N-gram analiza tekstualnih dokumenata na srpskom jeziku

Ulfeta Marovac, Aldina Pljasković, Adela Crnišanin, Ejub Kajan

***Sadržaj* — Savremeni način života, elektronsko poslovanje, velika količina podataka dostupna u elektronskoj formi, nametnuli su potrebu za analizom tekstualnih dokumenata pisanih na različitim prirodnim jezicima. Svaki prirodni jezik sadrži mnogo pravila i odstupanja, što otežava analizu dokumenta. N-gram analizom dokumenata može se doći do rezultata bez korišćenja velikih leksičkih resursa. U ovom radu prikazana je n-gram analiza tekstualnih dokumenata pisanih na srpskom jeziku i algoritam za izdvajanje ključnih reči (n-grama) iz dokumenta.**

***Ključne reči* — ključne reči, n-grami, normalizacija dokumenta.**

I. Uvod

DA bi se ubrzao proces traženja dokumenta sa odgovarajućim sadržajem u velikom skupu podataka, dokumenti moraju biti grupisani na osnovu sadržaja(Slika 1.). Složena gramatika sprskog jezika i upotreba dva zvanična pisma čine pretragu dokumenata na sprskom jeziku pravim izazovom. Višeznačnost reči i rečeničnih konstrukcija dodatno otežavaju problem analize sadržaja dokumenata. Za analizu su potrebni različiti leksički resursi: korpusi srpskog jezika, morfološki rečnik, stop-reči, rečnik skraćenica, rečnik vlastitih imena itd. Poteškoće oko pronalaženja odgovarajućih resursa su velike i zavise od vrste dokumenta koja se ispituje. Nije isto značenje reči u dokumentima koji predstavljaju pravne akte i u dokumentima koji opisuju prirodna bogastva jedne zemlje. Dodatni problemi koji prate analizu tekstualnih dokumenata na srpskom jeziku jesu i nepostojanje odrovarajućih resursa, stalne promene u prirodnim jezicima (npr. uvođenje novih termina) i veliko vreme obrade teksta kada se koriste veliki resursi. U ovom radu je prikazan način da se analiza dokumenata olakaša i omogući sa manjim brojem leksičkih resursa. Korišćenje morfološkog rečnika zamenjeno je n-gram analizom reči, i izložen je algoritam pretrage ključnih reči na jednom dokumentu.

Ovaj rad je delimično finansiralo Ministarstvo prosvete i nauke Republike Srbije po projektu III-44007

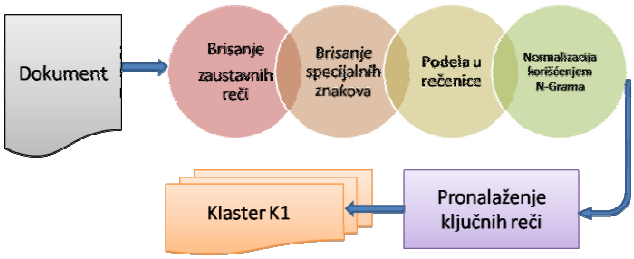
Ulfeta Marovac, *Državni univerzitet u Novom Pazaru*, Vuk Karadžića bb, 36300 Novi Pazar, Srbija (e-mail: umarovac@np.ac.rs)

Aldina Pljasković, *Državni univerzitet u Novom Pazaru*, Vuk Karadžića bb, 36300 Novi Pazar, Srbija (e-mail: apljaskovic@np.ac.rs)

Adela Crnišanin, *Državni univerzitet u Novom Pazaru*, Vuk Karadžića bb, 36300 Novi Pazar, Srbija (e-mail: acrnisanin@np.ac.rs)

Ejub Kajan, *Državni univerzitet u Novom Pazaru*, Vuk Karadžića bb, 36300 Novi Pazar, Srbija (e-mail: ekajan@ieee.org)

U ovom radu akcenat je stavljen na alternativne načine normalizacije dokumenata, posebno na n-gram analizu. Dati rezultati pokazuju kako odabrana normalizacija utiče na izdvajanje ključnih reči dokumenta. Ostatak rada je organizovan na sledeći način. U drugom poglavlju su navedeni potrebni leksički resursi (stop reči, skraćenice, simboli neformalne komunikacije...) i način pripreme dokumenta za analizu. Treće poglavlje sadrži klasifikaciju normalizacija dokumenata, sa posebnim naglaskom na n-gram analizu reči. Četvrto poglavlje demonstira proces izdvajanja ključnih reči. Na kraju, opisane su korišćene tehnologije, objašnjen je značaj dobijenih rezultata i prikazani pravci u daljem istraživanju.



