet e Machine Learning, bazohen në të dhënat trajnuese. Klasifikimi bazohet në një model i cili gjenerohet si rezultat i trajnimit nëpërmjet shembujve, në rastin tonë dokumenta trajnues të ndarë në kategori. Roli i testit të radhës lidhet me vlerësimin e ndikimit të madhësisë së korpusit trajnues në cilësinë e fjalëve kyçe të nxjerra. Për këtë është përsëritur testi i sesionit 6.3.3. me madhësi të korpusit përkatësisht: 80 dokumenta (10 për secilën klasë), 160 dokumenta (20 për secilën klasë), 240 dokumenta (30 për secilën klasë), 320 dokumenta (40 për secilën klasë) dhe 400 dokumenta (50 për secilën klasë). Vlerat e arritura për F-measure paraqiten në tabelën e mëposhtme:

Tabela 6.5. *Ndryshimi i F-measure sipas madhësisë së korpusit*

Madhësia e korpusit	F-measure
0	0
80	0.1825
160	0.2526
240	0.3137

320	0.3587
400	0.3602

Tendenca e kurbës në paraqitjen grafike tregon që rritja e cilësisë së fjalëve kyçe është me e dukshme në madhësi të vogla të korpusit, ndërsa për madhësi korpusi rreth 300 dokumenta rritja e cilësisë është më e qëndrueshme, ajo nuk ndryshon ose ndryshon pak me rritjen e mëtejshme të korpusit.

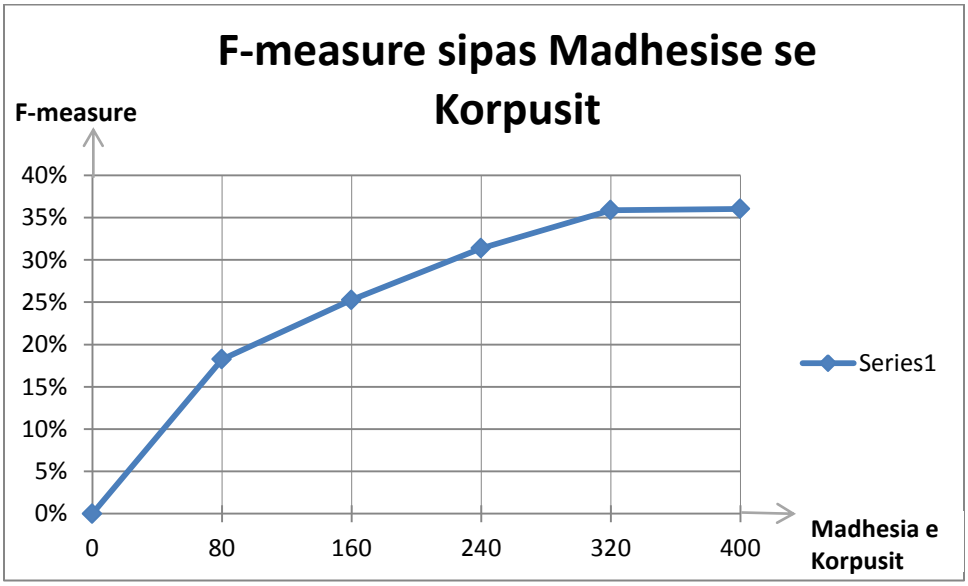

Figura 6.3. Paraqitja grafike e marëdhënies midis madhësisë së korpusit dhe F-measure

6.3.4. Testi i Vlerësimit të Shkallës së Përmirësimit ndaj Modelit Statistikor

Modeli statistikor TFIDF, konsiderohet si model me performancë të mirë dhe është i thjeshtë për t'u implementuar. Shpesh ai përdoret si model reference për modelet e tjera. Meqënëse algoritmi ynë kombinon modelin statistikor, TFIDF, me Machine Learning, eksperimenti në vijim do të aplikohet mbi të njëjtën bashkësi dokumentash testues, por duke përdorur për nxjerrjen e fjalëve kyçe vetëm komponentin statistikor. Meqënëse komponenti statistikor i thjeshtë vetëm i rendit fjalët kyçe kandidate sipas peshës, do të marrim 10 kandidatët me peshë më të lartë statistikore dhe do t'i konsiderojmë si fjalë kyçe të nxjerra nga komponenti statistikor, tfidf. Numri 10 është zgjedhur meqënëse mesatarisht ekspertët kanë përcaktuar 10 fjalë kyçe për çdo dokument. Po llogarisim vlerën e Precision dhe Recall për këtë rast, duke përdorur si sistem referimi të njëjtat fjalë kyçe të ekspertit si një eksperimentin e sesionit 6.3.2. Rezultatet e arritura janë paraqitur në tabelën e mëposhtme.

