Студије при Универзитету у Београду

Рачунарство у друштвеним наукама

Креирање и анализа корпуса текстова југословенских рок песама у периоду 1945-2003.

Мастер рад

Студент:
Људмила Петковић
Број индекса: 181/2017

Ментор:
Проф. др Ранка Станковић

9. фебруар 2019.

Садржај

Захвалница

Најсрдачније се захваљујем свом ментору, проф. др Ранки Станковић, на изузетном стрпљењу, константном усмеравању, пренетом знању и подршци коју ми је пружила у току израде овог мастер рада. Свесрдну захвалност дугујем и члановима комисије: проф. др Цветани Крстев, чији конструктивни савети су ми били од непроцењивог значаја за боље разумевање проблематике датог истраживања. Велико хвала и проф. Бранислави Шандрих, која је такође од самог почетка учествовала у целокупном процесу осмишљавања и разрађивања теме овог рада, и умногоме помогла у решавању техничких потешкоћа. Реч захвалности упућујем и проф. др Владану Девеџићу, руководиоцу мастер програма Рачунарство у друштвеним наукама, чија реализација је допринела развијању мог изразитог интересовања за овакву врсту интердисциплинарности. Коначно, желела бих да се захвалим и члановима своје породице на неисцрпном разумевању и благонаклоности у свим тренуцима мог бављења науком. Стога ова теза припада и вама.

Људмила Петковић

Сажетак

У раду се са теоријског и практичног аспекта анализира процес образовања и обраде корпуса текстова рок песама из бивше Југославије у периоду 1945-2003. Грађа је преузета са сајта LyricWiki коришћењем Пајтон библиотеке `lyricsmaster` у поступку *гребања веба* (енгл. *web scraping*). Добијени текстови су потом обједињени у јединствену XML датотеку и аутоматски анотирани помоћу Пајтон алата `yattag`. Касније је спроведено препроцесирање корпуса на формалном и садржинском нивоу. Уз то је демонстрирана и трансформација XML документа у XHTML формат коришћењем XSLT процесора, ради генерисања основних података о корпусу. Аутоматизован је и поступак рестаурације дијакритика помоћу апликације „Слово Мајстор", односно морфолошких електронских речника на српском језику у софтверу LeXimir. Рачунарска анализа текста обухватила је проналажење друштвено-политичких и патриотских тема применом `NLTK` библиотеке у Пајтону, док су љубавне и друге теме визуализоване у TreeCloud и WordItOut софтверу. Сличност између аутора заступљених у корпусу мерена је помоћу библиотеке `stylo` у програмском језику R. Најзад, дат је преглед најрелевантнијих програмских библиотека данашњице у области *обраде природних језика* (енгл. *natural language processing*), које истовремено служе аутору као смернице за будући рад.

Кључне речи: корпусна лингвистика, југословенски рокенрол, гребање веба, обрада природних језика, рударење текста.

Abstract

The thesis analyzes, from the theoretical and practical perspective, the creation and processing of corpus of rock songs' lyrics originating from the former Yugoslavia in the period 1945-2003. The lyrics are obtained from the LyricWiki website using the Python library `lyricsmaster` in the *web scraping* process. The collected texts are then merged into a single XML file and automatically annotated with the `yattag` Python tool. Afterwards, the data preprocessing was conducted at the formal and content level. Furthermore, the XML document is transformed into XHTML format applying XSLT processor, in order to generate basic corpus data. The diacritic restoration process with the "Slovo Majstor" application and morphological electronic dictionaries of Serbian language in the LeXimir software package, is also

automated. The *text mining* process encompassed retrieving socio-political and patriotic topics using `NLTK` library in Python, while romantic and other topics were visualized using the TreeCloud and WordItOut software. The similarity between authors represented in the corpus was measured using `stylo` package in the programming language R. Finally, an overview of the today's most relevant programming libraries in the field of *natural language processing* is provided, which, at the same time, serves as a guideline for the future work.

Keywords: corpus linguistics, yugoslavian rock and roll, web scraping, natural language processing, text mining.

Радна биографија

Људмила Петковић рођена је 14. августа 1994. године на Цетињу (Црна Гора). Дипломирала је 2017. године на Филолошком факултету Универзитета у Београду, модул *Грчки језик, књижевност, култура*. На истом факултету је 2018. године стекла диплому мастер академских студија, модул *Језик, књижевност, култура*. Исте године је уписала прву годину докторских академских студија на интердисциплинарном програму *Интелигентни системи* у оквиру Студија при Универзитету у Београду. Као аутор, до сада је објавила два научна рада у вези са примењеном лингвистиком и обрадом природних језика (један у домаћем часопису, а други у међународном зборнику радова 2018. године); трећи рад је исте године саопштен у коауторству на научном конгресу примењене лингвистике у Београду. Током основних и мастер студија била је стипендисткиња Министарства просвете и Фонда за младе таленте Републике Србије. Од 2016. године хонорарно је преводила финансијска документа са грчког, енглеског и српског језика. У школској 2015/2016. години држала је приватне часове грчког језика средњошколцима. У току основних студија допринела је уређивању дидактичког материјала у склопу Moodle платформе, под руководством надлежних професора Катедре за неохеленске студије Филолошког факултета УБ. Учествовала је и на семинару грчког језика и културе као стипендиста грчке државне фондације за стипендије „ИКИ”, 2014.

ВЕЋУ ЗА СТУДИЈЕ ПРИ УНИВЕРЗИТЕТУ
УНИВЕРЗИТЕТА У БЕОГРАДУ

Студијски програм: „РАЧУНАРСТВО У ДРУШТВЕНИМ НАУКАМА“
Наслов тезе: КРЕИРАЊЕ И АНАЛИЗА КОРПУСА ТЕКСТОВА ЈУГОСЛОВЕНСКИХ РОК ПЕСАМА У ПЕРИОДУ 1945-2003.

Ментор:

др Ранка Станковић
ванредни професор
Рударско-геолошки факултет
Универзитет у Београду

Кандидат:

мср Људмила Петковић
Број индекса: 181/2017

Чланови комисије:

Др Цветана Крстев, редовни професор, Филолошки факултет, Универзитет у Београду
Др Владан Девеџић, редовни професор, Факултет организационих наука, Универзитет у Београду

Изјава о академској честитости

Људмила Петковић, број индекса 181/2017, студенткиња мастер академских студија *Рачунарство у друштвеним наукама*. Ауторка мастер рада под називом: „Креирање и анализа корпуса текстова југословенских рок песама у периоду 1945-2003."
Потписивањем изјављујем:

- да је рад искључиво резултат мог сопственог истраживачког рада;
- да сам рад и мишљења других аутора које сам користила у овом раду назначила или цитирала у складу са Упутством;
- да су сви радови и мишљења других аутора наведени у списку литературе/референци који су саставни део овог рада и писани у складу са Упутством;
- да сам добила све дозволе за коришћење ауторског дела који се у потпуности/целости уносе у предати рад и да сам то јасно навела;
- да сам свесна да је плагијат коришћење туђих радова у било ком облику (као цитата, прафраза, слика, табела, дијаграма, дизајна, планова, фотографија, филма, музике, формула, веб сајтова, компјутерских програма и сл.) без навођења аутора или представљање туђих ауторских дела као мојих, кажњиво позакону (Закон о ауторском и сродним правима, Службени гласник Републике Србије, бр. 104/2009, 99/2011, 119/2012), као и других закона и одговарајућих аката Универзитета у Београду;
- да сам да сам свесна да плагијат укључује и представљање, употребу и дистрибуирање рада предавача или других студената као сопствених;
- да сам свесна последица које код доказаног плагијата могу проузроковати на предати мастер рад и мој статус;
- да је електронска верзија мастер рада идентична штампаном примерку и пристајем на његово објављивање под условима прописаним актима Универзитета.

Београд, _____ Потпис студенткиње _____

1 Увод

1.1 Циљ рада

Сврха рада јесте аутоматизација прикупљања текстова песама југословенских рок извођача, чији садржај је доступан на сајту LyricWiki. Предложени алгоритам би био применљив и у случају преузимања текстова за друге ауторе који су архивирани у бази података ове музичке веб стране. Како би се остварила унифицираност и већи степен квалитета корпуса, добијени подаци су прошли кроз фазу препроцесирања, док је анотацијом грађе омогућена даља анализа текста софтверима намењеним у те сврхе.

Други циљ тезе представља истраживање особености „југо-рока" са становишта рачунарске анализе текста, односно покушај доношења закључака по питању одлика текстуалних садржаја из тог жанра, који до сада, како се чини, нису обрађивани електронским путем. Такође, по узору на достигнућа из области корпусне лингвистике у Србији, тежимо и да допринесемо богаћењу постојећих колекција рачунарски обрадивих текстова на српском језику, чијим радом руководи српско Друштво за језичке ресурсе и технологије (JePTex)[1]. У ужем смислу, резултати датог пројекта могли би бити од користи онима који намеравају да кроз пројекте креирања и обраде корпуса текстова музичких дела спроводе истраживања на лингвистичком плану.

Најзад, кроз представљање исхода овог рада настојимо да повећамо интересовање и за примену метода *ископавања из текста* (енгл. *text mining*) када је реч о музичким делима, услед недостатка сличних пројеката на нашим просторима. Исто тако, верујемо да на тај начин можемо у извесној мери приближити југословенску рок сцену и оним љубитељима музике (првенствено млађим генерацијама) који нису довољно упознати са концептом „ex-Yu рока".

1.2 Предмет рада

Подручје истраживања овог рада представља рачунарска израда и анализа корпуса текстова југословенских рок песама, насталих у доба двеју југословенских држава – Социјалистичке Федеративне Републике Југославије (1945-1992) и Савезне Репу-

[1] http://jerteh.rs.

блике Југославије (1992-2003[2]. Најстарији албум обухваћен корпусом јесте *Наше доба* (1967) групе Индекси; са друге стране, најскорије објављени албуми заступљени на истом корпусу јесу албуми *Од неба до неба* групе Дивље Јагоде, односно *Collection* Нине Бадрић (оба издата 2003). Терминолошки посматрано, прва метода која се тиче прикупљања података назива се *гребање веба* или *стругање веба* (енгл. *web scraping*). Истраживање ће бити спроведено применом метода анализе текста у виду *обраде природних језика* (енгл. *natural language processing* – скр. *NLP*) и *ископавања из*/*истраживања*/*рударења текста* (енгл. *text mining*)[3].

1.3 Истраживачка питања

Идеја о развоју овог корпуса настала је из жеље да се кроз рачунарску анализу прикупљених текстова дубље проучи феномен југословенског рокенрола, који се, по мишљењу слушалаца и стручне јавности, у другој половини XX века истицао својим идеолошким, естетским и авангардним обележјима. Наиме, верујемо да би било значајно дати одговоре на следећа питања:

- Које су преовлађујуће теме у текстовима песама?
- Колики је удео заступљености речи које упућују на љубавне, друштвено-политичке или патриотске теме?
- Како можемо квантитативно измерити сличности међу извођачима према садржају текстова чије песме интерпретирају?
- У којој мери се у текстовима користе стране речи?
- Да ли се постојеће интерпретације текстова песама могу допунити квантитативним резултатима добијеним компјутерским путем?
- Због чега је „ex-Yu” рокенрол уживао велику популарност и шта нам добијени резултати говоре о том жанру?

[2]Према хронологији коју наводе Маловић и Јончић (исто, 2010: 5)

[3]Више о разлици између *рударења текста* и *обраде природних језика* в. Expert System Team, April 11, 2016.

1.4 Хипотезе истраживања

Иако је „ex-Yu" рок извршио огроман утицај на тадашњу југословенску омладину, инспирисану западњачким духом којим је ова врста музике одисала, власт у бившој Југославији се због тога озбиљно противила развоју новог музичког жанра.[4] Имајући у виду постојање бројних препрека и контроверзи са којима се ова врста музике суочавала, претпоставили смо да ће одређени текстови „ex-Yu" корпуса бити обојени друштвено-политичким темама. Са друге стране, пошто уплив музике са Запада у југословенски рокенрол није био занемарљив, у раду се заступа хипотеза да ће у текстовима бити присутан релативно велики број страних речи. Још једна претпоставка од које се полази јесте да се у текстовима „југо-рока" могу лоцирати и именовани ентитети географских подручја, чије референцирање осликава идеју јачања духа *братства и јединства* између народа на простору бивше Југославије.

1.5 Методологија истраживања

Полазну тачку овог пројекта чини структурирана организација кодова за аутоматско креирање колекције текстова песама помоћу технике *гребања веба*. Скрипт за прикупљање корпусне грађе је писан у програмском језику Python 3,[5] у интерактивном програмском окружењу Anaconda [6]. За примарни извор преузимања текстова песама одабран је музички сајт LyricWiki [7].

Определили смо се да будући корпус чине текстови нумера на некадашњем српскохрватском језику, који је стандардизован Бечким књижевним договором 1850. године, да би се са распадом Југославије 1991. разложио на српски, хрватски и босански језик (Hentschel, 2003: 277). У складу са тим, претраживали смо текстове песама на поменутим трима језицима, док су текстови песама на македонском и словеначком језику изостављени због мањег степена сличности са тадашњим српскохрватским језиком.

У првој фази издвајања и прикупљања података користићемо Пајтон библиотеку `lyricsmaster` [8], као погодно средство за екстракцију већег броја текстова који ће чинити

[4] Детаљнији приказ феномена *југословенског рокенрола* описан је у поглављу 3.
[5] https://www.python.org.
[6] https://www.anaconda.com.
[7] http://lyrics.wikia.com/wiki/LyricWiki.
[8] https://pypi.org/project/lyricsmaster/.

будући корпус[9]. Два основна критеријума за одабир извођача чије текстове преузимамо били су следећи:

- Да су текстови аутора објављени за време постојања двеју југословенских држава – Социјалистичке Федеративне Републике Југославије (1945-1992) и Савезне Републике Југославије (1992-2003);
- Да су исти аутори били музички релевантни у то доба, при чему су нам као смернице у оцени популарности послужили истакнути радови о југословенској рокенрол сцени држава из бивше Југославије; додатни показатељ овог параметра била је и листа најпопуларнијих извођача са комерцијалног сајта Last.fm[10].

1.6 Садржај рада

Рад покрива поступак аутоматског прикупљања грађе са интернета, полуаутоматског препроцесирања и конкретне анализе у програмским језицима и различитим софтверима. У првом, уводном поглављу представљен је циљ и предмет рада, истраживачка питања, хипотезе и методологија истраживања. Друго поглавље је посвећено сродним истраживањима у области корпусне лингвистике за српски језик, и *ископавања из текста* на примеру текстова песама на страном језику. У трећем поглављу анализира се феномен *југословенског рокенрола.*. Четврто поглавље демонстрира аутоматско *гребање веба* у Пајтону, генерисање јединствене XML датотеке корпуса и њено анотирање, уз проширивање могућности прикупљања података. У петом поглављу описује се полуаутоматско пречишћавање грађе на формалном и садржинском нивоу. Шесто поглавље објашњава трансформацију XML датотеке помоћу XSLT процесора у XHTML формат. Употреба двају рачунарских алата ради аутоматске рестаурације дијакритика је тема седмог поглавља. У осмом поглављу представљају се библиотеке за проналажење интересантних лингвистичких образаца и препознавање именованих ентитета којима се износе неке карактеристике корпуса. Девето поглавље обрађује три алата за визуализацију корпусне грађе, док се десетим поглављем заокружује дискусија о рачунарској изради и анализи „ex-Yu” корпуса. Током излагања техничких детаља у вези са

[9] Алтернативно средство гребања веба јесте и модул PyLyrics https://pypi.org/project/PyLyrics/, мада његове функционалности боље одговарају захтевима за прикупљање релативно малог броја текстова песама за једног аутора, при чему називе песама самостално дефинишемо.

[10] https://www.last.fm/tag/ex-yu+rock/artists.

формирањем и анализом корпуса, текст ће на одређеним местима бити праћен и репрезентативним деловима кодова, ради бољег разумевања наведених поступака. За преглед комплетних кодова који су коришћени за прикупљање и обраду текстова, може се консултовати електронски Прилог на крају рада, као и GitHub репозиторијум аутора [11].

2 Сродна истраживања

Генерално, *корпусна лингвистика* (енгл. *corpus linguistics*) представља развијену научну методологију за руковање структурираним, машински читљивим и наменски бираним текстовима, који представљају основу за анализу неког језичког аспекта. Учесталост коришћења одређене речи или фразе, проналажење речи у контексту (енг. *key word in context concordance*), стилистичка анализа или добијање информација на основу метаподатака (пола особе, жанра, аутора текста итд) само су неке од могућих задатака ове научне дисциплине (McEnery & Hardie, 2012: 1).

У последње време, аутоматско преузимање текстова песама са веба, њихова електронска обрада и анализа све више добијају на значају. Захваљујући конкретним резултатима које ове технике могу произвести, као што су генерисање листе најучесталијих речи (Laurier, Grivolla, & Herrera, 2008), *моделовање тема* – енгл. *topic modeling*, (Zhang, Caro Repetto, & Serra, 2017; Lukic, 2015), екстракција структуре из текста (Mahedero, MartÍnez, Cano, Koppenberger, & Gouyon, 2005: 476), аутоматско индексирање песама у зависности од садржине текстова, (Logan, Kositsky, & Moreno, 2004), примена *text mining*-а и *NLP*-а у музичком домену бива препозната као перспективна научна пракса, о чему сведочи и велики број пројеката и радова на исте теме.