Sl. 1. Proces obrade dokumenata za grupisanje na osnovu sadržaja

II. Priprema dokumenta za normalizaciju

Dokumenti koji se obrađuju često sadrže suvišne znake, nalaze se u ćirilićnom pismu ili sadrže neformalne znake. Dokument se pre analize mora pripremiti, kako bi bio spreman za normalizaciju.

Dokumenti koji se pretražuju su smešteni u bazi podataka i mogu biti u TXT, PDF ili HTML formatu. Zavisno od toga kojeg je formata dokument, primenjuju se odovarajuće funkcije za čišćenje dokumenta od suvišnih znakova. Bez obzira kojeg je formata, dokument koji se normalizuje mora biti u latiničnom pismu, bez suvišnih razmaka između reči. Rečenica se završava tačkom, znakom pitanja ili znakom uzvika. Nova rečenica počinje razmakom i velikim slovom. Format dokumenta treba da bude UTF8. Priprema dokumenta se odvija u sledeća tri koraka:

1. *Obrada dokumenta ćiriličnog i latiničnog pisma*
2. *Obrada dokumenta iz HTML formata i uklanjanje suvišnih znakova*
3. *Prebacivanje teksta iz ASCII u UTF8 format*

Ovakav dokument je spreman za normalizaciju reči. (Slika 2.)

Resursi koji su na raspolaganju su sledeći:

- Stop reči (prilozi, predlozi, veznici, uzvici, rečce, zamenice

- Nastavci (gramatički nastavci za oblik kod deklinacije imenica i prideva i konjugacije glagola[1]
- Izrazi neformalne komunikacije (emotikoni, sleng).

```
procedure PREPAREDOCUMENT(DocumentId)
ConectToDatabase(Project)
Document = GetDocumentFromDatabase(DocumentId)
StopWords = GetStopWordsFromDatabase()
CyrillicToAscii(Document)
LatinToAscii(Document)
HtmlToText(Document)
DeleteStopWords(Document)
DeleteIrregularSigns(Document)
AsciiToLatin(Document)
Sentences = SplitDocumentOnSentences(Document)
```

Sl. 2. Pseudokod algoritma pripreme dokumenta

III. NORMALIZACIJA REČI

U uvodnom delu rada naglašeno je da se dobijaju efikasniji rezultati pretraživanja ukoliko su dokumenti grupisani po nekom kriterijumu. Dokumenta je potrebno grupisati prema glavnim nosiocima njihovog značenja, tj. ključnim rečima. Kako je izdvajanje ključnih reči zahtevan posao, svaka prethodna obrada koja uklanja nepotrebne reči ili delove reči, čini ovaj posao bržim i efikasnijim. Kada je dokument pripremljen za analizu, potrebno je obezbediti da se u ključnim rečima ne pojave padežni/glagolski oblici jedne iste reči, već da se ona tretira kao jedna reč. Sličnost dveju reči se određuje na osnovu njihovog semantičkog značenja. Koren reči je nosilac tog značenja reči, a složena derivacija reči u srpskom jeziku (postojanje prefiksa, infiksa, sufiksa) otežava identifikaciju reči koje imaju zajednički koren. Stoga, pre izdvajanja ključnih reči, dokument prolazi kroz još jedan korak, koji se naziva *normalizacija reči.*

U širem smislu, pod normalizacijom teksta podrazumeva se transformacija teksta u neki drugi oblik koji je pogodniji za neku vrstu kompjuterske obrade. U našem slučaju to je pretraživanje. Svrha normalizacije reči je oslobađanje od suvišnih modifikacija reči koji ne unose izmene u značenju reči, odnosno svođenje tih modifikacija na zajednički, osnovni oblik.

Postoji nekoliko načina da se normalizacija reči sprovede u delo, i oni su nadalje opisani.