Tabela 6.6. *Precision dhe recall sipas komponentit statistikor*

Klasa	Nr.Fk.TFIDF	Nr.FK.Eksp	NrP	Precision	Recall	F-Measure
Biologji	200	200	4	0.0200	0.0200	0.0200
Ekonomi	200	198	55	0.2778	0.2750	0.2764
Filma	200	200	32	0.1600	0.1600	0.1600
Histori	200	219	60	0.2740	0.3000	0.2864
Kimi	200	208	61	0.2933	0.3050	0.2990
Letërsi	200	205	54	0.2634	0.2700	0.2667
Muzikë	200	200	33	0.1650	0.1650	0.1650
Sport	200	200	30	0.1500	0.1500	0.1500
Totali	**1600**	**1630**	**329**	**0.2018**	**0.2056**	**0.2037**

Për t'i dhënë më shumë kuptim vlerës së llogaritur për F-measure me anë të komponentit statistikor do ta krahasojmë këtë rezultat me zgjidhje të dhëna për gjuhë të tjera. Në një artikull i cili teston metodat për nxjerrjen e fjalëve kyçe, (Lemnitzer and Monachesi, 2007), po paraqesim pjesën e rezultateve që lidhen me vlerat e precision dhe recall për nxjerrjen e fjalëve kyçe nëpërmjet metodave statistikore, TFIDF, në disa gjuhë të ndryshme. Duket që rezultati më i mirë i arritur është për gjuhën Angleze me 32% përputhshmëri me fjalët kyçe të caktuara manualisht nga eksperti.

Tabela 6.7. *Precision dhe recall i gjetjes së fjalëve kyçe me metoda statistikore për disa gjuhë*

	TFIDF		
Gjuha	**Precision**	**Recall**	**F-Measure**
Anglisht	0.48	0.26	0.32
Bullgarisht	0.39	0.19	0.25
Cekisht	0.23	0.17	0.18
Gjermanisht	0.18	0.15	0.16
Hollandisht	0.36	0.25	0.29
Polonisht	0.42	0.19	0.25
Portugalisht	0.31	0.18	0.22
Rumanisht	0.26	0.11	0.15

6.3.5. Testi i Fjalëve Kyçe me Peshë Statistikore të Ulët

Qëllimi i eksperimentit të radhës është të vlerësojë përputhshmërinë midis fjalëve kyçe të gjeneruara automatikisht nga algoritmi ynë dhe fjalëve kyçe të gjeneruara nga komponenti linguistik. Duke iu referuar sesionit 5.4, është evidente për nga mënyra e projektimit, që për çdo dokument të testuar, do të kemi të paktën 5 fjalë kyçe të përbashkëta midis dy grupeve me fjalë kyçe. Duke ditur që algoritmi ynë gjeneron mesatarisht 7 fjalë kyçe për dokument, mbetet të vlerësojmë pjesën e mbetur nëqoftëse bën pjesë apo jo në grupin e termave kandidatë me TFIDF

të lartë. Na intereson të dimë sa përqind e fjalëve kyçe të nxjerra nga algoritmi ynë janë me TFIDF të lartë. Po përfshijmë në grupin e termave kandidatë me TFIDF të lartë, 10 termat e parë të renditur në rend zbritës sipas peshës. Në këtë rast do të përdorim të njëjtin testim që u realizua në sesionin 6.3.4, për TFIDF, vetëm se tani do ta krahasojmë për përputhshmëri me testet e realizuara në sesionin 6.3.2. për fjalët kyçe të nxjerra nga algoritmi ynë automatik. Për këto dy grupe me fjalë kyçe do të llogarisim Precision dhe Recall për të përputhshmërinë midis tyre.

Tabela 6.8. *Përputhshmëria midis Algoritmit të kombinuar me Machine Learning dhe atij statistikor*

Klasa	FK.Alg.Aut	FK.TFIDF	NrP	Precision	Recall	F-Measure
Biologji	158	200	144	0.7200	0.9114	0.8045
Ekonomi	123	200	101	0.5050	0.8211	0.6254
Filma	107	200	102	0.5100	0.9533	0.6645
Histori	203	200	161	0.8050	0.7931	0.7990
Kimi	124	200	112	0.5600	0.9032	0.6914
Letërsi	181	200	127	0.6350	0.7017	0.6667
Muzikë	115	200	106	0.5300	0.9217	0.6730
Sport	124	200	120	0.6000	0.9677	0.7407
Totali	**1135**	**1600**	**973**	**0.6081**	**0.8573**	**0.7115**

Mund të themi që 71% e fjalëve kyçe të nxjerra nga algoritmi ynë janë të njëjtat që nxjerr edhe algoritmi statistikor. Përmirësimi në përfomancë i algoritmit tonë vjen si rezultat i 29% të fjalë kyçe që i përzgjedh në bazë të veçorive të klasifikatorit Naive Bayes.