Примера ради, демонстрирано је коришћење Пајтон (енг. Python) програмске библиотеке `lyricsmaster` у циљу директног преузимања и чувања текстова песама из комплетне дискографије америчког репера Тупака Шакура [12], која је доступна на сајту LyricWiki. Затим, у другом истраживању су са веб-стране AZLyrics преузети текстови песама поп певачице Тејлор Свифт, како би се у њима пронашло и визуализовало двадесет најфреквентнијих речи (Chen, 2016). Истичемо и рад у вези са употребом техника

[11] https://github.com/ljpetkovic/ex-yu-songs-corpus.
[12] В. фусноту 8.

ископавања из текста ради праћења еволуције тематских мотива кроз лексичку анализу текстова песама кантаутора Леонарда Коена (Haslam, 2017). Коначно, постоји и корпус који обухвата текстове песама аутора на међународном нивоу, чије су музичке каријере остварене у различитим жанровима (The Open Lyrics Database, s.d).

Међу бројним пројектима српског Друштва JePTex у области израде корпуса и сродних лексичких ресурса на пољу корпусне и рачунарске лингвистике (Krstev & Vitas, 2005; Stanković, Obradović, & Trtovac, 2012; Krstev et al., 2011), посебну пажњу треба обратити и на развој веб-апликације Аурора[13] за генерисање конкорданце из књижевних дела. Омогућена је, између осталог, претрага конкорданци и проналажење речи у контексту, добијање података о фреквенцији токена и преглед пуног текста дела.

3 Појам југословенског рокенрола

Музички аналитичари, социолози и антрополози се једногласно слажу у оцени да је југословенски рокенрол оставио дубоког трага на просторима бивше Југославије. Списак одабраних текстова који се баве овом темом представља додатну потврду релевантности жанра који је под разматрањем [14]. Такође, развој југословенске рок сцене и креирање мултимедијалног документа инспирисаног наведеним жанром чини окосницу рада Обрадовић, Арсенијевић, & Шкорић, 2016. Због тога је у овом одељку фокус истраживања био усмерен ка сажетој анализи вишеслојног појма *југословенског рокенрола* кроз концептуалну, културолошку и феноменолошку визуру.

У питању је, дакле, музички стил који је зачет на просторима бивше Југославије двадесетих година прошлог века, са појавом америчког џеза (Kojić, s.d.). Пратећи хронологију развоја жанра коју наводи Раковић (2018: 746), сазнајемо да је „ex-Yu" рок педесетих година био обликован под утицајем других познатих америчких стилова музике блиских рокенролу, као што су *рокабили* и *твист*. Овај тренд ће наредне деценије бити замењен присвајањем британског начина извођења рокенрола, по узору на бенд The Beatles Раковић (исто: 747). Даљи утицај западњачког духа на југословенску верзију рокенрола се седамдесетих година огледао у виду јављања бендова насталих као омаж *хипи* покрету (Kojić, s.d.), али и италијанским *канцонама*, француским *шансонама*

[13] http://aurora.jerteh.rs.
[14] http://www.fil.bg.ac.rs/mmd_27/mmd_2015/knjige.php.

и немачким *шлагерима*, како се истиче у Petrov, 2017: 45; осамдесете године су, пак, донеле уплив британског *новог таласа* (Ristivojević, 2012: 214).

Поред афирмације модерног звука, прихватање западњачке музике је такође довело до радикалне промене тадашњег југословенског друштва, њиховог начина живота и стварања другачијих културних вредности. Наиме, Божиловић тврди да је развој југословенског рокенрола значајно утицао на прелаз Југословена из социјалистичког у капиталистичко друштво. Према истом аутору, овај музички феномен промовисао је и неговање пацифистичких и космополитских идеала на тим просторима, што се у датим околностима сматрало авангардном појавом Божиловић, 2016: 263).

Упркос томе што је музика пореклом са Запада била популарна међу младим Југословенима, она је противречила југословенском социјалистичком режиму, формираном 1945. године. Конкретно, џез жанр је наилазио на осуду власти све до педесетих година, када је Југославија почела слободније да прихвата утицаје западне културе (Божиловић, 2016: 265)[15]. Блиско томе, Раковић истиче да је рокенрол, као нови вид забаве, представљао предмет критике југословенске штампе до краја исте деценије (Раковић, 2018: 431). Естетски посматрано, половином шездесетих година је запажена промена у начину одевања, нарочито са појавом фармерки и мини-сукања; са друге стране, неговао се стил ношења дуге косе, премда су све наведене иновације изазивале негодовање чланова породице, школских директора и наставника (Раковић, 2011: 747; 753).

Међутим, оно по чему се југословенски рок посебно истицао били су слободарски текстови. Они су често били изложени критици, јер су представљали својеврстан отпор режиму, који је зазирао од прихватања утицаја популарне културе Запада. Наиме, будући да је садржај текстова неретко био политички ангажован, сатиричан или, пак, вулгаран, самим тим био је третиран као неподобан за тадашње југословенске друштвене прилике (Гајић, 2018: 23). Доказ да није реч о изолованој појави представља и систематизација случајева (ауто)цензуре у музичком свету југословенског рока (Kostić, 2005). У вези са претходним ставовима, парафразираћемо и одељак из Божиловићеве социолошке студије о култури сећања и југословенском рокенролу, којом се износе две

[15]Ипак, како наводи Раковић, неки љубитељи озбиљне музике су и даље веровали да је џез синоним за „дилетантску музику која залуђује омладину и одводи је од стицања префињеног музичког укуса" (исто: 430)

кључне одлике југословенског рока:

1. Када посматра југословенски рокенрол са идеолошког аспекта, аутор наглашава да тај жанр почива на бунту и борби рок извођача за демократију;

2. Социјални активизам је затим предочен у неконвенционалне, једноставне, непосредне или импровизоване музичке форме и текстове (Божиловић, 2016: 264).

Наредна поглавља нуде опис прикупљања и анализе корпусне грађе, у циљу покушаја да се методама корпусне лингвистике лоцирају неке од карактеристика овог жанра.

4 Рад у програмском језику Python

4.1 Коришћење Пајтона у сврхе *ископавања из текста*

Операције обраде се неретко примењују над скуповима неструктурираних података, којима је човек дигиталног доба данас засут. За квалитетније спровођење разнородних истраживачких пројеката који се тичу обраде и анализе велике количине података, информатичка заједница се често ослања на имплементацију програмског језика Пајтон. Пајтон је, између осталог, *језик високог нивоа*, што значи да је његова синтакса блиска стандардном енглеском језику, и самим тим пријемчивија за учење (Univerzitet u Beogradu, Matematički fakultet, s.d).

У овом истраживању искоришћене су могућности овог језика ради прикупљања жељеног садржаја са веба, тј. текстова песама. Иначе, овај језик опште примене се углавном користи ради проналажења информација, вршења статистичких прорачуна, машинског учења, обраде природних језика и данас све заступљеније научне методологије, познатије у англосаксонској литератури као *text mining* (Miller, 2015).

Глобално посматрано, можемо констатовати да се назив *text mining* усталио услед његовог доследног референцирања у бројним научним радовима. Насупрот томе, у славистичкој литератури није дошло до јасног консензуса по питању јединственог термина за означавање ове дисциплине, што потврђују и бројни синонимни. Термини који су у оптицају јесу *истраживање* (Pajić, 2012: 17), *ископавање из текста* (Vitas et al.,

2012), *ископавање текста* (Antić, 2011: 55), *интелигентна обрада текста* (Saračević, Mašović, & Kamberović, 2010: 1097), или *рударење текста* (Mirkov & Peranović, 2015: 586).

Посматрано са лингвистичког аспекта, занимљив увид у разноврсност именовања ове дисциплине нуде и преводи руских термина *интеллектуальный анализ текста* („интелектуална анализа текста"), француског *fouille de textes* („претраживање текста"), словеначког *rudarjenje besedil* („рударење текста") итд. који се генерално користе на интернет мрежи. Сви ови називи додатно истичу различита тумачења термина *text mining*, као надасве широке и популарне области напредне обраде података.

Свесни тешкоћа проналажења одговарајућег превода, у овом раду се као могући превод овог термина предлаже калковани израз *ископавање из текста*. Наиме, он је формални и садржински еквивалент називу *text mining*: формални зато што представља буквални превод термина, а садржински јер одражава сам циљ дисциплине, који се односи на проналажење или извлачење интересантних образаца из велике количине текста. Заправо, главни циљ наведене технике јесте „генерисање нових информација и претварање неструктурираног текста у структуриране податке, који ће бити коришћени у даљој анализи" (Linguamatics, para. 3).

Блиско томе, у опис концепта *ископавања из текста* уткане су и научноистраживачке тежње за препознавањем извесних текстуалних *патерна* (односно шаблона, трендова, образаца) о чему се опширно дискутује у књизи Miner, Elder, & Hill, 2012. У оквиру ове дисциплине, значењски опсег термина „текст" може обухватати садржај дигитализованих књига, порука електронске поште или, пак, коментара са друштвених мрежа (Irfan et al., 2015: 158; 164). Текстови песама, будући да су доступни на интернету, такође допуњавају поменуту листу веб-садржаја, и могу се на исти начин обрађивати ради спровођења анализе.

4.2 Прикупљање корпусне грађе – *гребање веба*

Када је реч о методи *гребања веба*, она се односи на употребу одређене компјутерске алатке за ефикасно преузимање и обраду дигитализованог садржаја[16]. На основу

[16]Ову методу не треба мешати са поменутим *web crawling*-ом, чије се средство функционисања, *web crawler*, односно „веб-индексер" (Petrović & Ivanović, 2011: 123), најопштије одређује као „систем за масовно преузимање веб страница" (Olston & Najork, 2010: 176). Иако по опису начина рада дели сличне

прегледа српске, хрватске и босанске литературе, стиче се утисак да не постоји јединствени терминолошки еквивалент термину *web scraping*. За означавање ове технике се најчешће употребљавају термини као што су *исецање* (Sweigart, 2015: 233), *налажење података на интернету* („Priručnik za data novinarstvo 1.0“, s.d, „Nalaženje podataka na internetu“) или *стругање података* („Priručnik za data novinarstvo 1.0“, s.d, „Scraping web sajtova – zašto?“, para. 9). У складу са називом методологије, садржај интернет сајтова који је прибављен на тај начин назива се „саструганим“ (Milić, 2015: 40). Поред тога, Јурић, Пехар и Заудер (Juric, Pehar, & Zauder, 2016: слајд 37) као сродне називе за поменуто веб стругање наводе црпљење (енг. *extracting*) и побирање (енгл. *harvesting*) података.

Суштина ове праксе јесте аутоматизација процеса свеобухватног екстраховања и складиштења података са веба. Штавише, описана метода може бити од изузетног значаја оним корисницима који треба да прикупе велику количину података у што краћем року ради реализације пословних или непословних активности[17]. Самим тим, примена гребања веба намеће се као далеко ефикасније решење у односу на методе ручног копирања и чувања сваке информационе јединице понаособ. Осим тога, наведена техника може представљати и једино могуће решење за екстракцију података у случају када су наведене опције copy/paste онемогућене у веб-читачу, или када сајт не обезбеђује API[18] или RSS feed-ове[19] (Panta, 2015: 13). Као што је већ наговештено, у овом истраживању биће демонстрирано функционисање гребања веба на примеру сајта LyricWiki.

особине са техником „стругања података”, кључна разлика лежи у томе што „веб пузање” индексира жељене стране са којих се могу преузимати подаци, док се „веб гребањем” издвајају циљани подаци са сајтова (Panta, 2015: 13).

[17] Одређеније, Шарма и Гупта истичу да коришћење „веб екстрактора за прикупљање садржаја” (енг. *web content extractor*) може помоћи новинарима у претраживању и одабиру релевантних садржаја за извештавање јавног мњења. Исти аутори додају да је ова алатка намењена и предузетницима који анализирају цене тржишних производа или податке о некретнинама. Осим тога, веб подаци могу бити екстраховани и ради информисања о дестинацијама за одмор, метаподацима књига итд. (Sharma & Gupta, 2012: 291).

[18] Акроним *API* (*Application Programming Interface*) означава софтвер који омогућава размену информација између различитих апликација. Више о API-ју в. “What is an API? (Application Programming Interface)”, s.d, para. 1.

[19] *RSS feed* (*Really Simple Sindication*) се користи за прикупљање релевантних информација, без директног приступања одређеном сајту (Cold, 2006).

4.2.1 Опис извора података: сајт LyricWiki

LyricWiki представља комерцијални музички веб-сајт, на коме су забележени текстови песама домаћих и страних аутора. Сајт је претражив према називу песама, језика на коме су оне написане, имену извођача, његовог родног места, продуцентске куће, албума и жанра коме извођач припада. Сам назив сајта сугерише да је он *wiki-based* (Woods, Thoeny, 2011: 76), што значи да корисници на њему могу прегледати и уређивати садржај, или дискутовати о истом, што је начелни принцип по коме функционише и општепозната онлајн енциклопедија Wikipedia. Како је наведено у опису[20] LyricWiki-ја, овај сајт је јавно доступан и има дозволу за објављивање поузданих и лиценцираних текстова песама.

База података сајта складишти више од 2,037,049[21] текстова песама домаћих и страних извођача. Будући да LyricWiki садржи релативно велики број песама и на некадашњем српскохрватском језику, у иницијалној фази истраживања смо преузимали текстове песама у извођењу одређених музичара из бивше Југославије који су обухваћени списком на веб-страни[22] посвећеној њима. Разлог више за одабир управо ове стране за гребање текстуалних података јесте чињеница да се са њега лако могу преузимати текстови песама захваљујући одговарајућем API-ју, тј. Пајтон модулу библиотеке `lyricsmaster`, о коме ће бити више речи у наредном пододељку.

4.2.2 Функционалности Пајтон библиотеке lyricsmaster

Аутоматско преузимање текстова песама са интернета представља ефикасну и релативно често коришћену методу која претходи електронској анализи корпуса[23]. Један од репрезентативних примера такве праксе јесте употреба програмске библиотеке `lyricsmaster`, која је доступна на PyPI[24] репозиторијуму програмских пакета за рад у

[20] LyricWiki, s.d, "Home", http://lyrics.wikia.com/wiki/LyricWiki.

[21] Статистички податак преузет 13.01.2019. са http://lyrics.wikia.com/wiki/Special:Statistics.

[22] http://lyrics.wikia.com/wiki/Category:Language/Serbian. Технички гледано, наслов стране (Category:Language/Serbian) није тачан, будући да се на истој страни не налазе искључиво текстови на српском језику, већ и они који су писани на другим југословенским и страним језицима.

[23] Ипак, етапа гребања веба може бити и изостављена уколико аналитичар текста већ располаже готовим скупом података. Примера ради, сајт Kaggle представља отворену базу података који припадају различитим доменима људске делатности, међу којима велики део заузимају и текстови песама. У оквиру наведене веб-стране налази се и корпус који садржи више од 380.000 текстова песама преузетих са сајта MetroLyrics https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics.

[24] https://pypi.org/project/lyricsmaster/.

Пајтону. Помоћу наведеног API-ја могу се прикупити текстови песама са неких од најпопуларних музичких сајтова (провајдера), као што су AZLyrics, Genius, Lyrics007, MusixMatch, The Original Hip-Hop (Rap) Lyrics Archive – OHHLA.com. Ради остваривања већег степена приватности и анонимизације IP адресе корисника који шаље захтев за прикупљање текстова, постоји опција гребања сајта преко Tor Proxy Server-а[25] помоћу TorController класе.

Имплементација функционалности модула библиотеке `lyricsmaster` резултује директним екстраховањем текстуалног садржаја са различитих циљаних страница у оквиру наведеног сајта, где је свака од тих страна[26] посвећена неком од популарних „ex-Yu" извођача из наше колекције. Наиме, дефинисаном функцијом за гребање веба проналажене су жељене странице и коришћене одговарајуће методе класе за прикупљање једино текстова песама, а не и других садржаја на истој страни (нпр. података у заглављу или бочној траци веб-стране). Имајући то у виду, определили смо се за конструисање алгоритма који би са тих страна преузимао садржаје текстова према именима извођача, и то на следећи начин:

1. Најпре је за провајдера текстуалних садржаја одабран сајт LyricWiki;

2. Затим је листом `izvodjaci` дефинисан скуп за 30 извођача чији текстови песама се прикупљају. Имена извођача референцирамо тачно онако како су наведена на сајту – ово се првенствено односи на случајеве у којима две групе деле исти назив (нпр. српска и финска група Negative). Наиме, уредници LyricWiki-ја су извршили дистинкцију између наведених група тако што су српском саставу придодали ознаку „RS", како би се знало да је реч о бенду из Републике Србије. Самим тим, у код за пребање веба је унет модификован назив Negative (RS);

3. Креирана је функција `korpus()` чији аргументи су елементи наведене листе извођача, и за сваког од њих смо потом покушали да преузмемо текстове помоћу функције `get_lyrics()`. Потом се формирају објекти `album` и `song`, док се увођењем параметара `title` и `lyrics` добијају наслови албума (`album.title`), песама (`song.title`) и сами текстови песама (`song.lyrics`).