A. *Normalizacija 1*

Jedino kod primene ovog tipa normalizacije, neophodno je posedovanje morfološkog rečnika. Korišćenjem ovog resursa, vrši se lematizacija reči – odklanjanje nastavaka za oblik i tvorbenih nastavaka. Prednost ove normalizacije je preciznost rezultata, ali se ona postiže uz ogromne napore, imajući u vidu glasovne promene i obim resursa koji se koriste, što rezultira sporijem dobijanju rezulata. Druga negativna osobina ove normalizacije je što je ona specifična samo za jedan jezik, kao i to što je teško dospeti do ovakvog resursa.

B. *Normalizacija 2*

Nastavci za oblik reči su neophodni resurs za ovaj tip normalizacije. Ovaj tip normalizacije detaljnije je opisan u radu[3]. Njeni nedostaci se sreću kod složene derivacije reči. Budući da je broj nastavaka za oblik neuporedivo manji od broja reči u morfološkom rečniku, složenost ovog algoritma je znatno manja, što čini ovaj algoritam mnogo bržim od prethodnog.

C. *Normalizacija 3*

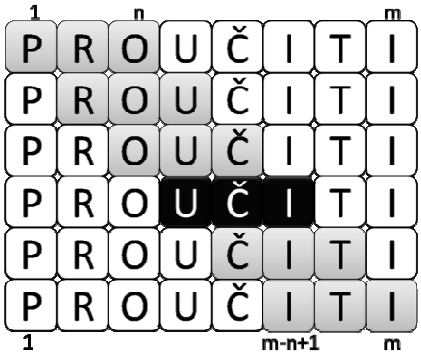
Ovaj algoritam je najjednostavniji, najbrži i ne zahteva nikakve dodatne resurse. Normalizacija reči ovim algoritmom vrši se izdvajanjem *k* prvih slova iz reči. Stoga, ostale reči čija je dužina manja od *k* ne ulaze u razmatranje. I ovde ostaje nerešen problem prefiksa i drugih gramatičkih specifičnosti.

D. *Normalizacija 4- n-gram analiza reči*

Za razliku od prethodna dva pristupa, ovaj tip normalizacije rešava problem prefiksa. Jedan je od jednostavnijih algoritama, ali se njegovom primenom broj normalizovanih reči značajno povećava. Budući da je najveći značaj u radu stavljen upravo na ovaj algoritam, n-gram analiza će biti nešto detaljnije opisana u odnosu na prethodne pristupe.

N-gram je podniz koji se sastoji od *n* elemenata datog niza. N-gram modeli se, u kompjuterskoj lingvistici, koriste najčešće za predviđanje reči ili slova u aplikacijama različitih namena. [3] Na primer, reč „učiti“ sastavljena je od sledećih ngrama: u-č-i-t-i (dužina 1), uč-či-it-ti (dužina 2), uči-čit-iti (dužina 3), učit- čiti (dužina 4) i učiti (dužina 5).

N-gram analiza je postupak koji se primenjuje na tekstu, a njegov rezultat je dobijanje skupa n-grama određene dužine. N-grami se dobijaju pomeranjem okvira dužine n, čiji početak može biti na pozicijama od 1 do m-n+1, pri čemu je m dužina stringa(Sl.3).



Sl. 3. Izdvajanje n-grama iz reči

N-gram analiza se može koristiti kao alternativno rešenje za pronalaženje reči čiji je koren isti (Sl. 4). Tako, naizgled računaru različite reči, „proučiti“, „naučim“, „uči“, imaju zajednički 3-gram „uči“, i semantički, sve su vezane za glagol „učiti“ od koga su i nastale.

U cilju dobijanja efikasnijih rezultata prilikom izbora broja *n*, sakupljeni su nastavci za oblik u srpskom jeziku za reči od važnosti [4]. Tom prilikom utvrđeno je da najveći procenat (92,70 %) nastavaka ima ispod 4 slova. [2] Stoga je u ovom radu vršena je 4-gram analiza.