6.3.6. Testi i Vlerësimit Manual të Fjalëve Kyçe

Me këtë test synojmë të kategorizojmë në dy grupe fjalët kyçe të gjeneruara nga sistemi automatik. Kemi përdorur një ekspert human i cili vlerëson fjalët kyçe të nxjerra nga algoritmi automatik. Ai duhet të vendosë në kategorinë NEGATIVE grupin e fjalëve kyçe të gjeneruara nga algoritmi automatik të cilat janë të papërshtatshme për atë dokument, dhe në kategorinë POZITIVE të gjitha fjalët kyçe që janë nxjerrë automatikisht dhe që konsiderohen si të përshtatshme për atë dokument. Në Tabelën 6.9 po japim rezultatet e këtij testi.

Tabela 6.9. *Rezultatet e vlerësimit manual të fjalëve kyçe nga një ekspert*

Klasa	FK.automatike/Klase	Pozitive	Negative
Biologji	158	146	12
Ekonomi	123	118	5
Filma	107	98	9
Histori	203	186	17
Kimi	124	118	6
Letersi	181	171	10

Muzike	115	104	11
Sport	124	102	22
Totali	**1135**	**1043**	**92**
Perqindje		**91.89%**	**8.11%**

Vihet re një rezultat shumë i mirë. Ky përfundim mund të interpretohet duke patur parasysh që edhe midis ekspertëve ka mospërputhje në vendosjen e fjalëve kyçe, megjithatë ata nuk i kundërshtojnë fjalët kyçe që vendos në mënyrë të pavarur eksperti tjetër.

E njejta situatë vihet re edhe me fjalët kyçe të caktuara automatikisht. Si përfundim mund të themi që nxjerrja e automatizuar e fjalëve kyçe ka performancë të krahasueshme me atë njerëzore.

7. Konkluzionet

Në këtë dizertacion trajtohet problemi i nxjerrjes së fjalëve kyçe nga tekste në gjuhën Shqipe. Algoritmi është një propozim i ri në literaturë i cili është projektuar duke respektuar rregullat strukturore të gjuhës Shqipe, por që mund të përshtatet edhe për gjuhë të tjera. Ky algoritëm mbështetet në krijimin e një modeli nëpërmjet klasifikatorit Naïve Bayes i cili përdoret për të rritur efiçensën e modelit statistikor të nxjerrjes së fjalëve kyçe. Klasifikatori jep kontribut në identifikimin e termave kyçe duke vlerësuar nëqoftëse një term i zgjedhur nga modeli statistikor është ose jo determinues për klasifikatorin. Një fjalë kyçe e përshtatshme për një dokument duhet të jetë njëkohësisht edhe identifikuese e klasës së cilës i përket dokumenti. Algoritmi aplikohet mbi dokumenta tekst të shkruara në gjuhën Shqipe dhe të paraproçesuara nga ana linguistike. Zgjidhja është specifike për gjuhën Shqipe dhe është implementuar në mënyrë të tillë që t'i përshtatet rregullave strukturore të kësaj gjuhe.

Si input është përdorur një bashkësi me lidhëzat e gjuhës Shqipe dhe një bashkësi dokumentash tekst në gjuhën Shqipe të klasifikuara sipas disa klasave të zgjedhura. Çdo dokument është i shoqëruar me: dy grupe me fjalë kyçe të zgjedhura manualisht nga dy ekspertë të ndryshëm; një bashkësi dokumentash trajnues, edhe këto të shkruar në gjuhën Shqipe dhe të klasifikuara paraprakisht. Këto dokumenta përdoren për të gjeneruar modelin.

Algoritmi përbëhet nga disa module të implementura në Java. Moduli i paraproçesimit linguistik është i bazuar në rregullat e gramatikës së gjuhës Shqipe dhe ka për qëllim t'i kthejë fjalët në formën e tyre më të shkrutër të pandryshueshme (rrënjët e fjalëve). Moduli statistikor llogarit peshat e termave të dokumentit dhe moduli i Machine Learning krijon modelin e klasifikatorit. Secili prej këtyre moduleve merr pjesë në tablonë e përgjithshme të nxjerrjes së fjalëve kyçe, siç ilustrohet në sesionin 5.4.

