Implementation of text mining techniques using NLTK in Python programming language

Ljudmila Petkovic

Implementacija tehnika analize teksta u programskom jeziku Python pomoću NLTK-a

Ljudmila Petković

***Sadržaj* — U ovom radu razmatraju se osnovne tehnike analize tekstova upotrebom NLTK (Natural Language ToolKit) biblioteka za obradu prirodnih jezika pomoću programskog jezika Python. Demonstriraćemo algoritme za izdvajanje dokumenata, generisanje konkordanse, leksičke frekvencije, diverziteta, disperzije, najučestalijih reči, pronalaženje bigrama, kolokacija i reči određene dužine, radi pružanja novog uvida u tekstove. Rezultati pomenutih metoda donose prikaz jednog dela širokog spektra funkcionalnosti NLTK paketa, na koji se informatička zajednica često oslanja tokom raznorodnih projekata u okviru računarske lingvistike.**

***Ključne reči* — NLP, NLTK, Python, računarska lingvistika, text mining.**

I. Uvod

Globalno posmatrano, analiza tekstova (*text mining*) pronalazi odgovarajuću primenu u obradi polustrukturiranih ili nestrukturiranih podataka, kojima je savremeni čovek digitalnog doba ophrvan. Okosnicu ove moderne i relativno mlade naučne discipline svakako čini izdvajanje relevantnih informacija iz velike količine raznolikih podataka, ali i uočavanje jezičkih šablona koji mogu biti od značaja za dalju sadržinsku analizu tekstualne jedinice [1]. Njena interdisciplinarnost ogleda se u sprovođenju metodoloških principa prevashodno računarske lingvistike, statistike, pronalaženja informacija, mašinskog učenja, *data mining*-a [2], ali i eksperimentalne psihologije, između ostalog [3]. Kao što sam naziv sugeriše, *text mining* predstavlja jedan vid „rudarenja teksta" [4], postupak koji je naročito zahvalan u situacijama kada se analitičar podataka susretne sa korpusima dokumenata i leksičkim resursima. Primera radi, ukoliko bi nas zanimalo da saznamo u kom se sve kontekstu javlja izvesna reč u knjizi, jasno je da bi linearna pretraga ciljane reči i njeno izdvajanje sa književnim okruženjem prerasli u mukotrpan proces. Iz tog razloga, u cilju automatizacije takvih i sličnih radnji, stvoreni su posebni alati za inteligentnu obradu teksta, poput *NLTK*-a *(Natural Language ToolKit*-a). Zvanična NLTK knjiga [5], uz prateću dokumentaciju za korišćenje ovog softvera otvorenog koda [6], predstavlja pogodnu polaznu osnovu za implementaciju osnovnih tehnika *text mining*-a. Shodno tome, u narednim poglavljima će kroz praktične primere biti opisane funkcionalnosti NLTK paketa u okviru šire discipline – obrade prirodnih jezika, poznatije kao *NLP* (*Natural Language Processing*). Kodovi su pisani u programskom jeziku Pajton (eng. *Python*) verzije 3.6.5 i testirani u Anaconda 4.5.4 interaktivnom razvojnom okruženju; izvorni kod za ovo istraživanje dostupan je na stranici https://nbviewer.jupyter.org/github/ljpetkovic/Telfor/blob/master/textmining.ipynb.

Ljudmila Petković, Filološki fakultet u Beogradu, Studentski trg 3; Studije pri Univerzitetu u Beogradu, Studentski trg 1, 11000 Beograd, Srbija (telefon: 381-62-8301625, e-mail: ljudmila.petkovic@gmail.com).

II. Pronalaženje informacija

Da bi se otpočeo proces manipulisanja tekstom, u interaktivno okruženje Pajtona najpre je neophodno uneti sledeće naredbe za preuzimanje celokupnog NLTK paketa:

```
import nltk
nltk.download()
from nltk.book import *
```

Sl. 1. Kod za preuzimanje NLTK paketa.

Poslednja linija koda omogućava uvoz raznih korpusa tekstova (od delâ beletristike do skupa čet poruka), nad kojima je zatim moguće vršiti odgovarajuće *text mining* metode, detaljnije izložene u nastavku.