[25] Tor, доступан на https://www.torproject.org.
[26] Пример стране за извођача Бајагу и Инструкторе http://lyrics.wikia.com/wiki/Bajaga_%26_Instruktor.

4. Након тога, преузети корпусни материјал би се чувао на локалној меморији рачунара методом `save()` [27];

∗ 5. Ипак, у току рада алгоритма, програм је пријављивао грешку услед које није успевао да приступи одређеној песми (а самим тим ни њеном тексту), тако што би вратио одговарајућу вредност функције за њен наслов. Штавише, иако је на LyricWiki странама са дискографијом одређених аутора (нпр. Горана Карана) већина наслова песама била форматирана у виду хипервеза које упућују на стране са доступним текстовима песама, за неколицину наслова уопште нису биле креиране наведене стране, упркос формалном постојању наслова (нпр. „Splitska serenada"). Другим речима, кад год би програм наишао на непостојећи наслов песме, обустављао би рад.

6. Ради руковања изузецима, увели смо исказ `try` за детекцију грешке на оном месту где је уочен проблем; њему смо придодали одредбе `except` и `continue` које отклањају грешке и омогућавају програму да настави са даљим радом.

4.2.3 Ограничења lyricsmaster-а

За одређене саставе и соло извођаче (Рибљу чорбу, Екатарину Велику, Азру, Филм, Џибонија, Ђорђа Балашевића и др.) није било могуће прикупити текстове помоћу API-ја `lyricsmaster`, било због онемогућеног приступа подацима или непостојања текстова песама за датог извођача.

Такође, у фази прикупљања текстова дошло до закључка да би постојећи корпус требало да буде допуњен и текстовима песама женских извођача. Па ипак, за одређене ауторе чији се музички стил може више окарактерисати као „рок" (Калиопи, Слађана Милошевић и сл.) није било могуће аутоматски преузети текстове са сајта LyricWiki; са друге стране, неки аутори се уопште нису налазили у бази података, а самим тим ни њихови текстови (као што је то случај са Мајом Оџаклијевском).

Како бисмо надоместили овај недостатак, применили смо алтернативне критеријуме за накнадни одабир извођача тако што смо проширили жанровски опсег који ће бити покривен нашом грађом. Прецизније, укључили смо у корпус и поп ауторе (нпр. Нину

[27] Подразумевана апсолутна путања јесте /{user}/Documents/lyricsmaster/.

Бадрић и Зану) у чијим музичким аранжманима се запажају и утицаји рока.

Поред тога, додали смо и Мадам Пиано, коју је стручњак за „ex-Yu" рок, Петар Јањатовић, уврстио у своју енциклопедију југословенског рока (Janjatović, 2007: 11; 214), због значаја који је ова представница џез и етно-звука имала за развој југословенске музике. Што се тиче мушких извођача, за Горана Карана се исто може тврдити да није репрезентативан пример рок извођача, али да не одступа превише од задатог оквира[28]. Следи тренутни списак од тридесеторо аутора чији текстови су обухваћени корпусом, уз могућност допуне:

Бајага	Електрични Оргазам	Неверне Бебе
Бајага и Инструктори	Забрањено Пушење	Негатив
Беби Дол	Зана	Нина Бадрић
Бијело Дугме	Идоли	Октобар 1864
Ван Гог	Индекси	Партибрејкерс
Галија	Јосипа Лисац	Прљаво Казалиште
Горан Бреговић	YU Група	Рани Мраз
Горан Каран	Кербер	Смак
Дино Мерлин	Мадам Пиано	Хари Мата Хари
Дивље Јагоде	Мирзино Јато	Хаустор

4.2.4 Формирање стабла директоријума за смештање грађе

Резултат извршавања претходног кода представљало је формирање хијерархијски организованих структура података, односно стабла директоријума са својим кореном и гранама. Тако добијену дрволику организацију можемо представити у виду извода са командне линије неког оперативног система, након покретања кода:

```
alias tree="find . -print | sed -e 's;[^/]*/;|____;g;s;____|; |;g'"
```

(Doyle, 2018).

У наставку текста је дат парцијални приказ наведене структуре који је преузет са апликације Terminal на оперативном систему Mac OS X:

[28]Наиме, иако је Каран стекао популарност извођењем поп песама „са далматинским призвуком", занимљив је податак да је он започео каријеру управо као рок певач ("Goran Karan u Beogradu", 2007, para. 4, http://mondo.rs/a48423/Zabava/Muzika/Goran-Karan-u-Beogradu-14.-i-15.-februara.html.)

Слика 1: Стабло директоријума „LyricsMaster“ са командне линије Terminal.

Као што се може закључити на основу горњег приказа, коренски директоријум носи подразумевани назив `LyricsMaster`, унутар којег су лоцирани поддиректоријуми са називима извођачâ (нпр. `Mirzino-Jato` или `YU-Grupa`); Затим се за сваког извођача наводе називи албума (`Šećer i med`, `YU-Grupa`, `Rim 1994`, `Ima Nade`, итд), у којима су садржани текстови песама. Крајњи чворови у структури дрвета представљају саме датотеке које садрже текстове песама у .txt формату (нпр. `Apsolutno-Tvoj.txt`, `Crni-Leptir.txt`). Приметићемо да се међу наведеним подацима формално јавља и датотека `.DS_Store`[29], која није ни од каквог значаја за даљу обраду и анализу корпуса, стога ће она у каснијим фазама бити елиминисана. Након процедуре гребања сајта, прикупљени садржаји текстова у .txt формату, који су ускладиштени у засебним директоријумима, у наредној етапи припреме корпуса биће обједињени у јединствену XML датотеку и аутоматски анотирани.

[29]Скр. од *Desktop Services Store*. Наведени фајл садржи информације о начину приказивања фолдера (било у виду иконица, листе или колоне), изгледу иконица и другим подешавањима фолдера на оперативном систему Mac OS X. Будући да назив фајла почиње симболом „.”, он је невидљив приликом класичне навигације кроз фолдере које га садрже, али се појављује приликом листања директоријума и фајлова (.DS Store, s.d), као у горњем примеру.

4.3 Генерисање корпуса као XML документа

4.3.1 О XML-у

XML представља *проширив језик означавања* (Krstev, s.d), што је скраћеница од *eXtensible Markup Language*. Сам назив језика говори да није реч о програмском језику као што су Python, Java, C++ и др, већ о врсти језика који има за циљ формално представљање структурираних података. Да би се то постигло, садржај неког документа је описан одговарајућим метаподацима или *ознакама* (*таговима*, енгл. *tags*). Поменути дескриптори могу бити *елементи* и/или *атрибути* који формирају дрволику структуру XML документа. Први елемент (*коренски елемент*, енгл. *root element*) нема родитеља, али садржи остале елементе, истовремено представљајући *родитељски елемент* (енгл. *parent node*) наредном *елементу потомка* (енгл. *child node*).

Једна од предности употребе овог језика јесте висок степен читљивости (W3schools.com, s.d), што га чини релативно лаким за коришћење. Флексибилност употребе овог језика огледа се у томе што се тагови произвољно дефинишу (што није случај са нпр. HTML-ом и CSS-ом, у којима су они стандардизовани), као и да се у документ може унети неограничен број тагова. Да би XML документ био исправан, он мора бити *добро формиран*, тј. треба да поштује принципе XML синтаксе; ипак, уколико желимо исти документ да представимо у неком другом формату, он онда мора бити и *валидан* по узору на *декларацију типа документа* (енгл. *Document Type Declaration* – скр. *DTD*)[30]. У следећим пододељцима представићемо поступак генерисања XML датотеке од текстуалних материјала прикупљених гребањем музичких страна.

4.3.2 Проналажење свих текстова коришћењем модула os

Први корак представљала је примена модула `os`[31] са методом `listdir()`, из стандардне програмске библиотеке Пајтона [32], са циљем правилног распоређивања садржаја у будућој анотираној датотеци, тј. прегледа текстова песама са свих албума за датог аутора. Наведена Пајтон метода враћа листу назива свих датотека и поддиректоријума у одређеном директоријуму. Да би се приступило свим називима, било је

[30] Више речи о исправној форми и валидности XML документа у пододељцима 5.1. и 6.2.
[31] https://docs.python.org/3/library/os.html.
[32] https://docs.python.org/3/library/.

потребно најпре формирати три апсолутне путање, почев од кореног директоријума `LyricsMaster`: две које садрже називе поддиректоријума (име аутора и назив албума) и крајњу, са насловом песама. Затим су употребљене функције `join()` и `isfile()` из `os.path`[33] модула. Прва функција је намењена спајању назива из генерисане листе директоријума и датотека (аутора, албума и песама) са крајевима текућих путања директоријума, како би им се приступило. Другу функцију смо комбиновали са исказом контроле тока `if (not)` како би се исправно утврдило да ли се генерише листа назива директоријума или фајлова.

4.3.3 Анотирање корпуса помоћу библиотеке Yattag

`Yattag` [34] представља Пајтон библиотеку којом се аутоматски могу додавати HTML и XML ознаке приликом структурирања докумената. Овај API аутоматски додаје отворене и затворене изломљене заграде, и сваки почетни таг је праћен затвореним тагом. На овај начин, анотатор може брже и лакше обележавати текстове, будући да програм смањује могућност јављања синтакстичких грешака. `Yattag` кроз модул `indent` такође подржава и аутоматску индентацију према општој хијерархијској структури XML фајла, при чему се величина индентација може модификовати. Будући да је поступак анотације требало спровести на великом корпусу, идеја је била искористити функционалности модула наведене библиотеке у склопу итеративног анотирања у складу са дефинисаним правилима означавања садржаја текстова. Та правила подразумевају укључивање XML ознака елемената и атрибута за описивање одговарајућег дела садржаја корпуса:

- Корени елемент је назив нашег корпуса, обележеног тагом `exYuPesme`;
- Аутори, односно извођачи, описани су тагом елемента `<autor>`, чији су атрибути наведени као `ime`, `brojAlbuma` и `pol` (текстописци се не бележе, а не наводе се ни на LyricWiki страницама песама);
- Албуме дефинишемо ознаком елемента `<album>`, који садржи атрибуте `naziv` и `godina`;

[33] https://docs.python.org/3/library/os.path.html.
[34] https://github.com/leforestier/yattag.

- Песма садржи таг елемента `<pesma>` и атрибут `naslovPesme`, док је за стихове резервисана ознака елемента `<li>`.

Пошто метод `tag()` креира XML ознаке, њему прослеђујемо као аргументе само жељене називе етикете у виду елемената и/или њихових атрибута.

```
with tag('autor', ime=author, brojAlbuma=len(albums), pol=""):
```

Са друге стране, метод `text()` генерише текстуални садржај који није назив етикете или атрибута. У циљу редукције кода, користили смо инстанцу класе `Doc` и здружену методу `tagtext()`, која тој инстанци додаје садржај који наведена метода производи (назив етикете и обичног текста). Након дефинисања свих неопходних параметара, применили смо `getvalue()` метод, како бисмо целокупан садржај претворили у велику ниску карактера. Овим путем текстови су смештени у нову XML датотеку.

4.4 Beautiful Soup

У потрази за одговарајућим алатом помоћу којег би могли бити прикупљени текстови песама за недостајуће ауторе, У процесу гребања интернет страница и, генерално, у области рударења текста, све је чешћа употреба Пајтон библиотеке `Beautiful Soup` (верзија 4)[35]. Помоћу ње је могуће ефикасно рашчланити и издвојити податке из текстова анотираних HTML и XML таговима. Имплементација модула ове библиотеке је нарочито погодна у случајевима када је потребно екстраховати само одређене садржаје са веб-страница, као што је то случај са текстовима песама. Друга важна ставка јесте поменуто избегавање ручног копирања или, пак, проналажење начина да се „саструже" текстуални садржај и у случајевима када веб прегледач не дозвољава ручно копирање. Суштински, `Beautiful Soup` приступа одређеном сајту, пролази кроз дрволику структуру документа, лоцира тагове између којих се налази садржај који желимо да прикупимо и преузима дати садржај. Тако ће текстови песама извођача који недостају нашем корпусу моћи да се преузму са неке од страна веб сајта MetroLyrics (нпр. в. страну за групу Рибља Чорба http://www.metrolyrics.com/riblja-corba-lyrics.html.. Програм ће прво приступити насловима песама, описаним тагом `<title>`, а затим и самим

[35] https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

текстовима нумера, смештеним између тагова <verse>. Прикупљени текстови би се чували на локалној меморији рачунара и били спремни за даљу обраду.

Овако би изгледао пример за прикупљање текстова песама за групе Азра, Рибља Чорба и Екатарина Велика:

Алгоритам 1: Издвајање текстова песама недостајућих аутора са сајта MetroLyrics.

```
from pyquery import PyQuery as pq
from lxml import etree
import requests
from bs4 import BeautifulSoup

with open('metrolyrics.txt', 'w') as m:
  linkovi = ['http://www.metrolyrics.com/azra-lyrics.html',
        'http://www.metrolyrics.com/riblja-corba-lyrics.html',
        'http://www.metrolyrics.com/ekv-lyrics.html']
  for link in linkovi:
    response = requests.get(link)
    doc = pq(response.content)
    titles = doc('.title')
    for title in titles:
      response_title = requests.get(title.attrib['href'])
      doc2 = pq(response_title.content)
      verse = doc2('.verse')
      print(verse.text())
      m.write((' ').join(verse.text().split()))
```

5 Полуаутоматско препроцесирање корпуса

5.1 Исправљање формалне структуре текстова

Користећи oXygen XML уређивач текста[36], уочен је недостатак адекватне XML структуре по питању присуства специјалних карактера у корпусу. Текстуални садржај је потом додатно обрађен како би у потпуности испоштовао принципе XML синтаксе. Наиме, у претходни код за генерисање XML корпуса је уметнута функција `html.escape()` модула `html`[37], чиме је карактер амперсенд (`&`, који нема своју излазну секвенцу у XML-у, и да се не би третирао као почетак *референце ентитета* (Means, 2006: 20) замењен својим карактером излазне секвенце `&`. По истом принципу је и симбол за апостроф (`'`) замењен својим HTML еквивалентом `&#x27`. Следећи код демонстрира генерисање *исправно формираног* (енгл. *well-formed*) XML документа:

[36] https://www.oxygenxml.com.
[37] https://docs.python.org/3/library/html.html.

Део резултата аутоматске анотације корпуса у XML формату представљен је у наставку текста. Наведен је пример песме „Апсолутно твој” групе Мирзино Јато:

```
exYuPesme   autor   album   pesma   li
1  <exYuPesme>
2      <autor ime="Mirzino Jato" brojAlbuma="1" pol="">
3          <album naziv="Sećer i med" godina="">
4              <pesma naslovPesme="Apsolutno Tvoj">
5                  <li>Apsolutno tvoj, apsolutno moj</li>
6                  <li>samo mi smo taj genijalan spoj</li>
7                  <li>apsolutno tvoj zivot je moj</li>
8                  <li>permanentno moj, permanentno tvoj</li>
9                  <li>samo ti si taj specijalan broj</li>
10                 <li>permanentno tvoj zivot je moj</li>
11                 <li>Ref.</li>
12                 <li>Apsolutno moj si broj</li>
13                 <li>apsolutno ja sam tvoj</li>
14                 <li>genijalan mi smo spoj</li>
15                 <li>permanentno bicu tvoj</li>
16                 <li>Apsolutno moj si broj</li>
17                 <li>apsolutno ja sam tvoj</li>
18                 <li>specijalan mi smo spoj</li>
19                 <li>permanentno bicu tvoj</li>
20                 <li>Ref.</li>
21                 <li>Apsolutno moj si broj, broj</li>
22                 <li>apsolutno ja sam tvoj, ti si moj</li>
23                 <li>specijalan mi smo spoj, spoj</li>
24                 <li>permanentno bicu tvoj</li>
25                 <li></li>
26             </pesma>
```

Слика 2: Одломак из правилно формираног XML корпуса.

Ради веће прегледности, аутоматски су уклоњене цртице између назива песама и албума које су остале приликом веб гребања. Ипак, у даљој обради корпуса акценат ће бити на аутоматској рестаурацији дијакритика, свођењу текстова написаних и латиницом и ћирилицом само на латинично писмо, исправљању штампарских, правописних и других грешака, као и на формирању листе српских *стоп-речи*.

5.2 Елиминација сувишног садржаја

Поновним прегледом корпуса запажен је и известан број текстова песама написаних искључиво на страним језицима. Будући да је тема овог рада обрада корпуса на југословенским језицима, било је извршено ручно уклањање наведених нумера или, пак, читавих албума из директоријума корпуса. Конкретно, из колекције текстова су избачене песме на грчком, македонском, ромском, енглеском, португалском и пољском језику. Иако је немали број аутора објавио барем једну песму на страном језику које

смо морали да елиминишемо, карактеристичан случај представља присуство мултилингвалности у дискографији Горана Бреговића као соло извођача, који је и аутор песама ”Kerna mas”, са албума *Alkohol: šljivovica & champagne*, ”7/8 & 11/8”, ”Ederlezi”, ”TV screen”, ”Ausência” (*Ederlezi*) и ”To nie ptak” (*Kayah & Bregović*). Албум са музиком за филм *Arizona Dream* садржи све песме на енглеском језику, тако да је и он уклоњен.