Vlerësimi i performancës bazohet në një faktor F-measure i cili përdor Precision dhe Recall. Kjo është një vlerë e matur në përqindje e cila tregon përputhmërinë e fjalëve kyçe të nxjerra automatikisht nga algoritmi ynë me fjalët kyçe të caktura manualisht nga një ekspert i jashtëm. Vlera e F-measure për algoritmin tonë rezultoi 36%, rezultat ky që konsiderohet relativisht i mirë duke patur parasysh që edhe midis dy ekspertëve të ndryshëm që zgjedhin manualisht fjalë kyçe nga i njëjti dokument, përputhshmëria është rreth 40%.

I krahasuar me performancën e sistemeve të ngjashme të nxjerrjes së fjalëve kyçe për gjuhën Angleze, algoritmi ynë ka efikasitet të krahasueshëm duke marrë në konsideratë që moduli ynë linguistik për Shqipen ka vend për përmirësime të mëtejshme për shkak se fusha e Linguistikës Komputacionale për gjuhën Shqipe është pak e studiuar.

Në aspektin e performancës i kemi kushtuar vëmendje disa përmirësimeve të mundshme nga ana llogaritëse:

1. Ripërdorim modelin që gjenerohet nga klasifikatori Machine Learning sepse proçedura e nxjerrjes së modelit ka numër të lartë veprimesh dhe ndikon ndjeshëm në përformancë. Modeli ruhet ne një dokument tekst dhe ripërdoret sa herë është e nevojshme të aplikohet algoritmi për nxjerrjen e fjalëve kyçe në një dokumenti të ri.

2. Të dhënat statistikore të dokumentave trajnues llogariten vetëm një herë dhe ruhen në një bazë të dhënash për t'u ripërdorur kur llogariten peshat statistikore të fjalëve kyçe kandidate të një dokumenti të ri për nxjerrjen e fjalëve kyçe.

Pavarësisht përpjekjeve tona, në rastin kur madhësia e dokumentit test arrin në disa dhjetra faqe të cilat duhet të paraproçesohen nga ana linguistike dhe statistikore dhe me pas t'i nxirren fjalët kyçe, koha e proçesimit bëhet e gjatë. Ky mund të jetë një problem që mund të anashkalohet nëqoftëse e mendojmë algoritmin automatik të nxjerrjes së fjalëve kyçe si një proçes i cili kryhet në background në një motorr kërkimi.

7.1. Kontributi i Këtij Disertacioni

Në këtë dizertacion jepen zgjidhje për disa probleme:

1. *Paraproçesimi linguistik i teksteve në gjuhën Shqipe:* Për këtë është ndërtuar një modul i i bazuar në rregullat strukturore të gjuhës Shqipe. Ai është komponenti bazë në zgjidhjen e shumë problemeve të Proçesimit të Gjuhëve Natyrale, NLP, si kërkimi nëpërmjet motorrëve të kërkimit, përgjigjia e pyetjeve, përmbledhja e tekstit, përkthimi i tekstit nga një gjuhë tjetër në Shqip ose anasjelltas, etj. Ky komponent mund t'u vijë në ndihmë të gjithë atyre që studiojnë NLP për gjuhën Shqipe.

2. *Nxjerrjen e fjalëve kyçe nëpërmjet metodave statistikore për herë të parë për tekste në gjuhën Shqipe:* pavarësisht se kjo nuk përbën një risi në kontekstin e modelit, mund të themi që rezultatet e performancës së modelit statistikor në gjuhën Shqipe tashmë mund të shfaqen përkrah rezultateve qe jep ky model në gjuhë të tjera.

3. *Implementimi i një modeli të ri, i cili shfrytëzon kapacitetin e modeleve Machine Learning për të mësuar nga shembujt, për të rritur efiçensën e modelit statistikor.*

4. *Implementimi i një metode që kombinon në një workflow potencialin përfaqësues (angl. representative) të fjalëve kyçe sipas TFIDF e bazuar në llogaritje statistikore dhe potencialin diskriminues/klasifikues të fjalëve kyçe sipas metodave të Machine Learning.*

7.2. Puna në të Ardhmen

Algoritmi ynë mund të përmirësohet me tej në aspekte të efiçensës duke u fokusuar në disa drejtime:

1. Përmirësimi i komponentit linguistik duke e pasuruar atë edhe me etiketat e pjesëve të ligjëratës. Kjo do të kërkojë një punë më të thelluar në aspektin gjuhësor por do të kontribuojë në mundësinë për t'i përzgjedhur fjalët kyçe edhe në bazë të funksionit të tyre në ligjëratë.