A. Konkordansiranje

Za početak, razrešićemo uvodni zahtev u vezi sa izdvajanjem određene reči (ili *tokena*, sa stanovišta programiranja, o čemu će biti više reči u trećem poglavlju), i konteksta u koji je smeštena. Metod poznat u stručnoj literaturi pod nazivom *KWIC* (*Key Word in Context*) *konkordansa* [7] ostvaruje se komandom `text2.concordance('affection')`, koja praktično zahteva da se iz romana *Razum i osećajnost* Džejn Ostin izdvoji u kojim se sve kontekstima javlja tražena reč 'affection'. Tako će u prvoj rečenici izlaznog rezultata levi kontekst te ključne reči glasiti: "*however, and, as a mark of his",* a desni *"for the three girls, he left them*", i ovaj postupak ekstrakcije jedne leksičke jedinice i njenog neposrednog jezičkog okruženja primenjuje se i na ostale konkordansirane rečenice.

Ipak, broj pojavljivanja ove reči ukazuje na njen relativno nizak značaj na nivou čitavog teksta – u procesu konkordansiranja pronađeno je ukupno 25 pogodaka, dok procentualna zastupljenost dostiže okvirno 0,06%. Poslednji parametar je prikazan sledećim kodovima:

```
def zastupljenost(string, substring):
    a = string.count(substring) / len(string) * 100
    print(a)
zastupljenost(text2, 'affection')
```

Sl. 2. Kod za prikaz zastupljenosti reči u tekstu.

Sa druge strane, dalja istraživanja dozvoljavaju

978-1-5386-7171-9/18/$31.00 ©2018 IEEE

upotrebu funkcije konkordanse i u svetlu konfirmativne analize podataka, kako bi se utvrdile ili osporile izvesne pretpostavke na osnovu predznanja analitičara o postojanju jezičkih varijanti jednog istog jezika. Preciznije, da bi se utvrdilo na koji način se menja značenje jedne iste reči u različitim kontekstima britanske i američke verzije engleskog jezika, korisno je uporediti tekstove napisane tim dijalektima – u našem slučaju, to će biti britanski *Monti Pajton i Sveti gral*: `text6.concordance('bloody')` odnosno *Mobi Dik* američkog književnika Hermana Melvila: `text1.concordance('bloody')`. Slede parcijalni rezultati konkordansiranja ove reči u britanskom i američkom engleskom jeziku u Tabeli 1:

TABELA 1: KONKORDANSA – BRITANSKI I AMERIČKI ENGLESKI.

Monti Pajton	***Mobi Dik***
Bloody peasant!	Bloody battle in Afghanistan
Bloody weather!	bloody hunt of whales
We live in a bloody swamp!	bloody deed he had planned
Oh, bloody hell!	bloody-minded soldadoes

Kao što se može zapaziti, dolazi do vidljive diferencijacije značenja reči 'bloody' u ova dva varijeteta, pri čemu se britanska verzija u ovom delu odlikuje značenjem „proklet", kao u rečenici "*Bloody* peasant!", ili „prokletstvo/đavola" u "Oh, *bloody* hell!". Sa druge strane, Melvilovo značenje ove leksičke jedinice stoji u vezi sa krvlju ili krvoločnošću: '*bloody* deed' (krvavi čin), '*bloody*-minded soldadoes' (krvoločni vojnici). Iz ovoga proizlazi da je u *Montiju Pajtonu* ova reč smeštena u neformalni kontekst, gde se ispoljava sekundarno pogrdno značenje (iako 'bloody' i u britanskom engleskom nosi osnovno značenje 'krvav'), dok američkoj varijanti ovaj diskursni fenomen nije svojstven, već će Melvil ovu reč upotrebiti samo u bukvalnom smislu. Prema tome, na osnovu rezultata konkordansiranja može se uvideti da određena reč često nosi drugačije značenje u zavisnosti od jezičkog varijeteta.

III. LEKSIČKA FREKVENCIJA

A. Značaj broja reči u tekstu

Najpre ćemo izvršiti terminološku distinkciju između *tokena* – sačinjenog od jednog ili niza karaktera (koji, zacelo, mogu predstavljati i konkretnu reč), gde se broje sva pojavljivanja tih grupa u tekstu, čak i ako su iste morfološke strukture – i *različitih leksičkih stavki* (*distinct vocabulary items*), čija frekvencija je uvek jednaka 1 prilikom njihovog prebrojavanja [5]. Ova razlika može se demonstrirati komandama, `print(sorted((text3)))` i `print(sorted(set(text3)))`, pri čemu bi se u prvom slučaju sva pojavljivanja reči npr. 'Adam' iz *Knjige postanja* tretirala kao 18 zasebnih leksičkih jedinica, a u drugom kao jedna reč.