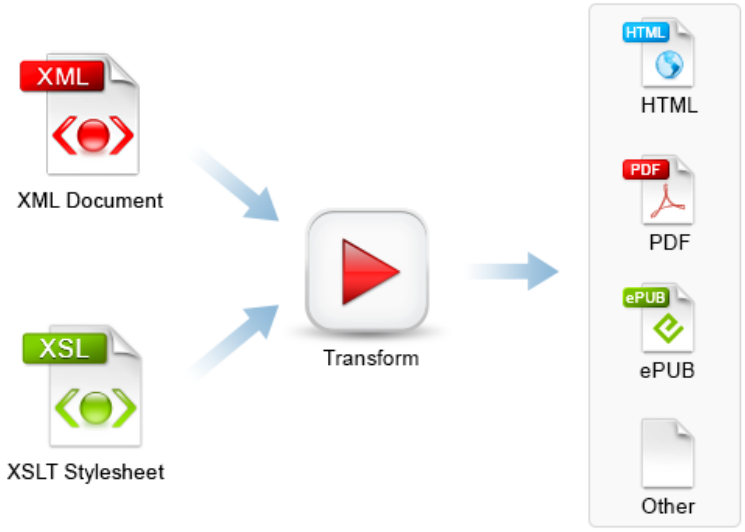
Упоредо са тим, водили смо рачуна и о томе да из корпуса искључимо текстове песама који су грешком приписани неком аутору. Примера ради, међу текстовима нумера на LyricWiki страни Нине Бадрић се налази и песма ”Ubila si del mene”, која припада словеначком *boy* бенду Game Over. Погрешно додељивање песме аутору такође је приметно за крагујевачку групу Смак, међу чијим песмама су биле присутне две песме финске истоимене групе (”Muistoja”, ”Myrsky”). Исто тако, на албуму *Zašto ne volim sneg* поменутог састава појављује се и неколико инструменталних песама. Нису у обзир узете ни песме без садржаја (тј. оне где су једино биле исписане три тачке), попут нумере „Ne mogu da kapiram”, бенда Партибрејкерс. Такође се дешавало да текст једне песме носи више различитих наслова (нпр. „Manijak”, „Vuk”, „Ujka Sam” за текст песме „Počasna salva” групе Забрањено Пушење).

Нарочиту пажњу је требало обратити на одабир песама са концертних или компилацијских албума. Наиме, првобитно је замишљено да се такве колекције одмах одстрањују, будући да су у њима већ постојале студијске верзије истих песама. Ипак, уочили смо присуство одређеног броја необјављених нумера на концертним албумима (нпр. „Na vrhovima prstiju” са албума *Neka svemir čuje nemir*); са друге стране, са албума највећих хитова Нине Бадрић (*Collection*) задржали смо мали број јединствених песама. Сличан начин селекције вршен је и за остале извођаче на чијим албумима се налазе репродукције старих песама. Другу врсту дупликата представљале су присутне и неке обраде песама, као за песму „Tako ti je mala moja kad ljubi Bosanac” (аутор: Бијело Дугме, 1975), коју је обрадила група Забрањено Пушење 1998. године.

6 XSLT трансформација XML датотеке у XHTML

6.1 Основе XSLT-а

XSLT (енгл. *eXtensible Stylesheet Language Transformation* је назив за апликацију која се користи у трансформацијама XML документа у неки други формат. Она се заснива на читању улазног XML и XSL документа. Друга врста фајла је писана *језиком проширивих листова* (енгл. *eXtensible Stylesheet Language*) који је намењен стилизовању улазног XML документа[38]. На основу *шаблона* (енгл. *templates*) присутних у XSL стилском листу, XSLT процесор врши поређење и рашчлањивање елемената XML документа, а излаз XSL конверзије јесте *стабло резултата* (Means, 2006: 15), или *излазно дрво*, (Krstev, s.d) (енгл. *result tree*). Добијени резултат потом бива претворен у нови формат документа (XML, HTML, чист текст итд). Наредна слика илуструје изложени концепт XSL трансформације:



[39].

Слика 3: Претварање XML и XSL датотека у жељени излазни формат

У случајевима када треба генерисати конкретне податке из улазног XML документа, користи се XPath упитни језик (XML Path Language) (Graovac, s.d, 7; 27). Прецизније, XPath изразима се проналазе одређени делови, тј. чворови (енгл. *nodes*) документа. Чворовима се означавају следеће компоненте у оквиру XPath модела података (Means, 2006: 164), који се могу рашчланити (*парсирати*, енгл. *parse*):

[38] Аналогни пример XSL-у би била употреба језика CSS за уређивање дизајна HTML страница.

1. Корен: `<exYuPesme>`;
2. Елементи: `<autor>`, `<album>`, `<pesma>`, `<li>`;
3. Атрибути: `ime`, `brojAlbuma`, `pol`, `naziv`, `godina` и `naslovPesme`;
4. Текст: нпр. `Sve će to, o mila moja Prekriti ruzmarin, snjegovi i šaš`;
5. Простор имена [40]: `xmlns:xsl="http://www.w3.org/1999/XSL/Transform"`, који се бележи како би се избегле евентуалне двосмислености у случају постојања истоимених назива елемената; помоћу простора имена се приступа XSLT компонентама;
6. `<?xml-stylesheet type="text/xsl" href="tabelecss.xsl"?>` је инструкција за обраду, која се придодаје постојећем XML документу да би био упарен са својим XSL еквивалентом (Fitzgerald, 2004) (`tabelecss.xsl`);
7. Коментар: опционо, текст коментара се умеће између симбола `<!--` и `-->`.

У циљу обогаћивања садржаја корпуса, ручно смо додали и вредности за атрибут `pol` извођача, с тим што смо у случају састава уносили вредност `Grupa`. Паралелно са тим, примећено је да програм за гребање LyricWiki страница није могао да региструје и годину издавања албума која се налазила иза наслова тих албума. Пошто је један од циљева будућег рада и испитивање еволуције мотива кроз године, определили смо се да за почетак додамо године издавања албума Бијелог Дугмета, групе са богатом колекцијом песама.

6.2 Валидација XML датотеке у складу са успостављеним DTD-ем

У претходном делу рада објаснили смо поступак правилног форматирања датотеке који испуњава критеријуме XML синтаксе. Други услов за креирање потпуно функционалног XML документа јесте поменута *валидност*. Конкретније, да би XML документ био валидан, он мора садржати *декларацију типа документа* којом се дефинише корени елемент документа. Она је истакнута у *прологу*[41] за обраду текста документа, тј. „између XML декларације и почетне етикете кореног елемента” (Krstev, s.d). Друга

[40] Енгл. *namespace*.
[41] Слично концепту *преамбуле* који постоји у LaTeXпрограму https://www.latex-project.org.

компонента DTD-а може бити URL преко које се декларација реферише, или, у нашем случају, *интерни* (*унутрашњи*) *подскуп DTD-а* (исто, s.d).

Њиме дефинишемо елементе и атрибуте XML корпуса. У декларацији се спецификује и да документ садржи више елемената (аутора, албума, песама и стихова). Једино се стихови обележавају резервисаном речју `#PCDATA`, која означава да елемент може садржати само рашчлањене знаковне податке (енгл. *Parsed Character Data*), али да не сме садржати елементе потомке; са друге стране, `CDATA` (енгл. *Character Data*) служи за описивање атрибута, чије вредности могу бити произвољне ниске текста. Наведене DTD ознаке такође указују на то да евентуални специјални знаци (', ", <, >, &) унутар елемената и атрибута морају бити замењени својим референцама ентитета. Следи опис DTD-а за цео корпус:

Алгоритам 2: Декларација типа документа за ex-Yu корпус.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE exYuPesme
  [
<!ELEMENT exYuPesme (autor)+>
<!ATTLIST exYuPesme
xmlns CDATA #FIXED ''>

<!ELEMENT autor (album)+>
<!ATTLIST autor
xmlns CDATA #FIXED ''
brojAlbuma CDATA #REQUIRED
ime CDATA #REQUIRED
pol CDATA #REQUIRED>

<!ELEMENT album (pesma)+>
<!ATTLIST album
xmlns CDATA #FIXED ''
godina CDATA #REQUIRED
naziv CDATA #REQUIRED>

<!ELEMENT pesma (li)+>
<!ATTLIST pesma
xmlns CDATA #FIXED ''
naslovPesme CDATA #REQUIRED>

<!ELEMENT li (#PCDATA)>
<!ATTLIST li
xmlns CDATA #FIXED ''>
  ]>
```

6.3 Приказ основних података о корпусу у XHTML формату

Пре него што изложимо процедуру представљања корпуса на вебу, кратко ћемо се осврнути на разлику између термина XHTML и HTML. Наиме, у Means, 2006: 109 је дато објашњење у вези са тим, где се наводи да је XHTML (енгл. *eXtensible HyperText Markup Language*) један од начина представљања XML документа на вебу, који користи језик прилагођен XML синтакси. Другим речима, XHTML представља „реформулацију HTML-а у XML синтакси" (Stachowiak, 2006), што такође подразумева и да XHTML *анализатор* (*парсер*, енгл. *parser*) намеће већа ограничења приликом рашчлањивања XML документа у односу на HTML парсер [42]. Ипак, разлог због којег је корисно обрађивати XHTML документа јесте у томе да се она онда лако могу поново претворити у XML формат, и тако анализирати у програмима за рад са XML документима; насупрот томе, структура HTML документа би захтевала додатне измене како би се постигла ваљаност и валидност XHTML структуре.

Међутим, од главног значаја за обраду корпуса о коме је реч јесте повезивање XML и XSL репрезентације садржаја текстова, како би се добио XHTML табеларни приказ релевантних података о:

- Именима и броју аутора, као и броју њихових песама и албума;
- Броју песама и албума за сваког аутора;
- Родној заступљености у корпусу.

XSL датотека садржи стандардне HTML тагове за опис табеле и документа у који ће табела са генерисаним подацима бити смештена. CSS синтаксом задали селекторе класе за стилизацију својствава који дефинишу изглед табеле и документа. У наставку је демонстрирана форматирање табеле овом методом:

```
table {
margin: 10px -1 330px;
border: 1;
background-color: cornflowerblue
}
```

[42] У погледу синтаксе, XHTML захтева да нпр. сваки почетни таг има и своју крајњу ознаку, што се у HTML-у толерише.

Ради реализације наведених упита, конструисали смо шаблонска правила која ће бити примењена на читав XML корпус. Да бисмо спецификовали елементе који ће бити исписани на излазном дрвету, односно на веб страни, користили смо неке од бројних XSLT елемената и функција за референцирање делова XML документа, према наредним принципима:

1. Будући да у корпусу већина аутора има више од једне песме, ауторе смо најпре груписали према њиховим именима. XSLT анализатор затим претражује све потомке контекстног чвора (енгл. *context node*, у нашем случају, текући чвор је коренски елемент) и сам контекстни чвор, од чијих јединствених ставки имена (вредности атрибута) креира листу аутора;

2. Од новог контекстног чвора (аутора из датог скупа) иде се ка његовим потомцима (песмама), чији се атрибути (наслови песама) броје. По сличном начелу врши се и пребројавање албума за сваког аутора;

3. Слично кораку бр. 1, бирају се и броје сви чворови чији атрибути имају вредност пола или групе.

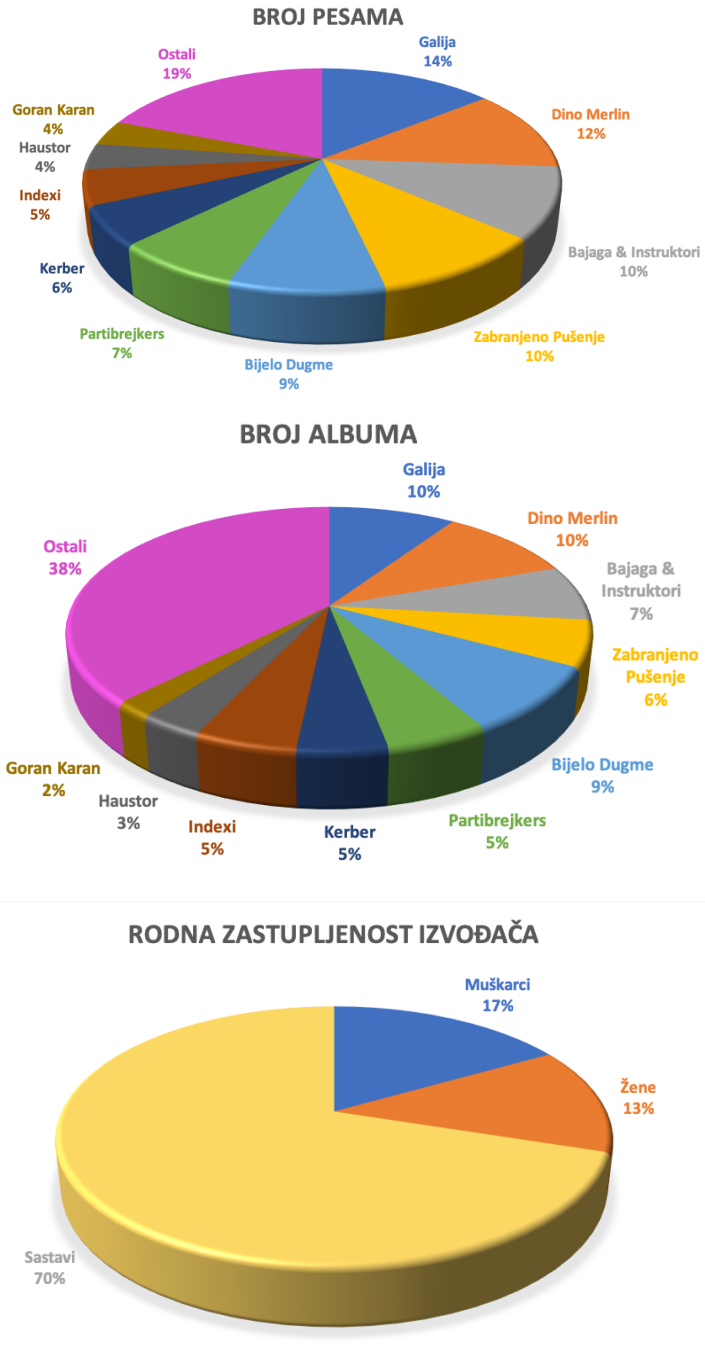
Код за XSLT трансформацију може се тестирати и на веб апликацији XSLT fiddle[43]. У уређиваче текстова је потребно унети XML и XSL улазне податке, како би се у излазним прозорима изворног (енгл. *native*), *Frameless* и *Saxon* процесора очитао резултат претварања. Резултат трансформације јесте табела представљена на слици 4.

[43] http://fiddle.frameless.io.

Statistika

Muskarci	Zene	Sastavi
5	4	21

Ukupno: 30	*Ukupno: 955*	*Ukupno: 148*
Autori	**Pesme**	**Albumi**
Galija	135	14
Dino Merlin	114	15
Bajaga Instruktori	98	11
Zabranjeno Pusenje	96	9
Bijelo Dugme	85	13
Partibrejkers	70	8
Kerber	54	7
Indexi	49	8
Haustor	36	5
Goran Karan	35	3
Prljavo Kazaliste	25	6
Negative	22	2
Hari Mata Hari	22	7
Nina Badric	21	5
Smak	21	6
Idoli	15	1
YU Grupa	14	7
Rani Mraz	13	2
Divlje Jagode	6	4
Goran Bregovic	5	3
Madame Piano	5	1
Van Gogh	2	2
Josipa Lisac	2	2
Bajaga	2	1
Zana	2	2
Oktobar 1864	2	1
Mirzino Jato	1	1
Neverne Bebe	1	1
Elektricni Orgazam	1	1
Bebi Dol	1	1



Слика 4: Статистички подаци о ex-Yu корпусу.

Као што се може приметити, табела је сортирана према броју песама за одређеног аутора, у опадајућем поретку. На основу ње закључујемо да група Галија има највише песама, а Дино Мерлин највише албума у корпусу; исто тако, постоји и известан број аутора са мањим бројем песама и албума забележених корпусом у односу на њихову

стварну музичку продукцију (нпр. српска група Ван Гог[44] или хрватска певачица Јосипа Лисац [45]. Стога, ова табела може представљати референтну тачку за даљи рад са корпусом у погледу процене заступљености аутора у њему. Ради јаснијег прегледа стања у корпусу, са десне стране је приложен и визуализовани удео заступљености аутора у корпусу према броју албума и песама, као и према полу.

Дађемо објашњење и поводом тога у којој мери су у корпусу присутни мушки, односно женски извођачи. Графички приказ расподеле мушких и женских соло-извођача и групних извођача указује на чињеницу да више од половине корпуса чине групни извођачи; ипак, за разлику од петнаест група састављених искључиво од мушких извођача (нпр. Бијело Дугме, Смак итд), постоје састави које чине и женски извођачи или на чијем челу су женски извођачи, које смо укључили у корпус (Негатив, Неверне Бебе, Зана, Октобар 1864, Мирзино Јато, Рани Мраз). На тај начин, покушали смо да постигнемо релативну балансираност корпуса по питању родне заступљености.

7 Аутоматска рестаурација дијакритика

Као што је већ напоменуто, прикупљени текстови песама у нашем корпусу били су прилично неуједначени по својој форми: већина текстова била је записана на латиници, а није мали број оних у којима специјална латинична слова (č, ć, ž, đ, š) не садрже потребне дијакритичке знаке. Са друге стране, мањи број песама је био првобитно забележен на ћирилици. Почетна замисао била је да се целокупан корпус транслитерира на ћирилицу; од те идеје се одустало из разлога што се у текстовима појављују неке речи и реченице на страним језицима (нпр. “la musique c’est fantastiqu prepare la revolution et la femme est tres jolie tre jolie comme un bonbon [sic]“ – из песме „Француска љубавна револуција” Бајаге & Инструктора).

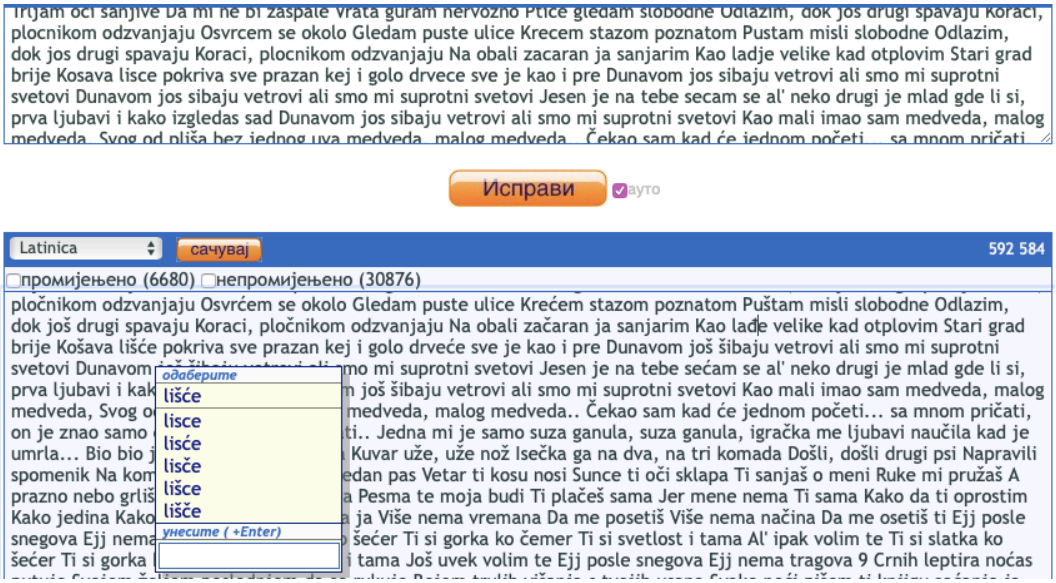
Стога смо се определили да у поступку транслитерације текстови написани *ошишаном латиницом*[46] и ћирилицом буду претворени у стандардну српску латиницу са дијакритичким знацима, и то аутоматски, услед обимности корпуса. У раду Krstev, Stan-

[44] http://www.musicvangogh.com/diskografija/.

[45] http://www.josipalisac.com/glazba/)

[46] Пре појаве Unicode стандарда за доследно енкодирање карактера, за писање на српском језику раније се прибегавало српској *ошишаној* или *ћелавој латиници*, која проистиче из ASCII кода Krstev, Stankovic, & Vitas, 2018: 1

kovic, & Vitas, 2018, који се бави истом темом, дат је линк ка софтверском алату „Слово Мајстор”, који аутоматски конвертује карактере одређеног српског писма на друго српско писмо [47]. Помоћу овог софтвера смо за изузетно кратко време генерисали жељени излаз на латиници. У великој већини случајева, алгоритам је ваљано исправио речи написане деградираном латиницом. Да би транслитерација у потпуности била успешна, може се применити метод ручног уноса додатних измене у трансформисани текст. Апликација нуди неколико могућих опција за исправљање погрешно транслитерираних речи – нпр. „lisce” може бити преправљено у неке од следећих јединица: „lisće”, „lišće”, „lišče”, „lišče” или „lišce”. Ефикаснија варијанта би била користити постојеће морфолошке речнике који би трансформисали речи према стандарду српског језика.



Слика 5: Транслитерација корпуса на латиницу – „Слово мајстор”.

7.1 Евалуација ефикасности рестаурације дијакритика

У овом одељку упоредићемо оригинални садржај корпуса, неуједначен по присуству дијакритичких знакова, са садржајем корпуса над којим је вршена рестаурација дијакритика у апликацији „Слово Мајстор”, као и у апликацији LeXimir помоћу електронских морфолошких речника, локалних граматика и Корпуса савременог српског језика (Krstev, Stankovic, & Vitas, 2018). Предност коришћења електронских морфолошких речника јесте могућност сагледавања врста речи, њихових фреквенција, лематизованих облика,

[47] Детаљније о пројекту рестаурације дијакритика из корпуса на јужнословенским језицима, в. Ljubešić, Erjavec, & Fišer, 2016.

метрике кључности, синтаксичко-семантичких ознака, апсолутних и релативних фреквенција који су дати у табеларном приказу. Прво су извршени су општи статистички прорачуни за „ex-Yu” корпус, на основу којих смо дошли до података о броју токена и јединственим лексичким ставкама за наведена три корпуса:

- Оригинални: 248807 токена, 17723 јединствених лексичких ставки, 116972 просте форме (17668 различитих), 268 цифара (10 различитих);
- Корпус са рестауриранim дијакритицима у апликацији „Слово Мајстор”, где су комбинована велика и мала слова: 248807 токена, 16953 јединствених лексичких ставки, 116972 просте форме (16898 различитих), 268 цифара (10 различитих);
- Корпус са рестауриранim дијакритицима у апликацији LeXimir: 248807 токена, 16964 јединствених лексичких ставки, 116972 простих форми (16909 различитих) и 268 цифара (10 различитих).

Ипак, оно што завређује нарочиту пажњу у овом одељку јесте постојање извесних речи које су погрешно кориговане у апликацији „Слово Мајстор”. Прецизније, одређеним речима, које су биле правилно написане у оригиналу (без дијакритичких знакова), применом постојећих алгоритама враћени су дијакритички знаци, чиме је дошло до нарушавања семантике речи. Репрезентативни примери наведене праксе јесу речи „zvjezdiće” или „šteko” – уместо исправних „zvjezdice” и „steko” из оригиналног корпуса (уколико се изузме непостојање апострофа из друге речи). Морфолошки речник је исправио неке исправио погрешно транслитериране речи, тако да је нпр. реч „isečka” враћена у „isecka”.

Muskarci	Zene	Sastavi
5	4	21

Text_Marked_As_Changed	Left_Context	Corrected_text	Right_Context
*7(milošti(1)_milosti(69))	zamku past će kad tada U trci ovoj	milosti	neeeeema Niko da čeka ko da drema
*7(milošti(1)_milosti(69))	********	milosti	********
*7(milošti(1)_milosti(69))	********	milosti	********
*7(baća(1)_bača(6)_baca(96))	čeka ko da drema Ulična svetiljka,	baca	svetlost u krug dok crni leptiri l
*7(bezi(1)_beži(93))	ila krila mi spržila crni leptiru,	beži	u noć jutro sačekaj svetlost će do
*7(uže(72)_uze(61))	sačuva No uzalud beše, more ga već	uze	Postavši crveno kao majske ruže Je
*7(šume(224)_sume(152))	uže Jednog toplog dana usred guste	sume	medved ispi med i *7(sok(68)_šok(1
*7(sok(68)_šok(111))	usred guste sume medved ispi med i	šok	uhvati za sovu, s njom poče igrati
*7(šuma(220)_suma(164))	igra ko dođe tigar zvani Kan sudna	suma	je to Na pečurki jedno bubamara ig
*7(miša(127)_misa(11))	prvi put slonica je opet uhvatila	misa	pa se vrti s njim u krug mišića sa
*7(što(36850)_sto(1268))	svoju čitavu svu ljubomorna mečka	sto	joj muž sa sovom igra ovaj ljuti r
*7(sok(68)_šok(111))	********	šok	********
*7(strašno(186)_strasno(31))	(S) Ona prođe brzo, najbrže a noge	strasno	duge, one najduže vešto se igra lj
*7(oči(818)_oci(24))	žurno Kafu pijem, nestajem Trljam	oci	sanjive Da mi ne bi zaspale Vrata
*7(Koraci(86)_Korači(1))	dne Odlazim, dok još drugi spavaju	Koraci	, pločnikom odzvanjaju Osvrćem se o
*7(puste(78)_puš te(1))	odzvanjaju Osvrćem se okolo Gledam	puste	ulice Krećem stazom poznatom Pušta
*7(Koraci(86)_Korači(1))	********	Koraci	********
*7(lisce(1)_lišce(1)_lišće(60))	d otplovim Stari grad brije Košava	lišće	pokriva sve prazan kej i golo drve
*7(uže(72)_uze(61))	********	uze	********
*7(uže(72)_uze(61))	********	uze	********
*7(Isečka(5)_Isecka(1))	s I ujeo kuvara Kuvar uze, uze nož	Isecka	ga na dva, na tri komada Došli, do

Слика 6: Рестаурација дијакритика морфолошким речницима.

На слици 7 могу се видети наведене речи настале погрешном рестаурацијом дијакритика које морфолошки речник (у табели означен са NemaSvalje) није успео да препозна (на тим местима налази се цртица):

token	out	freq	len	grouped	NemaOrig	NemaSvalje
šteko		2	5		-	-
šovom	low-freq	1	5		-	-
osiječam	low-freq	1	8	osiječam	-	-
špat	low-freq	1	4		-	-
zvjezdiće		2	9	zvjezdiće	-	-
kučaj		3	5		-	-
švabice	low-freq	1	7	švabice	-	-
gađat	low-freq	1	5		-	-
šnjegovi		2	8	šnjegovi	-	-
droče		2	5		-	-
nedotučan	low-freq	1	9	nedotučan	-	-
ćerkiće	low-freq	1	7	ćerkiće	-	-
šid	low-freq	1	3		-	-
murijači	low-freq	1	8	murijači	-	-
šnjeg	low-freq	1	5		-	-
tražit	low-freq	1	6		-	-

Слика 7: Речи изведене након погрешне рестаурације дијакритика – „Слово Мајстор”.

7.2 Анализа корпуса у LeXimir апликацији

LeXimir[48] апликација нуди подршку и за приказ токена, лематизованих облика, врста речи, фреквенција речи и колокација у DELAF формату. У програму Excel резултати фреквенцијске анализе се могу филтрирати према врсти речи – тако можемо утврдити нпр. које именице (или чак колокације чија је главни члан именица) су најфреквентније у корпусу. У питању је иста .xslx датотека корпуса обрађеног електронским речницима, при чему се у документу прво излистају сви токени, а затим и колокације, које се налазе при дну. Илустрација 8 представљена у наставку представља парцијални приказ резултата који сведочи о присуству колокација у вези са ратом („svetski rat", „novi svet", „vojnu muziku", „ratne filmove"). На основу слике 9 можемо видети да се међу најфреквентнијим именичким токенима јављају и речи „ljubav" и „srce".

15377	bistre oči	bistre oči	N	6
15378	nova godina	Nova godina	N	4
15379	Novi Sad	Novi Sad	N	4
15380	svetski rat	svetski rat	N	3
15381	novi svet	Novi svet	N	3
15382	noćne ptice	noćna ptica	N	3
15383	tam-tam	tam-tam	N	3
15384	prošlog vremena	prošlo vreme	N	2
15385	vojnu muziku	vojna muzika	N	2
15386	železnička stanica	železnička stanica	N	2
15387	crno grožđe	crno grožđe	N	2
15388	sunčev zrak	sunčev zrak	N	2
15389	malog medveda	Mali Medved	N	2
15390	zlatne medalje	zlatna medalja	N	2
15391	morske obale	morska obala	N	2
15392	svetla budućnost	svetla budućnost	N	2
15393	ratne filmove	ratni film	N	2
15394	filmove u boji	film u boji	N	2
15395	kišni oblak	kišni oblak	N	2
15396	bambusov štap	bambusov štap	N	2
15397	foto-modele	foto-model	N	2
15398	lošem glasu	loš glas	N	2
15399	bijele medvjede	bijeli medvjed	N	2
15400	Bosna i Hercegovina	Bosna i Hercegovina	N	2

Слика 8: Фреквенција именичких колокација.

[48] http://korpus.matf.bg.ac.rs/soft/LeXimir.html

Oblik	Lema	POS	Frequency
dan	dan	N	299
Al	Al	N	231
noć	noć	N	228
do	do	N	224
ljubav	ljubav	N	201
srce	srce	N	196
život	život	N	195
put	put	N	184
meni	mena	N	164
meni	meni	N	164
kraj	kraj	N	155
noći	noć	N	147
biti	bit	N	132
bila	bilo	N	124
grad	grad	N	120
san	san	N	115
sto	sto	N	114
dok	dok	N	113
oči	oči	N	112
kada	kada	N	111

Слика 9: Фреквенција именичких токена.

На основу свега до сада наведеног, следи да нормализација корпуса и његова припрема за напреднију лингвистичку анализу не представља нимало тривијалан задатак. Као што видимо, „шум" међу подацима може се аутоматски детектовати и регулисати применом електронских морфолошких речника који садрже валидне облике речи. Том методом се корисник ослобађа ручног преправљања погрешно рестаурираних речи у апликацијама као што је „Слово Мајстор". Важно је рећи и да аутоматска рестаурација дијакритика вршена помоћу морфолошких речника смањује број непрепознатих речи.

8 Идентификација друштвено-политичких и патриотских тема

8.1 NLTK – Natural Language ToolKit

Полазећи од хипотезе да ће приликом рачунарске анализе текстова нумера бити пронађени неки од лексичких индикатора за наведене теме, Пајтон програму смо задали скуп команди за проналажење токена који имају суфикс -ија. Наиме, очекивало се добијање токена попут *милиц**ија***, *полиц**ија***, *демонстрац**ија*** и сл. Проширили смо упит, тј. свели смо суфикс само на -ија, а не на -ција, који се јавља у наведеним речима.

У наставку ћемо видети да су се у резултату претраге пронашле и речи на -тија или -рија које такође испуњавају задате услове. Треба истаћи и да велики број излистаних речи са суфиксом -ија припада позајмљеницама, што је такође било обележје југо-рока (`'infekcija', 'inflacija', 'anarhija'` итд).

У овом процесу је од нестандардних Пајтон библиотека коришћена функционалност `NLTK`[49] пакета за токенизацију ниске текста, као и Пајтон библиотека `cyrtranslit`[50] ради транслитерације неких текстова са ћирилице на латинично писмо. У наставку је дат репрезентативни кôд који описује поступак проналажења лексичких јединица:

```
lista_tokena = sorted(w for w in set(tokens) if w.endswith('ija'))
```

Наведеном процедуром произведен је скуп од 189 јединствених лексичких ставки са суфиксом -ија, од којих издвајамо двадесет и једну која директно указује на друштвено-политичке аспекте (за разлику од речи *раскошнија, засија, сретнија* итд., које су се такође нашле у изводу и завршавају се на -ија, али нису део концептуалног кластера о коме је реч):

```
['akcija', 'anarhija', 'armija', 'avijacija', 'birokratija', 'delija', 'demagogija',
'demonstracija', 'industrija', 'inflacija', 'informacija', 'jugoslavija', 'legitimacija'
'malverzacija', 'milicija', 'murija', 'nacija', 'partija', 'policija', 'sankcija',
‘sudija’][…]
```

Из `NLTK` документације закључујемо да се помоћу овог алата не врши *стемирање* (проналажење корена речи) ни *лематизација* (свођење на речнички облик) речи на српском језику (наведене функционалности су подржане за 15 светских језика)[51]. Када је у питању рад са лематизованим облицима, препоручује се коришћење Unitex/GramLab[52] софтвера за обраду корпуса. Иначе, у резултатима претраге су се нашле и именице у номинативу јд. са суфиксом -ија, али и облици 3. л. јд. глагола (нпр. „одбија”) или присвојне придевске заменице (нпр. „чија”). Закључак је да добијени резултати могу у извесној мери да одговоре на наше захтеве, али и да за собом повлаче одређене непрецизности.

На основу датих резултата и провере контекстâ који дефинишу добијене токене (кроз конкорданцу) бива јасније да се аутори нису устручавали да у својим нумерама изразе и реакционарске ставове. Овакво испољавање наведених ставова се нарочито односи

[49] https://www.nltk.org.
[50] https://pypi.org/project/cyrtranslit/.
[51] https://www.nltk.org/api/nltk.stem.html.
[52] http://unitexgramlab.org.

на контракултурне покрете *панк* и *нови талас*, којима су се инспирисали бројни југословенски рок музичари. Штавише, они су отворено, неретко уз коришћење погрдних речи, критиковали власт и доводили у питање устаљени систем вредности који она пропагира (Обрадовић, Арсенијевић & Шкорић, 2016: 114-115; 122). Један од значајнијих бендова југословенског панка, тј. *новог примитивизма*, јесте Забрањено Пушење, у чијој се нумери „Од историјског АВНОЈ-а" јасно манифестују наведене карактеристике (речи са суфиксом -ија које се бележе у тој песми јесу *парт**ија*** и *акц**ија***). Са друге стране, представници новог таласа такође су били политички ангажовани (попут састава Азра, како се наводи у Brkić, 2011: 447), с тим што се новоталасна музика супротстављала постојећем друштвеном поретку кроз херметичнији и софистицираније стил. Прецизније, Божиловићева помиње да се у песми „Маљчики" групе Идоли пародира совјетски социјалистички реализам путем ироније, метафора и алузија (Božilović, 2013: 73-74).

Стихови који следе издвојени су из нумере „Манифест" групе Кербер, и на илустративан начин осликавају дух „југо-рока" у виду употребе речи са суфиксом -ија:

> Kako trune drzava - **milicija**, **armija**,
> kad nestane **partija** - republika, pokrajina,
> svako od nas bice i predsednik i **sudija**,
> svako od nas govorice preko **radija**. [53]

Божиловић у свом раду анализира и концепт *југословенског идентитета*, који се сматра често обрађиваном темом у сфери „ex-Yu рока" (Божиловић, 2016: 268). Патриотизам заузима значајно место у том корпусу, будући да се у великом броју композиција опева управо Југославија. На основу претходно изложеног кода могу се уочити одређене речи које упућују на наведену идеју (неке више, а неке мање директно)[54].