2. Proçesi i gjetjes së rrënjëve të fjalëve mund të zgjerohet duke shtuar edhe elementë të Machine Learning përvec elementëve të bazuara në rregulla.

3. Përmirësimi i komponentit linguistik me gjetjen e entiteve të emëruara (Named Entity Recognition). Kjo mund të rrisë mundësinë që fjalë me frekuencë të ulët të cilat potencialisht do përjashtoheshin nga lista e fjalëve kyçe kandidate sipas TFIDF, të mund të përdoren si fjalë kyçe me karakter diskriminues.

4. Përmirësimi i komponentit statistikor mund të realizohet duke shtuar edhe elementë të tjerë, përveç TFIDF-së, të cilët të kontribuojnë në peshën e fjalëve. Mund të përmendim këtu pozicionet e ndeshjes së fjalëve në një dokument. Për shembull, ka më shumë probabilitet që një term të jetë fjalë kyçe, nëqoftëse ai ndeshet në fillim të paragrafit ose në titujt e sesioneve.

5. Zgjerimi i fjalëve kyçe në shprehje kyçe duke përdorur n-gram. Ky zgjerim ka nevojë për një paraproçesim linguistik të etiketimit të pjesëve të ligjëratës, siç sugjerohet në pikën 1.

6. Të merren korpuse më të mëdha për trajnim dhe të testohet gjetja e fjalëve kyçe në tekste të plota apo libra në format elektronik.

7. Të investigohet një klasifikator tjetër i cili potencialisht mund të performojë më mirë në disa fusha sesa Naïve Bayes.

8. Algoritmi mund të paralelizohet nëqoftëse kërkohet të arrihet përformancë e tipit real time.

Aneks A

RREGULLAT E NXJERRJES SË RRËNJËJVE TË FJALËVE (Agalliu et al., 2002).

***Tabela A.1**. Tebala e eleminimit të zanoreve*

Rregullat e eleminimit te Zanoreve	
Zanoret per eleminim	**Shembuj**
-a	makina → makin
-e	punëtore → punëtor
-ë	makinë → makin
-i,	djali → djal
-o	djalo → djal
-u	nisu → nis

***Tabela A.2**. Tebala e lidhësave të gjuhës Shqipe*

Lista e Lidhesave
a, e, i, o, u, dhe, sepse, kur, edhe, në, qoftë, s, se, as, po, nga, të, me, sesa, prej, apo, ose, ku, sa, për, si, tek, veç, por, daç, pa, jo, nuk, mbasi, krye, de, seç, ndoshta, bri, ngaqë, nja, bile, porse, nëse, sikurse, jose, oburrani, që, porsi, meqë, kurse, ua, bobo, shët, oburra, teksa, dot, vallë, sidomos, le, mos, pale, posi, thuajse, pothuajse, siç, kushedi, ngase, ndërsa, prveçse, veçse, derisa, megjithatë, deri, qysh, atë, qyqja, këtë, of, rreth, ja, uf, bubu, më, te, pthu, sikur, ashtu, ndonëse, ptu, oh, bërr, mbase, çdo, deri, mu, ndaj, prandaj, pra, megjithëse, vuu, kuku, bam, madje, mosni, ëhë, bah, aha, oho, uhu, eu, ah, oj, hopa, na, moj, sus, hë, hop, iu, ia, sipas, nën, një, pas, etj, ende, ana, ta, herë, here, ti, to, qe, do, dy, duke, la, mbi, ii, shek, m, tu, tre, tri, cili, t'ia, tja, psh, pse

***Tabela A.3**. Tebala e rregullave për fshirjen e parashtesave*

Rregullat e fshirjes se parashtesave			
Parashtesa	**Shembuj**	**Parashtesa**	**Shembuj**
mos-	mosmarveshje → marveshje	poshtë-	i poshtëshënuar → shënuar
para-	parandjenjë → ndjenjë	drejt-	drejtshkrim → shkrim
kundër-	kundërmasë → masë	gjysmë-	gjysmëvjeçar → vjeçar
sipër-	sipërfaqe → faqe	gjithë-	gjithëfuqishëm → fuqishëm
mbi-	mbishkrim → shkrim	jashtë-	jashtëfisnor → fisnor
ndaj-	ndajshtes →shtes	plot-	i plotfuqishëm → fuqishëm
për-	prmas →mas	shumë-	shumëshekullor →shekullor
prej-	prejardhje → ardhje	vetë-	vetëkritik → kritik
ultra-	ultratingull → tingull	hidro-	hidroenergjitik →energjitik
anti-	antifetar → fetar	pseudo-	pseudokulturë → kulturë
super-	superfuqi → fuqi	ndër-	ndërlidh → lidh
poli-	polifoni → foni	de-	demaskoj →maskoj
lart-	i lartshënuar → shënuar	pa-	i paafrueshëm → afrueshëm
jo-	jonormal → normal	gjeo-	gjeomagnetik → magnetik
para-	parashkollor → shkollor	ri-	rizgjedh →gjedh
pro-	proamerikan →amerikan	neo-	neolatin →latin
pas-	pasardhës → ardhës	zh-	zhvlerësoj → vlerësoj
prapa-	prapavendosur →vendosur	ndër-	ndërsektorial → sektorial
tej-	i tejdukshëm → dukshëm	stër-	i stërmadh →madh
sapo-	i sapolindur → lindur	bashkë-	bashkëkohor → kohor
ç-	çnjerëor → njerëor	nën-	nënligjor → ligjor