N-gram analiza primenjena je nad rečima, a ne na rečenicama, tako da se praznine ne mogu naći u okviru

jednog n-grama. To bi povećalo broj podataka u bazi podataka, a samim tim i vreme obrade zahteva. Mada takav tip n-gram analize ima, u nekim jezicima, više smisla, zbog pronalaženja čestih fraza ili sintagmi. Ovaj problem je rešen na drugi način, računanjem matrice uzajamnog ponavljivanja reči, koja je detaljnije opisana u narednom poglavlju.

```
procedure GENERATENGRAMS(Sentence,N)
NGrams[0]={∅}
NGramCount=0
foreach Word ∈ Sentence
        while (NumOfLetters(Word) >=N)
                NGarms[NGramCount]= FirstNLetters(Word)
                Word = RemoveFirstLetter(Word)
                NGramCount=NGramCount + 1
return NGrams
```

Sl. 4. Pseudo kod n-gram analize

IV. IZDVAJANJE KLJUČNIH REČI I N-GRAMA

Ključne reči su termini koji sadrže najvažnije informacije iz dokumenta. Automatsko izdvajanje ključnih reči predstavlja pronalaženje malog skupa reči i fraza koje opisuju sadržaj dokumenta. Postojeće metode za automatsko izdvajanje ključnih reči mogu se podeliti u četiri grupe:

- jednostavan statistički pristup (statističke informacije o rečima zasnovane na učestalosti pojavljivanja reči, tf*itf, matrici uzajamnog pojavljivanja itd.)
- lingvistički pristup (leksička analiza, sintaksička analiza)
- pristup zasnovan na mašinskom učenju (na osnovu skupa dokumenata sa već izdvojenim ključnim rečima generiše se model za pronalaženje ključnih reči u novom dokumentu)
- poziciono težinski pristup (reči na različitim pozicijama imaju različit stepen značajnosti)[5]

Često korišćeni algoritam za indeksiranje tf*idf zahteva postojanje korpusa dokumenata, da bi se pomoću njega dobile ključne reči. I ostali navedeni pristupi za automatsko izdvanje reči zahtevaju odgovarajuće leksičke resurse ili rad sa označenim dokumentima.

U nedostatku korpusa ključne reči se mogu izdvojiti na osnovu broja pojavljivanja termina u posmatranom dokumentu. Dalje je opisan algoritam za izdvajanje ključnih reči iz jednog dokumenta, ne zahtevajući korpus.

Da bi se pristupilo izdvajanju ključnih reči, tekst mora biti normalizovan. Kao što je prethodno navedeno, tekst je očišćen od stop reči i svih neslovnih oznaka i podeljen na rečence. Najčešće pojavljivani termini se izdvajaju. Izdvajanje ključnih reči je vršeno na dokumentu koji predstavlja opis sistema za automatsko davanje odgovora na pitanja građana koje je zasnovano na klasterovanju.[6] U tabeli 1. prikazano je deset najzastupljenijih termina u dokumentu pri čemu je primenjena normalizacija 1.

Rezultati normalizacije vršene odsecanjem na n-grame dužine četiri (normalizacija 3.) prikazani su u Tabeli 2. Prikazano je prvih 10 ngrama čiji broj pojavljivanja odstupa za dve standardne devijacije od prosečnog pojavljivanja ngrama u celom dokumentu. Ovom analizom su sve reči kraće od četiri slova isključene iz dalje analize, izuzev reči koja sadrže (ć, č, š, đ i ž) koja se pri obradi tretiraju kao dva znaka (rešenje ovog problema će biti predmet daljeg istraživanja) .

TABELA 1. KLJUČNE REČI (NORMALIZACIJA 1)

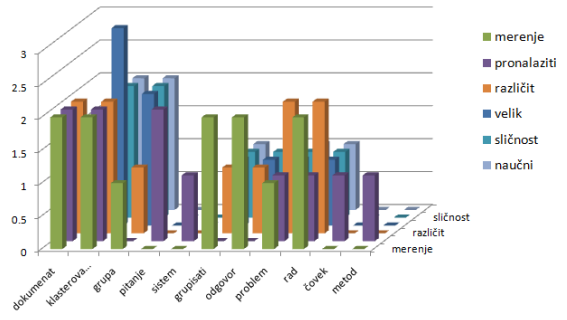
Termin	Broj pojava
dokument	12
klasterovanje	9
grupa	6
pitanje	5
sistem	5
grupisati	3
odgovor	4
problem	4
rad	4
sličnost	4

TABELA 2. KLJUČNE REČI (NORMALIZACIJA 3)

Termin	Broj pojava
doku	12
grup	9
klas	9
slič	6
sist	5
razl	5
prim	4
pita	4
odgo	4
pred	3

Poređenjem tabela 1 i 2 može se uočiti da se početni tetragrami ključnih reči prikazanih u prvoj tabeli nalaze i u drugoj. Skraćivanjem reči na 4 karaktera broj pojmova obuhvaćenih ključnim rečima se povećao, jer, na primer, reči „grupa“ i „grupisati“ su skraćene na istu osnovu, a nova ključna reč je n-gram „pred“ koji je početak reči „predstaviti“ i „predmet“.