```
['jugoslavija', 'romanija', 'sarajlija', 'slovenija', 'srbija, 'šumadija']
```

Видимо, дакле, да су присутни именовани ентитети у виду топонима (*Југославија*) и демонима (*Сарајлија*). Што се тиче речи *авлија*, која се такође наводи у резултату, она се јавља у композицији под називом „Цијела Југа, једна авлија", у извођењу Дина Мерлина. Већ из наслова нумере закључујемо да реч *авлија* служи као метафора за

[53] http://lyrics.wikia.com/wiki/Kerber:Manifest.

[54] У корпусу је присутна и реч *bukurija* [sic], што највероватније представља старо хрватско име. За више детаља о овом имену, в. https://actacroatica.com/en/name/Bukurija/.

мултинационалну југословенску државу. Поред тога, у наставку ћемо видети да се у стиховима који претходе наведеној фрази јављају још неки именовани ентитети. За њихово откривање могао би се користити модел за препознавање именованих ентитета, обучаван помоћу програмског алата `spaCy` и корпуса за евалуацију именованих ентитета у српском језику (Krstev et al., 2011). Тим поступком би се побољшало тумачење према којем се у југословенском рокенролу промовише култ *братства и јединства*, једне од „тековина Народноослободилачке борбе” ” (Кољанин, 2011: 448), а касније и крилатице СФРЈ[55] (Dimitrijević, 2001: 141).

> Šizi **Beograd**, šizi **Novi Sad**
> **Tuzla**, **Sombor**, **Zagreb**, **Titograd**
> Cijela **Juga** jedna avlija […] [56]

Информацију о постојању друштвено-политичких тема у корпусу можемо претраживати и кроз колокације:

```
bez tebe; sve što; neki novi; zbog tebe; novi klinci; toplim vjetrom;
sretno dijete; vjetrom juga; prvi put; jedan dan; voli ženu; dernek
utihne; dan republike; rock cirkus; ulični hodač; vozi mercedes; ove
noći; haile selasije; sine mitre; poskočiću drugu
```

Од 21 понуђеног резултата, уочене су 3 колокативне јединице које се могу довести у везу са југословенским елементом:

```
dizem zastavu; dan republike; dernek[57] utihne
```

Такође је занимљива употреба колокације `haile selasije` у истоименој песми групе Забрањено Пушење, посвећеној последњем етиопском цару. Подсећања ради, Селасије је остао упамћен по томе што је бранио домовину од италијанске инвазије на Етиопију 1935. године; међутим, због свог начина владавине сматран је и за диктатора (Slobodna Bosna, 2018). Са друге стране, постоје сведочења да је ауторитативни режим био изузетно заступљен и у Југославији за време владавине комунистичког државника

[55] Такође, Роксандић сматра да је половином осамдесетих година и даље владао дух братсва и јединства, као и да се родољубље испољавало управо кроз популарну културу (Roksandić, 2017: 15).

[56] DinoMerlin:CijelaJugaJednaAvlijaLyrics.http://lyrics.wikia.com/wiki/Dino_Merlin:Cijela_Juga_Jedna_Avlija.

[57] Према Московљевићу(2000: 148), дернек је врста „народног сабора, вашара”, и познато је да се тај термин користи на подручју Босне и Херцеговине и Црне Горе.

Јосипа Броза Тита (Anđelic, 2004: 35). Исто тако, сматра се да се босанскохерцеговачка група у својим песмама често служила иронијом и пародијом како би изразила незадовољство конзервативним вредностима у друштву (haler, 2008). Стога, иако песма наизглед слави долазак етиопског владара у Југославију, на нивоу ширег контекста можемо претпоставити да нумера заправо носи са собом извесну дозу бунта и политичке критике.

8.2 spaCy

`spaCy`[58] представља специјализовану програмску библиотеку за обраду природних језика. Сматра се најбржим програмским оквиром за обраду природних језика, при чему се за обучавање неких модела користе и неуронске мреже. Поседује богату палету функционалности, као што су *токенизација* (енгл. *tokenization*), *препознавање именованих ентитета* (енгл. *named-entity recognition – NER*), *граматичко тагирање* (енгл. *part-of-speech tagging)* и др.[59] Статистички модели `spaCy`-ја се увелико користе за седам светских језика, док је развој система за српски језик у фази развоја[60]. Ипак, у оквиру нашег истраживања тестирали смо могућности NER модела који се обучава на новинским текстовима за српски језик. Поред испитивања ефикасности модела, циљ је такође био олакшати претрагу властитих имена из домена политике, који се јављају у нашем корпусу.

Дакле, одабрали смо текст поменуте песме „Хаиле Селасије" од Забрањеног Пушења, где је алгоритам успевао да препозна одређене именоване ентитете, као што су „Хаиле Селасије", „Насер"[61] и „Ганди"[62]. Што се тиче географских подручја, тачно је препозната реч „Азија", али не и „Африка". Са друге стране, обучавани модел региструје као именоване ентитете и неке лексичке јединице које почињу великим словом јер се налазе на почетку стиха, иако не припадају домену властитих имена (нпр. „Obuko [sic] sam bijelu kapicu/**I** crveni šal/Učitelj je rek'o/**Dolazi** nam car"). То значи да ће у будућности

[58] https://spacy.io.

[59] У равни са овим софтвером стоји и Stanford CoreNLP библиотека, настала под окриљем Групе за обраду природних језика при Универзитету Стенфорд (Manning et al., 2014). Без даљег разматрања овог алата у овом раду, споменућемо само да је помоћу ње такође могуће вршити наведене анализе текста.

[60] http://ner.jerteh.rs/.

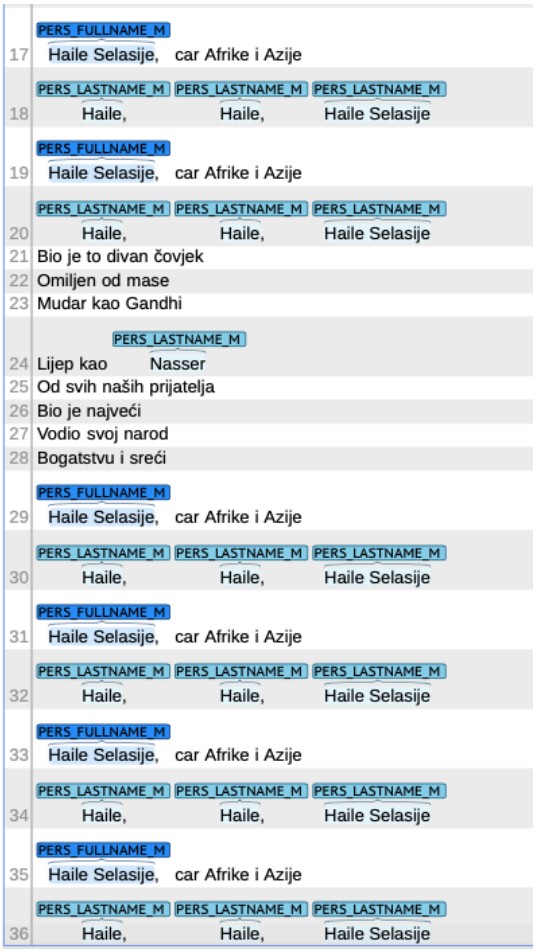
[61] Гамал Абдел Насер, председник Египта који је учествовао у стварању Покрета несврстаних https://www.jewishvirtuallibrary.org/

[62] Махатма Ганди, индијски политичар и друштвени активиста, чувен по својој борби за индијску независност од британског колонијализма https://www.britannica.com/biography/Mahatma-Gandhi

бити потребно поставити додатна ограничења за спецификацију именованих ентитета како би се даље усавршио постојећи NER модел за српски језик.

Занимљиво је споменути да је су именовани ентитети у текстовима на српском језику визуализовани помоћу BRAT веб апликације за анотирање докумената, која се користи у области обраде природних језика (Stenetorp et al., 2012). На тај начин се могу добити прецизнији метаподаци о именованом ентитету и утврдити до ког нивоа се препознају ентитети. Дакле, тренутно се обучавају три различита модела препознавања која обухватају:

- Имена особа уопште;
- Имена особа: пуна, само име и само презиме;
- Имена особа: пуна, само име и само презиме, уз информацију о полу.



Слика 10: Примена обученог NER модела за српски језик помоћу библиотеке spaCy.

9 Визуализација корпуса

9.1 Формирање листе функционалних речи српског језика

Речи које нису појмовно богате и не доприносе даљој анализи, као што су предлози, чланови, помоћни глаголи и др, називају се *функционалне (стоп) речи*, и углавном се изостављају из рачунарске анализе текста (Way, Somers, & Carl, 2003: 22). Стога се креира листа функционалних речи која је произвољног карактера, будући да њен садржај зависи од врста речи заступљених у неком језику (нпр. у неким језицима, као што је српски, не постоје чланови).

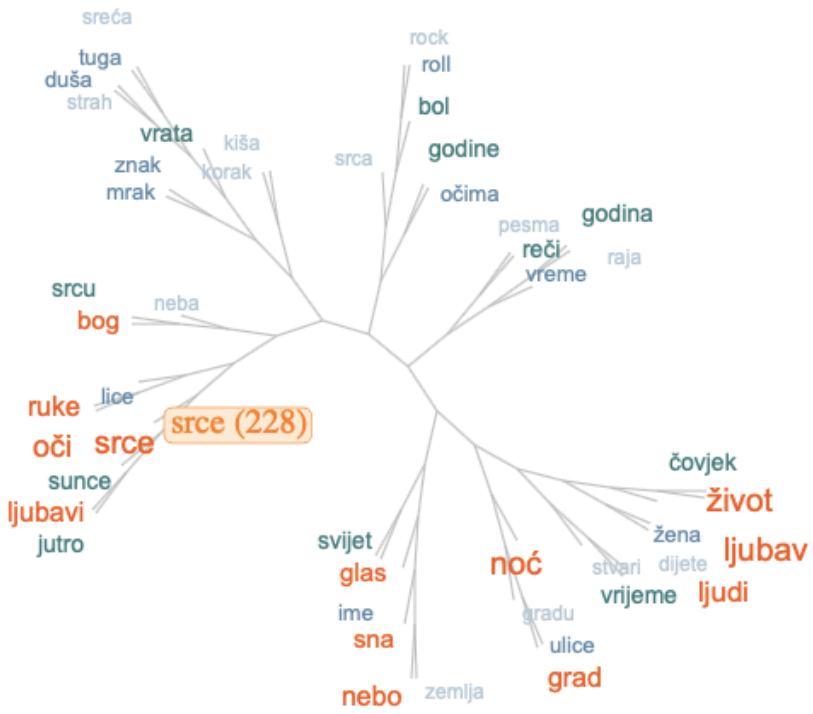
Са друге стране, приликом креирања стоп-листе се узима у обзир и природа текстова у којима се функционалне речи јављају. Ова тврдња је поткрепљена наводима Лиондиса (Leondes, 2005: 125) по питању начина екстракције стоп-речи из информатичке литературе. Наиме, јасно је да ће у том случају и реч „компјутер" бити третирана као функционална реч, јер ће се она јављати велики број пута, стога аутор такве речи назива *доменски зависним* (енгл. *domain dependent*). У контексту обраде текстова песама југословенског корпуса, стандардне стоп-речи српског језика биће допуњене и лексичким јединицама типичним за овакву врсту садржаја (нпр. варијантама за ознаку рефрена: „ref", „ref.", „Ref.", „Refren", „REF." и „chorus").

9.2 Генерисање облака дрвета помоћу алата TreeCloud и WordItOut

Под руководством француских професора са Универзитета Париз-Исток Марн ла Вале (франц. L'Université Paris-Est Marne-la-Vallée), креирана је TreeCloud веб апликација, намењена визуализацији и анализи текстуалних података (Gambette & Véronis, 2009). Као што сам назив алата сугерише, сврха оваквог начина представљања је двојака, будући да се комбинује графички приказ *облака речи* (франц. *nuage de mots*) и дрволике структуре речи (франц. *arbre de mots*). Штавише, корисник може лако закључити не само које су најфреквентније лексичке јединице, већ и које су преовлађујуће теме у документу. Величина речи указује на фреквенцију лексичке јединице, док се концептуално сличне речи („čovjek", „žena", „dijete") гушће концентришу на релативно јасно издвојеним гранама. У вези са тим, Парезановић (2017: 46-47) истиче да ће семантичка сличност двеју речи бити представљена растојањем између њих у тексту.

Ради квалитетније анализе корпуса, неопходно је да улазни материјал буде препроцесиран барем на неком нивоу (било у виду вршења *лематизације*, *стемирања* или *уклањања стоп-речи из текста*). Наш задатак се састојао у томе да, за почетак, из постојећег корпуса елиминишемо српске стоп-речи. Постоје пројекти (као што је Пајтон библиотека `stopwords` [Savsd] или `NLTK`, у оквиру којих су развијене стоп-листе речи за неке светске језике, али не и за српски. У потрази за приближним решењем, открили смо да постоји једна стоп-листа и за српски језик (Champion, 2009). Ипак, скрећемо пажњу на недостатке у погледу садржајности ове листе. Пре свега, јавља се недоследност у приказивању специјалних слова са дијакритичким знацима (нпр. слово „č" се замењује са „è", као у речи „uèinio", док слова попут „š" остају непромењена). Затим, у листу нису укључени сви флективни облици помоћног глагола „бити", али ни неких других променљивих врста речи (нпр. личних заменица: „он", „она" итд.).

Да бисмо додатно побољшали резултат каснијег проналажења најфреквентнијих речи или доминантих тема, допунили смо досадашњу верзију наведене листе функционалних речи уносом доменски зависних и често коришћених речи (неких прилога, предлога, речци итд). Облак речи на нивоу пречишћеног „ex-Yu" корпус може се видети на слици 11.



Слика 11: Кластеризација најфреквентнијих речи у TreeCloud апликацији.

На основу приказаног облака дрвета, бива јасно да су у корпусу често присутни мотиви љубав – издвојиле су се речи „ljubav”, „ljubavi”, „srce” (број 228 представља фреквенцију њеног појављивања на нивоу читавог корпуса) и „srca” у нешто мањој мери. Очекивало се да ће одређени удео у облаку дрвета заузети и друштвено-политичке теме; ипак, на основу ове визуализације се изводи закључак да мотиви који потичу из тог домена нису у довољној мери заступљени у корпусу да би били статистички значајни, већ да су ограничени на мали број песама. Исто важи и за стране речи, међу којима су једино присутне речи „rock” и „roll”.

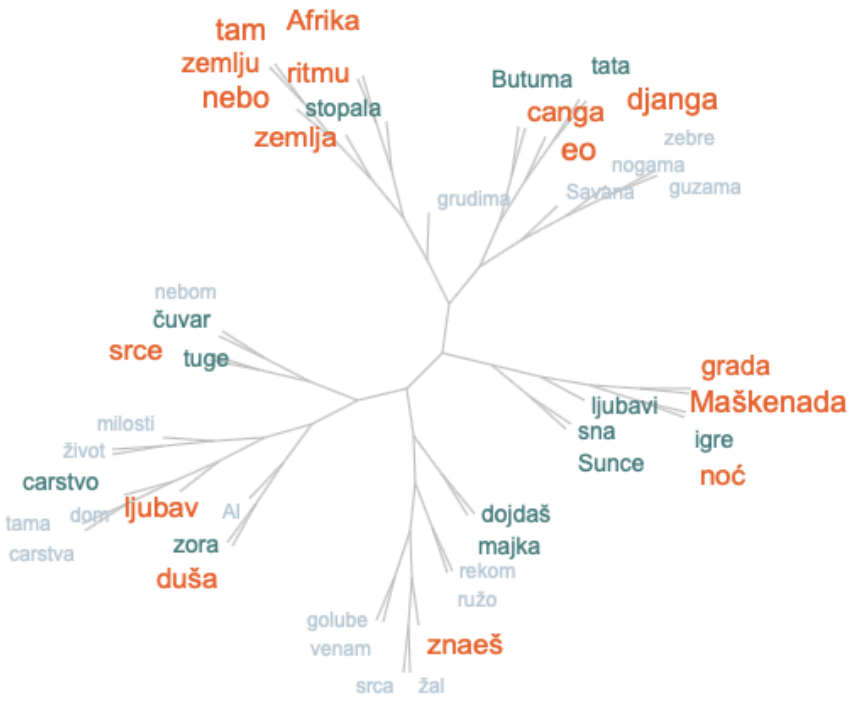
`TreeCloud` пружа занимљив увид у грану дрвета на којој су смештене лексичке јединице „duša”, „tuga”, „strah”, „sreća”, „mrak” (и „бол”, недалеко од њих), које би могле описати концепт *осећања*. Са друге стране „žena”, „dijete”, „čovjek”, „ljudi” означавају жива бића. Делови тела су представљени речима „ruke”, „oči”, „lice”, „srce”, док речи „noć”, „gradu”, „grad”, „ulice” указују на опевање урбаног живота. Посматрано у бројкама, десет најфреквентнијих речи су *noć*, (251 појављивање) *ljubav* (232), *život* (225), *oči* (161), *ljudi* (124), *grad* (122), *ljubavi* (115), *nebo* (103), *ruke* (97) и *sunce* (69).

Други вид визуализације целокупног корпуса је спроведен у софтверу WordItOut[63], на основу које се могу уочити само најфреквентније речи у колекцији текстова (слика 12).

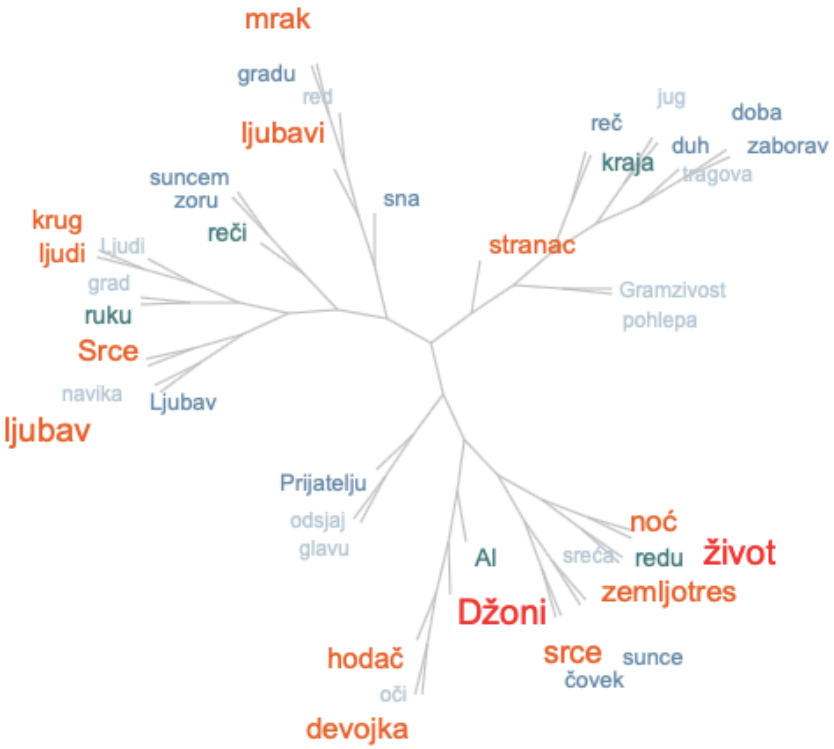
Слика 12: Креирање облака речи корпуса у алату WordItOut.

Исти софтвер можемо користити и за упоредну анализу текстова песама и њихових извођача:

[63] https://worditout.com.



Слика 13: Облак речи за Мадам Пиано.



Слика 14: Облак речи за Партибрејкерсе.

9.3 Стилометријска анализа у R-у

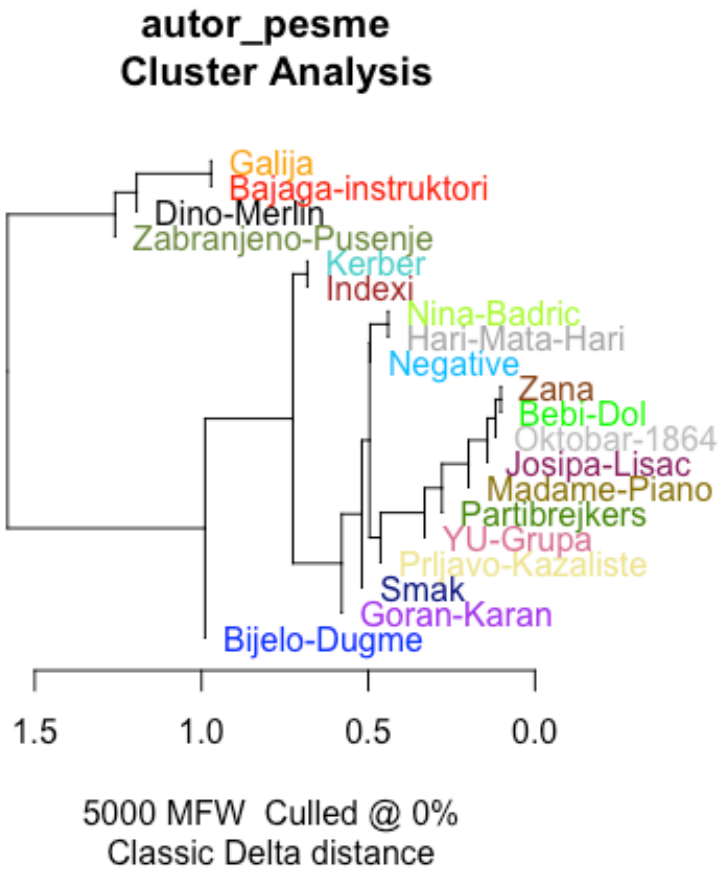
За крај, представићемо и имплементацију библиотеке `stylo` у програмском језику R [64], ради спровођења *стилометријске* анализе текстова песама. *Стилометрија* се у начелу примењује када је потребно извршити верификацију ауторства, и често се користи и у форензичкој лингвистици. Касније су њени методолошки конструкти развијани и у правцу одређивања сличности међу документима на основу њихових стилистичких карактеристика. У овом истраживачком контексту, библиотека `stylo` омогућава вршење различитих статистичких прорачуна, попут *кластер анализе, вишедимензионалног скалирања или анализе основних компоненти*).

Основни разлог због којег уводимо ову врсту анализе у наш рад јесу резултати добијени овим путем. У овом случају, скуп података представљају појединачни .txt фајлови свих песама одређеног аутора, са циљем израчунавања растојања међу документима, која служе као показатељ сличности, односно различитости текстова и аутора. На основу спроведене кластер анализе, за коју се генерише и пратећи визуални приказ, долазимо до закључка да су аутори груписани у оделите кластере према жанровској сличности. Другим речима, групе Галија, Бајага и Инструктори и Забрањено Пушење, или Кербер и Индекси, међусобно су слични по текстовима песама. Са друге стране, женски извођачи Зана, Беби Дол, Јосипа Лисац и Мадам Пиано су оформиле сопствени кластер, и знатно се разликују од наведених аутора, Бијелог Дугмета, Прљавог Казалишта итд. Разлог томе можемо тражити у томе што текстови песама поменутих женских извођача (осим Јосипе Лисац и групе Негатив) не носе у довољној мери дозу бунтовничког става којима се одликују рок бендови. За генерисање графика стилометријске анализе потребно је исписати четири линије кода, које врше учитавање библиотеке `stylo` (уколико је она већ инсталирана), подешавање радног директоријума, креирање корпуса од појединачних докумената, уз активирање графичког интерфејса и формирање корпуса од скупа датотека и генерисање резултата.

Алгоритам 3: Стилометријска анализа песама у програмском језику R.