Imajući u vidu tu razliku, u ovom odeljku posvetićemo više pažnje izračunavanju leksičke frekvencije, prebrojavanjem svih tokena i različitih reči iz teksta romana *The Man Who Was Thursday* od G. K. Čestertona, koji je prvobitno prošao kroz etapu preprocesiranja u vidu tokenizacije, uklanjanja interpunkcijskih znakova i funkcionalnih reči, uz svođenje velikih na mala slova. Pomenutoj metodi smo pridodali i računanje diverziteta teksta, kao odnosa različitih reči i svih tokena. Objedinjeni algoritam je dat u nastavku:

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

def statistika(text):
    tokens = [word_tokenize(word) for word in gutenberg.words(text)]
    tokens = [token.lower() for token in gutenberg.words(text)]
    tokens = [token for token in tokens if token.isalpha()]
    stoplist = stopwords.words('english')
    tokens = [w for w in tokens if not w in stoplist]
    print(len(tokens), len(set(tokens)))
    a = len(set(tokens)) / len(tokens) * 100
    print(a)
statistika('chesterton-thursday.txt')
```

Sl. 3. Kod za računanje broja reči i leksičkog diverziteta.

Ukupan broj tokena prečišćenog teksta iznosi 28328, koji sadrži 6159 različitih reči. Procena raznovrsnosti vokabulara romana izražena je u procentima – 21,74%. Vredno pomena jeste statističko poređenje Čestertonovog romana sa *Knjigom postanja*, za koju je istim principom izračunato da ima 18335 tokena i 2495 različitih leksičkih jedinica, što rezultira manjim leksičkim diverzitetom od približno 13,61%.

Paralelno, skrećemo pažnju i na proces izračunavanja procentualnog udela koji konkretni token zauzima u informacionoj stavki, što predstavlja dodatni kriterijum za određivanje raznovrsnosti vokabulara. U prethodnom pasusu zaključili smo da se *Knjiga postanja* odlikuje relativno niskim stepenom leksičkog diverziteta, a u svrhu njegovog dodatnog isticanja primenili smo istu funkciju kao u prošlom poglavlju koji se tiče zastupljenosti jedinice u tekstu: `zastupljenost(text3, 'God')`, na osnovu koje saznajemo da token 'God' zauzima 0,52% u celokupnoj tekstualnoj jedinici. Zatim, klasično prebrojavanje apsolutnih pojavljivanja ovog tokena unutar teksta formulisano je sledećom funkcijom, koja proizvodi rezultat od 231 javljanja:

```
def frekvencija(string, substring):
    words = string.count(substring)
    print(words)
frekvencija(text3, 'God')
```

Sl. 4. Kod za računanje frekvencije reči 'God' u tekstu.

Gledano sa statističke strane, *Knjiga postanja* se ne odlikuje bogatstvom reči, ali to svakako ne umanjuje njenu moralno-religijsku vrednost koja se ostvaruje drugim književnoumetničkim sredstvima, što donekle govori i o varljivoj prirodi statističkih podataka.

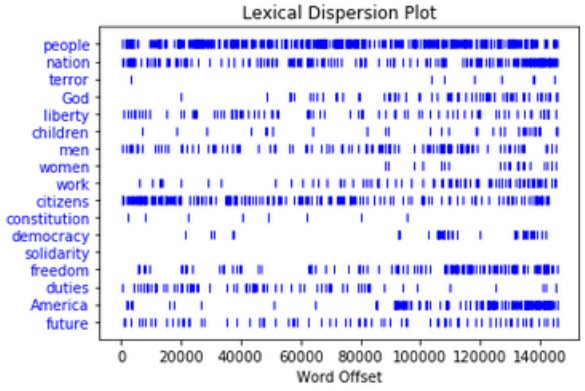
B. Vizualizacija leksičke disperzije unutar teksta

Pored prethodno navedenih komandi, *Pajton* može grafički predstaviti i leksičke tendencije prisutne u tekstovima, što ćemo pojasniti na primeru korpusa inauguracionih govora američkih predsednika (*Inaugural Address Corpus 1789-2009*). U sklopu konfirmativne analize, odabrano je šesnaest reči za koje se pretpostavilo da bi mogle biti zastupljene u govorima predsednika:

```
text4.dispersion_plot(['people', 'nation', 'terror', 'God', 'liberty', 'children',
'men', 'women', 'work', 'citizens', 'constitution', 'democracy', 'solidarity',
'freedom', 'duties', 'America', 'future'])
```

Sl. 5. Kod za predstavljanje leksičke disperzije (16 reči).

i čija je leksička disperzija ilustrovana *grafikom leksičke disperzije* (eng. *lexical dispersion plot*), prikazanom na Sl. 6.



Sl. 6. Grafik leksičke disperzije (inauguracioni govori).

Beleži se svako pojavljivanje i tačno mesto ponuđenih reči u čitavom korpusu, predstavljenom kao neprekinuti niz svih tekstova govora. Osim toga, iz priložene vizualizacije se vidi da su predsednici SAD-a veoma često upotrebljavali reč 'people', 'nation', 'citizens', 'freedom', 'America', kao i da se veća učestalost pojavljivanja tih reči na slici predstavlja zadebljanim i gusto zbijenim plavim uspravnim linijicama, koje sa većom koncentracijom naizgled poprimaju oblik kvadrata ili pravougaonika.

Algoritmizacijom u vidu prebrojavanja biranih tokena pomoću funkcije `frekvencija` iz prethodnog potpoglavlja možemo ispitati učestalost njihovog javljanja (npr. 'God' se javlja 97 puta, 'democracy' 52 itd.). Međutim, pažljivijom analizom gornjeg grafika iz suprotne perspektive, odnosno, potragom za ređe korišćenim rečima, može se ustanoviti još jedan zanimljiv trend, a on se ogleda u primetno smanjenoj upotrebi reči 'women' (28 pojavljivanja). Ova informacija bi mogla da se posmatra kroz prizmu korenitih promena po pitanju prava glasa žena u SAD-u, koje su obeležile dug period od kraja XVIII veka sve do 1984. godine, kada je taj proces okončan stupanjem na snagu tog prava i u državi Misisipi [8]. Imajući u vidu pomenuta političko-istorijska previranja, ne čudi i relativna marginalizacija žene u predsedničkim govorima, u kojima prednost nosi reč 'men' (139 javljanja).

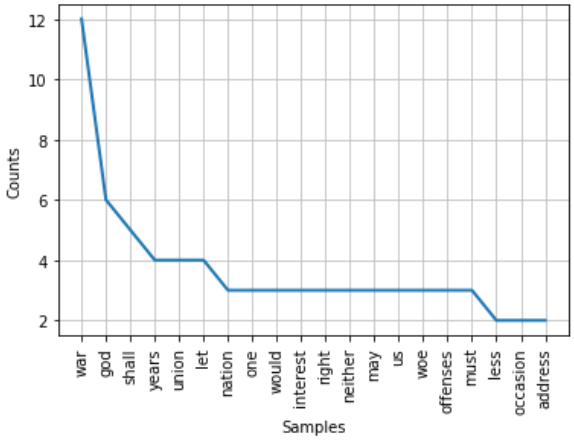
C. *Najfrekventnije reči*

Prethodni grafik leksičke disperzije može poslužiti kao uvod u demonstriranje algoritma za generisanje liste dvadeset najčešćih reči u drugom inauguracionom govoru Abrahama Linkolna iz 1865. godine.