***Tabela A.4**. Tebala e rregullave për fshirjen e prapashtesave*

Rregullat e fshirjes se prapashtesave			
Parashtesa	**Shembuj**	**Parashtesa**	**Shembuj**
-ës	mbledhës → mbledh	-sirë	ëmbëlsirë → ëmbël
-ar	argjendar → argjend	-ore	mëngjesore → mëngjes
-tar	gjahtar → gjah	-tore	ëmbëltore → ëmbël
-or	këmbësor → këmbës	-ëtirë	shkretëtirë → shkret
-tor	fajtor → faj	-inë	cmendinë → cmend
-ak	fierak → fier	-ësi	gjatësi → gjat
-an	dibran → dibr	-ëri	djalëri → djal
-jan	matjan → mat	-shëm	gatshëm → gat
-as	tiranas → tiran	-shmëri	gatishmëri → gati
-it	gjirokastrit → gjirokastr	-im	besim → bes
-iot	mallakastriot → mallakastr	-esë	jetesë → jet
-ist	kitarist → kitar	-atë	ligjëratë → ligjër
-ant	kursant → kurs	-imë	i lartshënuar → shënuar
-më	blegërimë → blegëri	-azi	djathtazi → djatht
-ent	asistent → asist	-llk	avokatllk → avokat
-ier	bankier → bank	-izëm	kapitalizëm → kapital
-xhi	qiraxhi → qira	-azh	spiunazh → spiun
-xheshë	qiraxheshë → qira	-urë	agjenturë → agjent
-je	mbathje → mbath	-thi	kalimthi → kalim
-isht	barisht → bar	-ac	qullac → qull
-urin	qelqurin → qelq	-acak	morracak → morr
-ësirë	hapësirë → hap	-acuk	rrjepacuk → rrjep
-alaq	ngordhalaq → ngordh	-iv	objekt-iv → objekt
-aluq	trashaluq → trash	-ueshëm	i trishtueshëm → trisht
-alec	barkalec → bark	-të	rëndomtë → rëndom
-avec	grindavec → grind	-ur	hapur → hap
-atar	gjakatar → gjak	-imisht	detyrimisht → detyr
-ues	pastrues → pastr	-toj	ledhatoj → ledha
-ësh	rrugësh → rrug	-ullon	qarkullon → qark
-anik	qelbanik → qelb	-ëroj	lajmëroj → lajm
-ik	atomik → atom	-ësoj	lartësoj → lart
-al	embrional → embrion	on-	kullon → kull
-ual	gradual → grad	-ëzoj	pyllëzoj → pyll
-os	burgos → burg	-oj	punoj → pun
-is	arratis → arrati	-uar	punuar → pun
-ova	punova → pun	-akisht	ushtarakisht → ushtar
-as	barkas → bark		

Referenca

Abney, S. P. (1990). Rapid icremental parsing with repair. Proceedings of the 6th New OED Conference: Electronic Text Research 1-9, University of Waterloo, Waterloo, Ontario.

Agalliu, F., Angoni, E., Demiraj, Sh., Dhrimo, A., Hysa, E., Lafe, E., and Likaj, E. (2002). Gramatika e Gjuhes Shqipe. Akademia e Shkencave.

Aggarwal, Ch. C. and Zhai, Ch. (2012). Mining Text Data. Springer.

Ananiadou, S. (1994). A methodology for automatic term recognition, COLING '94 Proceedings of the 15th conference on Computational linguistics - Volume 2 Pages 1034-1038.

Bach, N. and Badaskar, S. (2007). A Review of Relation Extraction. Canergie Mellon University Pittsburgh: Language Technologies Institute School of Computer Science, 2-11.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. New York: Addison-Wesley-Longman.