```
library("stylo")
setwd("/Users/ljudmilapetkovic/stilometrij/autor_pesme_album/autor_pesme")
stylo()
stylo(gui = F, mfw.min = 1000, mfw.max = 5000)
```

[64] https://www.r-project.org.



Слика 15: Рачунање сличности међу извођачима стилометријском анализом у R-у.

10 Закључак

10.1 Закључна разматрања

У раду је представљен пројекат формирања и корпусне анализе текстова песама извођача из жанра југословенског рокенрола у периоду 1945-2003. Имплементиране су Пајтон библиотеке `lyricsmaster` и `yattag` ради аутоматског прикупљања текстуалног садржаја и његовог анотирања по правилима XML синтаксе. Извршено је неколико етапа препроцесирања садржаја по питању садржаја и форме (уклањање сувишних ниски и аутоматска рестаурација дијакритика). Приказани су и основни облици рачунарске обраде текстова – најпре је XML датотека корпуса трансформисана у XHTML формат, погодан за објављивање основних статистичких података на вебу, који су потом визуализовани у Excel програму. У раду са `NLTK` алатом у Пајтону идентификоване су друштвено-политичке и патриотске теме у корпусу, као и релативно честа употреба

страних речи; на основу приказа облака речи у софтверу `TreeCloud`, као преовлађујући мотиви су се издвојили појмови у вези са љубављу, градом, осећањима и деловима тела. Примена LeXimir апликације пружила је увид у фреквенцијску анализу токена/колокација из корпуса, податке о врстама речи, метрику кључности токена, синтаксичко-семантичке и друге ознаке. Упоредном анализом начина рестаурације слова са дијакритичким знацима у софтверима „Слово Мајстор" и LeXimir, дошло се до закључка да је за добијање квалитетнијих резултата сврсисходнија употреба српских морфолошких речника. Општи закључак јесте да се доступним рачунарским алатима могу уочити извесни лингвистички шаблони, што даје простора за дубља тумачења садржаја текстова. Најзад, на основу стилометријске анализе показало се да су изразито рок извођачи (Партибрејкерси, Бајага и Инструктори, Бијело Дугме) међусобно сличнији по садржају текстова песама које изводе у односу на „мање рок" извођаче.

10.2 Будући рад

Пошто није било могуће прикупити текстове песама за све жељене групне или соло извођаче (нпр. Екатарину Велику, Рибљу чорбу, Ђорђа Балашевића, Џибонија итд.), остављамо простора за даље обогаћивање постојећег списка извођача. Исто тако, планирано је да корпус буде допуњен и текстовима песама женских југословенских рок-извођача, попут Маје Оџаклијевске, Слађане Милошевић, Калиопи итд. Накнадно ће бити додати и текстови нумера са веб-страна и посвећених песмама на хрватском[65] и босанском[66] језику, које се нису нашле на претходно наведеној страни са обједињеним југословенским текстовима песама (Category:Language/Serbian).

Конкретно, даљи рад би евентуално укључивао коришћење измењеније методологије за проналажење текстова недостајућих аутора. Једно од могућих решења тог проблема била би имплементација модула библиотеке `BeautifulSoup` за парсирање HTML страница. Извори рашчлањивања података са интернета ради екстракције жељених података биле би странице MetroLyrics и Tekstovi.net[67], које садрже текстове нумера наведених извођача. Иначе, други сајт је окарактерисан управо као „галерија музичких текстова са подручја БиХ, Црне Горе, Хрватске и Србије".

[65] http://lyrics.wikia.com/wiki/Category:Language/Croatian.
[66] http://lyrics.wikia.com/wiki/Category:Language/Bosnian.
[67] https://tekstovi.net/2,0,0.html.

Такође, будући рад би се тицао и допуне постојећег корпуса текстовима и у случајевима где постоји диспропорционалност између укупног броја песама које је аутор објавио у току каријере и броја доступних песама на сајту LyricWiki. Пример тога јесте група Ван Гог, која иза себе има богату дискографију, али је корпусом забележено свега два текста песама[68]. На тај начин би накнадно богаћење колекције повећало степен унифицираности података проширеног корпуса.

Даљи правци развоја тичу се и коришћења библиотеке `stylo` у програмском језику R, где је могуће је вршити *стилометријску анализу* у оквиру *компутационе стилистике*. Предмет изучавања *стилометрије* (енгл. *stylometry*) тиче се утврђивања ауторства и примене квантитативних метода у истраживању стила којим је неки текст написан. Кључна компонента овог алата огледа се у проналажењу сличности и разлика међу улазним документима кроз различите статистичке метрике сличности (*кластер анализа, вишедимензионално скалирање, анализа основних компоненти*).

На примеру нашег корпуса, хтели смо да утврдимо колика је сличност између извођача на основу садржаја текстова њихових песама. Пајтон кодом смо генерисали појединачне текстове песама за све ауторе, а сличност између аутора је визуализована у програму RStudio[69]. Резултат анализе показује да су текстови песама (а самим тим, и њихови извођачи) Партибрејкерса, Бајаге и Инструктора и Бијелог Дугмета сличнији у односу на извођаче попут Мадам Пиано или Нине Бадрић.

Такође је креирана и рудиментарна функција за прикупљање свих текстова песама на одређеном албуму за одређеног извођача. Очекује се да тај кôд буде у будућности коригован, како би се само на основу аутора као задатог аргумента аутоматски генерисали текстови на свим његовим албумима. Самим тим, проучавање сличности албума и еволуције мотива кроз време биће остављено за будуће истраживање.

[68] Упоредити http://www.musicvangogh.com/diskografija/ и http://lyrics.wikia.com/wiki/Van_Gogh.
[69] `https://www.rstudio.com.`.

Литература

1. .DS Store. (s.d). https://www.revolvy.com/page/.DS-Store (приступљено 15.11.2018).
2. Acta Croatica. Name Bukurija. https://actacroatica.com/en/name/Bukurija/. Датум приступа: 9. фебруар 2019.
3. Alspaugh, T. A. (s.d). XSL Transformations (XSLT). https://thomasalspaugh. org/pub/fnd/xslt.html. (приступљено 16.10.2018).
4. Anaconda. (s.d). https://www.anaconda.com (приступљено 09.11.2018).
5. Antić, M. (2011). Informacioni sistemi za podršku alternativnom poslovnom odlučivanju na bazi ekonomskih pokazatelja (Neobjavljeni master rad). Univerzitet Singidunum, Beograd, Srbija.
6. Аурора. http://aurora. jerteh.rs//. (приступљено 09.02.2019).
7. Beautiful Soup Documentation. (2015). https://www.crummy.com/software/BeautifulSoup/bs4/doc/ (приступљено 05.12 2018).
8. Божиловић, Н. (2016). Култура сећања и југословенски рокенрол. *Култура* (152), 257-280.
9. Božilović, J. (2013). New wave in Yugoslavia: Socio-political context. *Facta universitatis-series: Philosophy, Sociology, Psychology and History, 12*(1), 69-83.
10. Brkić, A. (2011). Mogućnost društvene promene i popularna muzika u Srbiji: Studija slučaja – Šobaja i "Tri poljupca".
11. Carl, M., & Way, A. (Eds.). (2003). *Recent advances in example-based machine translation (Vol. 21)*. Springer Science & Business Media. *Етноантрополошки проблеми н.с. 6*(1).
12. Champion, J. (Sep 9, 2018). extra-stopwords. https://github.com/Xangis/extra-stopwords/blob/master/serbian. (приступљено 09.11.2018).
13. Chen, I. (2016). Analyze Taylor Swift lyrics using Python (GitHub repository). Доступно на https://github.com/irenetrampoline/taylor-swift-lyrics (приступљено 16.10.2018).

14. Cold, S. J. (2006). Using Really Simple Syndication (RSS) to enhance student research. *SIGITE Newsletter. 3(*1), 6- 9.

15. cyrtranslit. (s.d). PyPI. https://pypi.org/project/cyrtranslit/ (приступљено: 9. фебруар 2019).

16. Dimitrijević, B. B. (2001). Army and the Yugoslav identity 1945-1992 [Abstract]. Vojno delo, 53(2), 141-154.

17. Doyle, J. (2018). View Folder Tree in Mac OSX Terminal https://coderwall.com/p/owb6eg/view-folder-tree-in-macosx-terminal (приступљено 16.10.2018).

18. Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *R journal, 8*(1). Harold, E. R., & Means, W. S. (2006). *XML za programere* (prevod). Mikro knjiga.

19. Expert System Team. (April 11, 2016). Natural language processing and text mining. https://www.expertsystem.com/natural-language-processing-and-text-mining/ (приступљено 11.11.2018).

20. Fitzgerald, M. (2004). XML Hacks: 100 Industrial-Strength Tips and Tools. "O'Reilly Media, Inc.". https://www.oreilly.com/library/view/xml-hacks/0596007116/ch01s02.html (приступљено: 9. фебруар 2019).

21. Gambette, P., & Véronis, J. (2010). Visualising a text with a tree cloud. In Classification as a Tool for Research (pp. 561-569). Springer, Berlin, Heidelberg.

22. Graovac, J. Uvod u XML i XML baze podataka. http://poincare.matf.bg. ac.rs/ ivana-/courses/pbp/pbp.cas10.XMLiXMLBaze.pdf. (приступљено 16.10.2018).

23. haler. Jedinstvena Bosna i Hercegovina. Sarajevski novi primitivizam-Pokret za destrukciju muslimanskog nacionalnog bica. http://haler.blogger.ba/arhiva/ 2008/10/08/1831190 (приступљено 09.11.2018).

24. Haslam, T. J. (2017). Mapping the Great Divide in the Lyrics of Leonard Cohen. *Rupkatha Journal on Interdisciplinary Studies in Humanities, IX*(1), 1-10.

25. Hentschel, E. (2003). The expression of gender in Serbian. *Gender across languages, 3,* 287-309.

26. Horvat, A. S., & Muhvić-Dimanovski, V. (2014). Today we are, tomorrow we are not – A Contribution to the Problem of Hapaxes in Croatian. [Abstract]. In: *Međunarodni skup Riječki filološki dani, Vol. 9* (511-520). Rijeka: Filozofski fakultet.

27. html – HyperText Markup Language support. (s.d). https://docs.python.org/3/ library/html.html. (приступљено 04.10.2018).

28. Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., Kolodziej, J., Wang, L., Chen, D., Rayes, A., Tziritas, N., Xu, C., Zomaya, A. Y., Alzahrani, A. S., & Li, H. (2015). A survey on text mining in social networks. *The Knowledge Engineering Review, 30*(2), 157-170.

29. Janjatović, P. (1998). *Ilustrovana YU rock enciklopedija*: 1960-1997. Beograd: Geopoetika.

30. JePTex – Друштво за језичке ресурсе и технологије. (2018). Лексички ресурси. http://jerteh.rs/?page_id=112#. (приступљено 05.12.2018).

31. Jewish Virtual Library. Gamal Abdel-Nasser (1918-1970). https://www.jewishvirtuallibrary.org/gamal-abdel-nasser (приступљено 09.11.2018).

32. Josipa Lisac. Albumi. (s.d). http://www.josipalisac.com/glazba/. Датум приступа: 9. фебруар 2019.

33. Juric, M., Pehar, F. & Zauder, K. (2016). Računalna pismenost za suvremeno novinarstvo. U: *Zbornik Informacijska tehnologija i mediji 2016*. Zagreb: Sveučilište u Zagrebu, Hrvatski studiji.

34. Kaggle. (s.d). 380,000+ lyrics from MetroLyrics. https://www.kaggle.com/gyani95/380000-lyrics-from-metrolyrics (приступљено 23.11.2018).

35. Kim, W., Ding, Y., & Kim, H. G. (Eds.). (2014). *Semantic Technology: Third Joint International Conference, JIST 2013*, Seoul, South Korea, November 28–30, 2013, Revised Selected Papers, Vol. 8388. Springer.

36. Kojić, A. (2014). Yu rok scena – Srbija. http://www.fil.bg.ac.rs/mmd_27/mmd_2015/rep.php?x=1 (приступљено 26.11.2018). Last.fm. (2018). Ex-Yu Rock Artists. https://www.last.fm/tag/ex-yu+rock/artists (приступљено 11.11.2018).

37. Колањин, Д. (2011). Обликовање новог концепта наставе историје у основним школама у Србији (1948-1952). *Истраживања: часопис за историју*(22), 441-454.

38. Kostić, P. (16.3.2005). Hoćemo cenzuru: YU rockeri i (auto)cenzura. Преузето 12.11.2018. са: https://balkanrock.com/autorski-clanci/kolumne-i-clanci/hocemo-cenzuru-yu-rockeri-i-autocenzura

39. Krstev, C. (s.d.). Kurs iz XML-a. Преузето 25.11.2018. sa http://poincare.matf.bg.ac.rs/čvetana/kurs-xml/.

40. Krstev, C., & Vitas, D. (2005). Corpus and Lexicon-Mutual Incompletness. In: *Proceedings of the Corpus Linguistics Conference* (pp. 14-17). Birmingham: University of Birmingham.

41. Krstev, C., Stankovic, R., & Vitas, D. Knowledge and Rule-Based Diacritic Restoration in Serbian. In: *Proceedings of the Third International Conference Computational Linguistics in Bulgaria (CLIB 2018).*

42. Krstev, C., Vitas, D., Obradović, I., & Utvić, M. (2011, July). E-Dictionaries and finite-state automata for the recognition of named entities. In: *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing* (pp. 48-56). Association for Computational Linguistics.

43. Laurier, C., Grivolla, J., & Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In: *Machine Learning and Applications, 2008.* (pp. 688-693). IEEE.

44. Leondes, C. T. (Ed.). (2005). *Intelligent knowledge-based systems: business and technology in the new millennium*. Kluwer Academic.

45. Linguamatics. (s.d.). What is NLP Text Mining? Преузето 22.11.2018. са https://www.linguamatics.com/what-is-text-mining-nlp-machine-learning.

46. Lukic, A. (2015). A Comparison of Topic Modeling Approaches for a Comprehensive Corpus of Song Lyrics. Pittsburgh: Carnegie Mellon University.

47. Ljubešic, N., Erjavec, T., & Fišer, D. (2016). Corpus-based diacritic restoration for south slavic languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA)(may 2016).

48. Logan, B., Kositsky, A., & Moreno, P. (2004). Semantic analysis of song lyrics. In: Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, Vol. 2 (pp. 827-830). IEEE.

49. lyricsmaster 2.7.21. (2018). https://pypi.org/project/lyricsmaster/ (приступљено 09.11.2018).

50. LyricWiki. (s.d.). Bajaga & Instruktori. http://lyrics.wikia.com/wiki/Bajaga_%26_Instruktori (приступљено 09.11.2018).

51. LyricWiki. (s.d.). Dino Merlin:Cijela Juga Jedna Avlija Lyrics. http://lyrics.wikia.com/wiki/Dino_Merlin:Cijela_Juga_Jedna_Avlija (приступљено 26.11.2018).

52. LyricWiki. (s.d.). http://lyrics.wikia.com/wiki/LyricWiki (приступљено 04.10.2018).

53. LyricWiki. (s.d.). Kerber:Manifest Lyrics. http://lyrics.wikia.com/wiki/Kerber:Manifest (приступљено 26.11.2018).

54. LyricWiki. (s.d.). Language/Bosnian. http://lyrics.wikia.com/wiki/Category:Language/Bosnian (приступљено 05.12.2018).

55. LyricWiki. (s.d.). Language/Croatian. http://lyrics.wikia.com/wiki/Category:Language/Croatian (приступљено 05.12.2018).

56. LyricWiki. (s.d.). Language/Serbian. http://lyrics.wikia.com/wiki/Category:Language/Serbian (приступљено 10.11.2018).

57. LyricWiki. (s.d.). Statistics. http://lyrics.wikia.com/wiki/Special:Statistics (приступљено 09.11.2018).

58. Mahedero, J. P., MartÍnez, Á., Cano, P., Koppenberger, M., & Gouyon, F. (2005). Natural language processing of lyrics. In: Proceedings of the 13th annual ACM international conference on Multimedia (pp. 475-478). Singapore: ACM.

59. Маловић, Г., & Јончић, Д. (2010). Југословенске владе 1918-2006: каталог изложбе. Београд: Архив Југославије.

60. Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

61. MetroLyrics. (s.d). Riblja Corba Lyrics. http://www.metrolyrics.com/riblja-corba- lyrics.html. (приступљено 15.11.2018).

62. Milić, D. (2015). SEO optimizacija i faktori rangiranja veb lokacija (Neobjavljeni master rad). Matematički fakultet, Beograd, Srbija.

63. Miller, T. W. (2015). Web and Network Data Science: Modeling Techniques in Predictive Analytics. Pearson Education.

64. Miner, G., Elder IV, J., & Hill, T. (2012). Practical text mining and statistical analysis for non-structured text data applications. Academic Press.

65. Mirkov, N., & Peranović, M. (2015). Rudarenje teksta sa društvenih mreža API pristupom. INFOTEH-JAHORINA, 14, 584–588. Andjelic, N. (2004). Bosnia-Herzegovina: The end of a legacy. Routledge.

66. Mondo. (25.01.2007). Goran Karan u Beogradu 14. i 15. februara. http://mondo.rs/a48423/Zabava/Muzika/Goran-Karan-u-Beogradu-14.-i-15.-februara.html (приступљено 15.11.2018).

67. Московљевић, М. (2000). Дернек. У: *Речник савременог српског књижевног језика с језичким саветником* (148). Београд: Гутенбергова галаксија.

68. Nanda, B.R. Mahatma Gandhi. https : / / www . britannica . com / biography / Mahatma-Gandhi. (приступљено 09.11.2018).

69. NLTK 3.3 documentation. (2019). Natural Language Toolkit. https://www.nltk.org. (приступљено 26.8.2018.).

70. nltk.stem package. (s.d). https://www.nltk.org/api/nltk.stem.html (приступљено 26.8.2018.).

71. Olston, C., & Najork, M. (2010). Web crawling. Foundations and Trends® in Information Retrieval, 4(3), 175-246.

72. Обрадовић, М., Арсенијевић, А., & Шкорић, М. (2016). Израда мултимедијалног документа 'YU рок сцена'. *Инфотека, 16*(1-2), 113-129.

73. os. https://docs.python.org/3/library/os.html. (приступљено 09.11.2018).

74. os.path (s.d). https://docs.python.org/3/library/os.path.html (приступљено 09.11.2018).

75. oXygen XML Editor. (2018). https://www.oxygenxml.com (приступљено 11.11.2018).

76. Pajić, V. (2012). Modeli konačnih stanja u ekstrakciji informacija (Neobjavljena doktorska disertacija). Matematički fakultet, Beograd, Srbija.

77. Parezanović, V. (2017). Razvoj korpusa tekstova prevedenih sa japanskog jezika na srpski i engleski jezik (Neobjavljeni master rad, preuzet uz dozvolu mentora). Studije pri Univerzitetu u Beogradu, Beograd, Srbija.

78. Paumier, S. (March 27, 2016). Unitex 3.1. User Manual. Marne-la-Vallée: Université Paris-Est Marne-la-Vallée.

79. Petković, Lj. ex-yu-song-corpus. https://github.com/ljpetkovic/ex- yu-songs-corpus (приступљено 09.02.2019).

80. Petrov, A. (2017). In Search of 'Authentic' Yugoslav Rock: The Life and Afterlife of Bijelo Dugme. AM Journal of Art and Media Studies, (13), 43-59.

81. Petrović, J., & Ivanović, M. (2011). Internet–sredstvo komunikacije i distribucije u hotelskoj industriji. CM – časopis za upravljanje komuniciranjem (20), 117-129.

82. Priručnik za data novinarstvo. (s.d.). Nalaženje podataka na internetu. http://prirucnik-datanovinarstvo.media.ba/getting_data_3.html (приступљено 15.10.2018).

83. PyLyrics. (2018). https://pypi.org/project/PyLyrics/ (приступљено 09.11.2018).

84. Python. (s.d). https://www.python.org (приступљено 04.10.2018).

85. Раковић, А. (2011). Бит мода, рокенрол и генерацијски сукоб у Југославији 1965−1967. *Етноантрополошки проблеми, 6*(3), 745-762.

86. Ristivojević, M. (2012). Rokenrol kao lokalni muzički fenomen. Issues in Ethnology Anthropology, 7(1).

87. Roksandić, D. (2017). „Jugoslavenstvo prije Jugoslavije“. U: Jugoslavija u istorijskoj perspektivi (27-54). Beograd: Helsinški odbor za ljudska prava u Srbiji.

88. RStudio. (s.d). https://www.rstudio.com. (приступљено 09.11.2018).

89. Saračević, M., Mašović, S., & Kamberović, H. (2010). Tehnike Text Mining-a i njihova realizacija primenom objektno-orijentisane analize. U: 18. Telekomunikacioni forum TELFOR 2010 (1097-1100). Beograd: TELFOR.

90. Savand. A. (s.d). stop-words.7.23.2018. https://pypi.org/project/stop-words/. (приступљено 09.11.2018).

91. Sharma, A. K., & Gupta, P. C. (2012). Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining. International Journal of Advanced Research in Computer Engineering & Technology, 1(8), 287-293.

92. Slobodna Bosna. BIO JE POSEBAN, NEPONOVLJIV, PRVI: O njemu su pjevali Bob Marley i Elvis J. Kurtović, a njegov narod smatrao ga je MESIJOM.

    https://www.slobodna-bosna.ba/vijest/96397/https://www.slobodna-bosna.ba/vijest/963

    Датум приступа: 9. фебруар 2019.

93. spaCy. Industrial-Strength Natural Language Processing in Python. https : / / spacy.io. (приступљено 09.11.2018).

94. Stachowiak, M. Understanding HTML, XML and XHTML. https://webkit. org/blog/68/understanding-html-xml-and-xhtml/. Датум приступа: 9. фебруар 2019. Sep 20, 2006 (cit. on p. 30).

95. Stanković, R., Obradović, I., & Trtovac, A. (2012). An Approach to Development of Bilingual Lexical Resources. In: Proceedings of the Fifth Balkan Conference in Informatics BCI 2012 (pp. 101-104). Novi Sad.

96. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 102-107). Association for Computational Linguistics.

97. Studenti Katedre za bibliotekarstvo i informatiku Filološkog fakulteta Univerziteta u Beogradu. (2014). Multimedijalni dokument – Yu rock scena. http://www.fil.bg.ac.rs/mmd_27/mmd_2015/knjige.php. (приступљено 15.11.2018).

98. Sweigart, A. (2015). Uvod u Python, automatizovanje dosadnih poslova (S. Prudkov, prev.). Beograd: Kompjuter biblioteka. McEnery, T., & Hardie, A. (2011). Corpus linguistics: Method, theory and practice. Cambridge University Press.

99. Tekstovi.net. (s.d). https://tekstovi.net/2,0,0.html.(приступљено 29.11.2018).

100. The LaTeX project. (s.d.). https://www.latex-project.org (приступљено 11.11.2018).

101. The Open Lyrics Database. (s.d.). https://lyrics.github.io (приступљено 04.10.2018).

102. The R Foundation. The R Project for Statistical Computing. https://www.r- project.org. Датум приступа: 9. фебруар 2019. s.d (cit. on p. 48).

103. The Standard Python Library. (2013). https://docs.python.org/3/library/. (приступљено 09.11.2018).

104. Tor (s.d). https://www.torproject.org. (приступљено 05.12.2018).

105. Unitex/GramLab. (2018). http://unitexgramlab.org (приступљено 11.11.2018).

106. Univerzitet u Beogradu, Matematički fakultet. (s.d). Sintaksa jezika Pajton. http : / / www.edusoft.matf.bg.ac.rs/python/lessons/uvod2.php. (приступљено 02.10. 2018).

107. Van Gogh. (s.d). LyricWiki. http://lyrics.wikia.com/wiki/Van_Gogh

108. Van Gogh. Diskografija. (2016). http://www.musicvangogh.com/diskografija/ (приступљено 09.11.2018).

109. Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., & Stanojević, M. (2012). The serbian language in the digital age. Heidelberg: Springer.

110. W3schools.com. XML Tutorial. (s.d). https://www.w3schools.com/xml/default.asp. (приступљено 09.11.2018). os. (s.d).

111. “What is an API? (Application Programming Interface)” (s.d). //www.mulesoft.com/resources/api/what-is-an-api (приступљено 05.12.2018).

112. Woods, D., & Thoeny, P. (2011). Wikis for dummies. John Wiley & Sons.

113. WordItOut. (s.d). https://worditout.com (приступљено 15.11.2018).

114. XSLT fiddle. http://fiddle.frameless.io. Датум приступа: 9. фебруар 2019. s.d (cit. on p. 31).

115. Zhang, S., Caro Repetto, R., & Serra, X. (2017). Understanding the expressive functions of jingju metrical patterns through lyrics text mining. In: *Proceedings of the 18th International Society for Music Information Retrieval conference* (pp. 397-403). Suzhou: ISMIR.

116. Раковић, А. (2018). Рокенрол у Социјалистичкој Југославији: од забаве градске омладине до националне културе. У: *Сан о граду: зборник радова* (427-439). Вишеград: Андрићев институт.

Попис слика

Прилог

Алгоритам 4: Гребање LyricWiki страна за жељене извођаче.