Dokument se sastoji iz rečenica i pojavljivanje termina se računa na nivou rečenice. Izdvojene ključne reči su osnova daljeg traženja ključnih reči i izraza. Ukoliko se neki n-gram javlja češće u rečenicama sa određenim ključnim rečima onda on bliže određuje te ključne reči i kandidat je takođe za ključnu reč. Prikazan je grafik raspodele broja rečenica u kojima se neki neključni termini javlja u istoj rečenici u kojoj se javalja određena ključna reč.(Slika.5)



Sl. 5. Uzajamno pojavljivanje ključnih i neključnih termina u rečenici

Za izdvajanje ovakvih n-grama izračunava se broj rečenica u kojima se zajedno javljaju neka od ključnih reči i ostali n-grami i pravi se matrica uzajamnog pojavljivanja. Ukoliko se n-gram pojavljuje selektivno uz samo neke ključne reči to je indikacija da je n-gram sa većim značajem. Računanje odstupanja od očekovanih vrednosti vršeno je hi-kvadrat testom. Očekivana vrednost se računa kao proizvod n_w, ukupnog broja reči u rečenicama u kojima se pojavljuje posmatrani n-gram(w) i p_q, procenta ukupnog broja reči u rečenicama u kojima se pojavljuje ključni n-gram(q) u celom dokumentu (Formula 1.).[7]

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,q) - n_w * p_q)^2}{n_w * p_q} \qquad (1)$$

Ukoliko ima nekih grupa reči koje se često koriste zajedno njihovi n-grami će biti izdvojeni. N-grami sa najvećim hi-kavdrat testom su dodatni kandidati za ključne reči. Izdvajanjem samo prvih n-grama zanemaruje se ostatak reči i može se desiti da neke reči koje se često pojavljuju budu zanemarene zbog različitih prefiksa, dok s druge strane reči različitog značenja mogu imati isti prefiks pa će i taj ngram biti uključen u ključne reči. Da bi se mogle uočiti i ovakve specifičnosti rađena je n-gram analiza reči (normalizacija 3). Za prvi skup ključnih n-grama uzimaju se samo početni n-grami, ali broj njihovog pojavljivanja se računa nezavisno od pozicije u tekstu. Stoga ukoliko je i neka duža reč od četiri slova značajna reč, ona će biti identifikovana izdvajanjem svih n-grama koji se u njoj pojavljuju jer će oni pokazati veći broj pojava u rečenicama u kojima se pojavljuje prvi n-gram te reči.

Za prikaz rezultata uzet je za ulazni dokument tekst ovog rada. Na slici 6. prikazani su dobijeni rezultati početnog skupa ključnih reči (najzastupljeniji n-grami) i 10 neključnih n-grama sa najvećom vrednošću hi-kvadrat testa. Za izdvajanje oba skupa ključnih reči tekst je normalizovan normalizacijom 3.

Database: ceotekst, Table: kljucne

rec	br_pojava	p	id_dokumenta
reč	125	0.7559006211180124	1
doku	62	0.5670807453416149	1
klju	39	0.41055900621118013	1
n-gr	32	0.33354037267080744	1
anal	28	0.33043478260869563	1
norm	28	0.28819875776397513	1
izdv	23	0.3167701863354037	1
poja	22	0.27391304347826084	1
broj	19	0.2906832298136646	1
kori	16	0.2683229813664596	1

Database: ceotekst, Table: hi_kvadrat

rec	id_dokumenta	hi
grup	1	306.3440311044516
teks	1	285.09953533722427
osno	1	280.6863346662924
pret	1	278.645280519061
term	1	245.89603836551552
nač	1	241.5807678709709
algo	1	229.03222967046855
rezu	1	220.36668406940188
zaje	1	218.74763868203075
razl	1	212.46184413878026

Sl. 6. Ključne reči (normalizacija 3)

Rezultati proširivanja skupa ključnih reči primenom hi-kvadrat testa na tekstu koji je normalizovan n-gram analizom (normalizacija 4) prikazani su na slici 7.