```
def najcesce_reci(text):
    tokens = [word_tokenize(word) for word in inaugural.words(text)]
    tokens = [token.lower() for token in inaugural.words(text)]
    tokens = [token for token in tokens if token.isalpha()]
    stoplist = stopwords.words('english')
    tokens = [w for w in tokens if not w in stoplist]
    fdist = FreqDist(tokens)
    fdist1 = fdist.most_common(20)
    fdist.plot(20, cumulative  = False)
    print(fdist1)
najcesce_reci('1865-Lincoln.txt')
```

Sl. 7. Kod za generisanje liste najčešćih reči u tekstu.

Generisana je lista najfrekventnijih reči: [('war', 12), ('god', 6), ('shall', 5), ('years', 4), ('union', 4), ('let', 4), ('nation', 3), ('one', 3), ('would', 3), ('interest', 3), ('right', 3), ('neither', 3), ('may', 3), ('us', 3), ('woe', 3), ('offenses', 3), ('must', 3), ('less', 2), ('occasion', 2), ('address', 2)], upotpunjena vizualnim prikazom u koordinatnom sistemu, kao na Sl. 8:



Sl. 8. 20 najučestalijih reči, *Inaugural Address Corpus*.

Poznavalac američke političke istorije primetiće jasne naznake centralnih tema Linkolnovog govora, u kome se, između ostalog, navodi da je Američki građanski rat sredinom XIX veka bio posledica Božjeg proviđenja, kojim se Amerikanci kažnjavaju zbog svoje robovlasničke prošlosti. Pored najfrekventnijih reči 'war' i 'God', glavnih nosilaca te ideje, govor je poznat i po obilju biblijskih referenci koje sadrže ključne reči 'woe', 'God', 'offenses', ali i arhaičniji glagolski oblik u američkom engleskom 'shall', dok je u velikoj meri prisutan i termin 'Union', koji denotira Linkolnovu vladu (za iscrpnija tumačenja ovog govora v. [9]). Prema tome, prethodna razmatranja mogu u velikoj meri otkriti preovlađujuće tematske šablone, koji olakšavaju put do lociranja relevantnih sadržaja za određeni dokument.

IV. Stilistička analiza

A. *Značaj dužine reči u tekstu*

Za kraj, obratićemo pažnju na piščev stil kojim se odlikuje neko književnoumetničko delo, a čije osobenosti možda nisu odmah uočljive prilikom uobičajenog sekvencijalnog čitanja. Naime, jedno od istraživačkih pitanja može glasiti: da li se određeni roman teže ili lakše čita u zavisnosti od izbora reči? Pretpostavimo da analitičar vrši poređenje dela *Mobi Dik* i *Razum i osećajnost*, na osnovu zastupljenosti veoma dugačkih reči koje sadrže petnaest ili više karaktera. Takav zahtev podrazumeva izdvajanje setova različitih reči, od kojih treba prikazati samo one koje zadovoljavaju dati kriterijum dužine, uz navođenje njihove frekvencije u tekstu. Ova ideja je jezgrovito izražena *Pajton* sintaksom:

```
def duge_reci(text):
    V = set(text)
    long_words = [w for w in V if len(w) >= 15]
    print('Broj dugačkih reči u tekstu: ', len(long_words), '\n''\n',
sorted(long_words))
duge_reci(text2)
```

Sl. 9. Kod za prikaz rečī dužine 15 karaktera u tekstu.

Veliki broj reči, od dvadeset pronađenih koje ispunjavaju zadati uslov za roman Ostinove, relativno su frekventne u standardnom engleskom jeziku, poput ‘acknowledgments’, ‘congratulations’, ‘disqualifications’, ‘incomprehensible’ i dr. Istim postupkom za *Mobija Dika* dobijamo znatno više dugačkih reči (ukupno 72), usled većeg broja strana Melvilovog romana. Gledano iz ovog ugla, moglo bi se reći da ovakva slika nudi jedan pouzdan pokazatelj nesrazmernosti između stilističkih osobenosti u književnom pisanju Ostinove i Melvila, što bi impliciralo da je Melvilov način pisanja visokoparniji ili, pak, poetičniji i nekonvencionalniji, usled krajnje neočekivanih formi kojima je sadržaj reči izražen, poput ‘uninterpenetratingly’, 'skrimshandering', 'notwithstanding', ‘indomitableness’ itd. Pored toga, mogao bi se ustanoviti i visok stepen slikovitosti ovih složenih reči, za koje se ponekad teško pronalazi adekvatan prevod, naročito za apstraktne pojmove, kao što je reč ‘passionlessness’, prevedena na hrvatski jezik kao „bešćutnost” [10] . Stoga, barem nakon ove vrste analize, deluje da bi Melvil mogao da se okarakteriše kao kompleksniji pisac od Ostinove.

B. *Bigrami i kolokacije*

Još jedna zanimljiva pojava koja se tiče analize korpusa jeste i način ulančavanja dveju reči koje se učestalo javljaju zajedno i formiraju jedinstveno značenje. Ukratko, ovakvi ustaljeni izrazi u lingvističkom diskursu poznati su pod nazivom *kolokacije*, kao što su „hemijska olovka” ili „topao doček”. Primena ekstrakcije kolokacija iz dokumenta NLTK korpusa započinje se uvođenjem funkcije *bigramizacije*, tj. uparivanjem dveju susednih reči iz liste tokena, što je preduslov za kasnije prepoznavanje njihovih čestih spojeva komandom `collocations()`.