Ball, C.N. (1996). ”Tutorial: concordances and corpora”. Text in electronic format in http://www.georgetown.edu/cball/corpora/tutorial.html; (Last Retrieved: 15 February 2014).

Banko, M. and Etzioni, O. (2008). The tradeoffs between open and traditional relation extraction. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics.

Belkin and Croft (1992). Information filtering and information retrieval: Two sides of the same coin? Communications of the ACM, v. 35, no. 12: 29-38.

Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V. (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. In Proc. Int. Conf. Research and Development in Information Retrieval, 192-199.

Berry, M.W. and Kogan, J. (2010). Text Mining: Applications and Theory, John Wiley & Sons, Ltd.

Bourigault, D. (1996). LEXTER, a Natural Language tool for terminology extraction”, Proceedings, 7th EURALEX International Congress, pages 771-779, Goteborg.

Cabré Castellví, M.T., Bagot, R.E. and Palatresi, J.V. (2001). Automatic Term Detection: A Review of Current Systems, Recent Advances in Computational Terminology, chapter 3, John Benjamins Publishing Company, Philadelphia, PA, USA.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), pp. 249–254.

Chambers, N. and Jurafsky, D. (2011). Template-based information extraction without the templates. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 976–986.

Daille, B., (1996). Study and implementation of combined techniques for automatic extraction of terminology. In J. L. Klavans & P. Resnik (eds.), The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pp. 49–66. Cambridge, MA: MIT Press.

David, S. and Plante, P. (1991). Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes. Proceedings of the Montreal Colloquium Les industries de la langue: perspectives des années 1990, 1: 71-88.

Enguehard, C. and Pantera, L., (1995). Automatic Natural acquisition of a terminology, Journal of Quantitative Linguistics, 2(1):27-32.

Feldman, R. and Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press.

Finkel, J., Dingare, S., Nguyen, H., Nissim, M., Manning, C. and Sinclair G. (2004). Exploiting context for biomedical entity recognition: From syntax to the web. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pages 88–91.

Frakes, W. B. And Baeza-Yates, R. (1992). Information Retrieval: Data Structures and Algorithms, Prentice Hall, 1 edition.

Frank, E., Paynter, G. W., Witten, I.H., Gurwin, C., Nevill-Manning, C. C., (1999). Domain Specific Keyphrase Extraction, Proceeding IJCAI'99 Proceedings of the 16th international joint Conference on Artificial Intelligence - Volume 2, Pages 668-673.

Frank, E., Paynter, G., Witten, I.H., Gutwin, C. and Nevill-Manning, C. (1999). Domain-specific keyphrase extraction. Proc Int Joint Conf on Artificial Intelligence IJCAI-99. Stockholm, Sweden, pp. 668-673.

Frantzi, K., Ananiadou, S. and Tsujii, J. (1998). The C-value/NC-valueMethod of Automatic Recognition for Multi-Word Terms. Lecture Notes in Computer Science. Springer-VerlagBerlin Heidelberg. pp 585-604.

Gjati, E. (2013). Handling Composite Words in Stemming Albanian in a Text Classification Approach. Master Thesis, University of Greenwich.

Hippisley, A., Cheng, D. and Ahmad, K. (2005). The head-modifier principle and multilingual term extraction. Natural Language Engineering 11(2):129–157.

Hulth, A. (2004), Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. Doctoral Dissertation, Department of Computer and Systems Sciences, Stockholm University, April 2004.

Jacquemin, C. and Borigault, D. (2003). Term extraction and automatic indexing. In R. Mitkov (ed.), The Oxford Handbook of Computational Linguistics, pp. 599–615. Oxford University Press.

Jin Huang; Lu, J.; Ling, C.X., (2003). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy, Data Mining, 2003. ICDM 2003. Third IEEE International Conference on , vol., no., pp.553,556, 19-22 Nov.

Jolliffe, I. T. (2002). Principal Component Analysis, 2nd ed. Springer, New York.

Jones, S. and Paynter, G.W. (2000). Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. Jam Soc Inf Sci Technol;53(8):653brar

Justeson, J. S. and Katz, S. M., (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering 1(1):9–27.

Kadiu, S. (2012). Tipologjia e fjalëformimit të emrit në gjuhën turke dhe në gjuhën shqipe. Punim doktorature, Universiteti i Tiranës, Fakulteti i Gjuhëve të Huaja, Tiranë 2012.

Krovetz, R. (1993). Viewing morphology as an inference process. Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'93, 191-203.

Kuropka, D (2004). Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und -Retrieval mit relationalen Datenbanken. Advances in Information Systems and Management Science, bd.10, 264 pages, ISBN 3-8325-0514-8.