Database: ngramceotekst, Table: kljucne

rec	br_pojava	p	id_dokumenta
reč	127	0.7835913312693498	1
doku	63	0.576625386996904	1
klju	42	0.38544891640866874	1
n-gr	32	0.31160990712074305	1
norm	28	0.3109907120743034	1
anal	28	0.323374613003096	1
izdv	23	0.31037151702786375	1
poja	22	0.22275541795665635	1
broj	19	0.23823529411764705	1
kori	16	0.24148606811145512	1

Database: ngramceotekst, Table: hi_kvadrat

rec	id_dokumenta	hi
eči	1	1897.8706675417066
okum	1	1462.2431217080425
kume	1	1462.243121708042
umen	1	1462.243121708042
aliz	1	1425.1672591305557
anje	1	1380.2153867541047
acij	1	1275.7967025136315
gram	1	1236.749936453375
juč	1	1213.5560150656222
janj	1	1189.6367482049006

Sl. 7. Ključne reči (normalizacija 4)

Može se primetiti da je skup ključnih reči u prvom slučaju dopunjen novim terminima („grup“,“tekst“,“algo“) dok u drugom slučaju su dodati n-grami koji dopunjavaju ključne termine (“reč”, ”eči” ; ”doku”, ”okum”, ”umen”, ”ment”).

V. TEHNOLOGIJE I ALATI

Aplikacija za analizu dokumenata je Web aplikacija pa je shodno tome implementacija algoritama urađena u programskom jeziku PHP. Podaci su smešteni u MySQL bazu podataka. Korišćen je Apache HTTP server.

VI. PROŠIRENJA RADA U BUDUĆNOSTI

Algoritmi koji su opisani predhodno imaju za cilj da olakšaju pripremu dokumenata za pretraživanje. Prikazane normalizacije dokumenta i algoritam izdvajanja ključnih reči smanjuju zahteve za velikim leksičkim resursima. Algoritam treba dalje dopuniti klasterovanjem ključnih reči na osnovu njihove frekvencije i raspodele pojavljivanja u rečenicama zajedno sa drugim terminima, čime će se izdvojiti oni n-grami koji imaju veliku frekvenciju i veliku vrednost hi-kvadrat testa, kao glavni kandidati za ključne reči. Nakon dobijenog skupa ključnih reči dokumenti se mogu grupisati na osnovu sadržaja.

LITERATURA

[1] Živojin Stanojčić, Ljubomir Popović,GRAMATIKA SRPSKOG JEZIKA za gimnazije i srednje škole od 1. do 4. Razreda, Zavod za udžbenike

[2] Munirul Mansur, Naushad UzZaman ,Mumit Khan “Analysis of n-gram based text categorization for Bangla in a newspaper corpus” , ICCIT 2006, Dhaka, Bangladesh, decembar 2006.

[3] Ejub Kajan, Aldina Pljasković, Adela Crnišanin „Normalization of text documents in serbian language for eficient searching in e-government systems“,ETRAN 2012, Zlatibor , jun 2012

[4] D. Subotić, N. Forbes, “*Serbo-Croatian language – Grammar*”, Oxford : The Clarendon press, str.25-31, 61-64, 101-113

[5] Kaur, J. and V. Gupta (2010). Effective Approaches for Extraction of Keywords. *International Journal of Computer Science Issues*, 7(6), pp. 144-148.

[6] Ulfeta Marovac, Ejub Kajan, Goran Šimić,” A solution of semantic clustering of text documents ”,CPPMI 2012, Novi Pazar, jun 2012

[7] Matsuo, Y. and M. Ishizika (2004). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information. *International Journal on Artificial Intelligence Tools.* Vol. 13, No 1, pp. 157-169.

ABSTRACT

The modern way of life, e-business, a large amount of data available in electronic form imposed the need for analysis of textual documents written in different natural languages. Every natural language has many rules and variations which makes analysis of the document more difficult. By N-gram analysis of documents, the results can be obtained without specific lexical resources. In this paper, the n-gram analysis of textual documents written in Serbian language is shown and also the algorithm for extracting keywords (n-grams) from a document.

N-GRAM ANALYSIS OF TEXT DOCUMENTS IN SERBIAN LANGUAGE

Ulfeta Marovac, Aldina Pljasković, Adela Crnišanin, Ejub Kajan