```
from nltk.util import bigrams
list(bigrams(['Jedan', 'primer','bigramizacije']))
```

Sl. 10. Kod za formiranje bigrama od reči u tekstu.

Nakon testiranja alatke `bigrams()`, rezultat ovog ilustrativnog primera proizvodi bigrame **[(**'Jedan'**,** 'primer'**),** **(**'primer'**,** 'bigramizacije'**)]**. Dalje pronalaženje kolokacija može biti veoma informativno i za analizu političkih govora. Štaviše, naredbom `text4.collocations()` neke od kolokacija su sledeće: *United States; fellow citizens; years ago; Federal Government; American people; Vice President; Old World; Almighty God; Chief Justice; God bless; Indian tribes; public debt; one another; foreign nations; political parties* itd. Uporedo sa standardnim političkim terminima, kao što su ‘Vice President’ ili ‘Chief Justice’, primećujemo i kolokativne strukture sa religijsko-emotivnom konotacijom, poput ‘Almighty God’ i ‘God bless’. Za kraj, napomenućemo da ovakvi leksički skupovi mogu biti predmet opsežnijih istraživanja i u vezi sa temom pažljivog odabira reči koje služe kao posebno sredstvo vršenja uticaja na građansko društvo u toku saopštavanja političkih govora.

V. Zaključak

Odeljak koji se tiče fragmentiranja teksta putem taksativnog nabrajanja svih pojavljivanja konkretne reči u kontekstu ostavlja prostora za generisanje drugih specifičnih informacija u okviru eksplorativnog pristupa, poput odgovora na pitanje koliki je životni vek mitoloških ličnosti iz *Knjige postanja*: `text3.concordance('lived')`.

U kontekstu srpskog jezika, funkcionalnosti *NLTK*-a su primenjene i radi analize korpusa tekstova pesama iz doba bivše SFRJ. Polazeći od hipoteze da je to doba obeleženo brojnim društveno-političkim previranjima, vršena je pretraga odgovarajućih leksičkih indikatora. Kroz proces pronalaženja reči koje se završavaju npr. sufiksom –ija izdvojile su se reči poput „sankcija“, „malverzacija“, „birokratija“, „milicija“, „partija“ itd. (za proveru konteksta je korišćena metoda konkordansiranja). U vezi as tim, spomenućemo još da su se izdvojile i kolokacije „dan Republike“ i „dižem zastavu“. Pošto ova tema prevazilazi okvire ovog rada, ostavlja se prostora za njenu iscrpniju analizu.

Literatura

[1] G. Miner, J. Elder IV, and T. Hill, *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012, p. XV.

[2] A. Moreno and T. Redondo, “Text analytics: the convergence of big data and artificial intelligence,” *IJIMAI*, vol. 3, no. 6, pp. 57–64, Mar. 2016.

[3] T. W. Miller, *Web and Network Data Science: Modeling Techniques in Predictive Analytics*. Pearson Education, 2014, p. 253.

[4] N. Mirkov and M. Peranović, “Rudarenje teksta sa društvenih mreža API pristupom,” *INFOTEH-JAHORINA,* vol. 14, pp. 584–588, Mar. 2015.

[5] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. Sebastopol, CA: O'Reilly Media, 2009.

[6] S. Bird, NLTK 3.3. documentation [Online]. Dostupno na: http://www.nltk.org (Last updated on May 06, 2018).

[7] S. Th. Gries and J. Newman, “Creating and using corpora,” *Research methods in linguistics*, pp. 257-287, 2013.

[8] M. H. Swain, E. A. Payne, and M. J. Spruill, Eds., *Mississippi women: Their Histories, Their Lives*, *Vol. 2*. University of Georgia Press, 2010, p. 154.

[9] M. Leff, “Dimensions of temporality in Lincoln's second inaugural,” *Communication Reports*, vol. 1, no. 1, pp. 26-31, Winter 1988.

[10] H. Melville, prev. Z. Gorjan, J. Tabak, i P. Mardešić, *Moby Dick ili Bijeli kit*. Zagreb: Školska knjiga, 1999, p. 553.

Abstract

This paper examined text mining techniques using NLTK (Natural Language ToolKit) library for natural language processing in Python programming language. Algorithms hereby proposed addressed this aspect by extracting documents, generating concordance, lexical frequency, diversity, dispersion, list of most frequently used words, as well as by retrieving bigrams, collocations and words of a certain length in order to provide new insight into texts. These methods form only a small part of wide range of NLTK functionalities.

IMPLEMENTATION OF TEXT MINING TECHNIQUES USING NLTK IN PYTHON PROGRAMMING LANGUAGE

Ljudmila Petković