Lauriston, A. (1994). Automatic recognition of complex terms: Problems and the TERMINO solution, Terminology, 1(1):147-170.

Lemnitzer, L. and Monachesi, P. (2007). Keyword extraction for metadata annotation of Learning Objects, RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments. Borovets, 26. September 2007.

Lewis, D.D. (1995). Evaluating and optimizing autonomous text classifcation systems. In SIGIR.

Manning, C.D. and Raghavan, P. and Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge University Press. Cambridge, England.

Marx, Z., Dagan, I. and Shamir, E. (2002). Cross-component clustering for template learning. In Proceedings of the 2002 ICML Workshop on Text Learning.

Mihalcea, R., and Tarau, P. (2004), Textrank: Bringing order into texts, In Lin, D., & Wu, D. (Eds.), Proceedings of EMNLP 2004, pp. 404{411 Barcelona, Spain. Association for Computational Linguistics.

Moens, M.F. (2006), Information Extraction: Algorithms and Prospects in a Retrieval Context Katholieke Universiteit Leuven, Belgium, Published by Springer.

Nenkova, A. and McKeown, K. (2012). A Survey of Text Summarization Techniques. In C. Aggarwal & C. Zhai (ed.) Mining Text Data (Book chapter) pp. 43-76. ISBN: 978-1-4614-3222-7.

O'Conor, B.C. (1996). Explorations in Indexing and Abstracting: Pointing, Virtue, and Power. Library and Information Science Text. Libraries Unlimited, Inc.

Ohsawa, Y. Benson, N. E. and Yachida, M., (1998). KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor. In Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries (ADLlume 2:27-32.1–18, April 1998).

Park, Y., Byrd, R.Y. and Boguraev, B. (2002). Automatic Glossary Extraction: Beyond Terminology Identification, Proceedings of the 19th international conference on Computational linguistics; 01/2002.

Rosenfeld, B. and Feldman, R. (2007). Clustering for unsupervised relation identification. In Proceedings of the 16th ACM conference on Information and Knowledge Management, pages 411.

Sadiku, J. (2011). A Novel Stemming Algorithm for Albanian in a Data Mining Approach for Document Classification. Master Thesis, University of Greenwich.

Salton, G. and Buckley, C. (1988). Term-weighing approache s in automatic text retrieval. In Information Processing & Management, 24(5): 513-523.

Smadja, F. (1993). Retrieving collocations from text: XTract, Journal of Computacional Linguistics - Special issue on using large corpora. 19(1):143-177.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Book: Documentat retrieval system, pp 132-142.

Shinyama, Y. and Sekine, S. (2006). Preemptive information extraction using nrestricted relation discovery. In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 304-311.

Turney, P.D., (1999a). Learning to extract keyphrases from text, Technical report, National Research Council, Institute for Information Technology.

Turney, P.D., (1999b), Learning algorithms for keyphrase extraction, Information Retrieval—INRT 34-99.

Turtle, H.R. and Croft, W.B. (1992). A comparison of text retrieval models. The Computer Journal, 35, 3, 279-290.

Van Rijsbergen, C. J. (1979). Information Retrieval 2nd edition, London: Butterworths, 1979.

Wan, X. and Xiao, J. (2008), CollabRank: Towards a collaborative approach to single-document keyphrase extraction. In Proceedings of 22nd International Conference on Computational Linguistics COLINGgital Libraries (ADLlume 2:27-3–976).

Witten, I. H., (2005). Text mining. In Practical handbook of internet computing, edited by M.P. Singh, pp. 14-1 -14-22. Chapman & Hall/CRC Press, Boca Raton, Florida.

Witten, I.H., Paynter, G.W., Frank,E., Gutwin,C,. and Nevill-Manning, C., (1999). KEA: practical automatic keyphrase extraction, Proceedings of the fourth ACM conference on Digital libraries, Pages 254-255.

Witten, J. and Frank, E. (2012). Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Series in Data Management Systems.

Wong, W. et al. (2007a). Determining Termhood for Learning Domain Ontologies in a Probabilistic Framework. In Proceedings of the 6th Aistralasian Conference on Data Mining (AusDM), pp. 55–63.

Wong, W. et al. (2007b). Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency. In Proceedings of the 6th Aistralasian Conference on Data Mining (AusDM), pp. 47–54.

Zhang, A. (2004). The Optimality of Naive Bayes, Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference FLAIRS 2004, AAAI Press.

Zhang, C. (2008). Automatic Keyword Extraction From Documents Using Conditional Random Fields. Journal of Computational and Information Systems.