```
from lyricsmaster import LyricWiki

provider = LyricWiki()
izvodjaci = ['Bajaga', 'Bajaga␣&␣Instruktori', 'Bebi␣Dol', 'Bijelo␣Dugme',
    'Dino␣Merlin', 'Divlje␣Jagode', 'Elektricni␣Orgazam', 'Galija', '
   Goran␣Bregovic', 'Goran␣Karan', 'Hari␣Mata␣Hari', 'Haustor', 'Idoli',
   'Indexi', 'Josipa␣Lisac', 'Kerber', 'Madame␣Piano', 'Mirzino␣Jato', '
   Negative␣(RS)', 'Neverne␣Bebe', 'Nina␣Badric', 'Oktobar␣1864', '
   Partibrejkers', 'Prljavo␣Kazaliste', 'Rani␣Mraz', 'Smak', 'Van␣Gogh',
   'Vesna␣Pisarovic', 'YU␣Grupa', 'Zabranjeno␣Pusenje', 'Zana', 'Zdravko␣
   Colic', 'Zeljko␣Joksimovic']

def korpus(izvodjaci):
  for izvodjac in izvodjaci:
    try:
      discography = provider.get_lyrics(izvodjac)
      for album in discography:
        print('Album:␣', album.title)
        for song in album:
          print('Song:␣', song.title)
          print('Lyrics:␣', song.lyrics)
          discography.save()
    except:
      continue

korpus(izvodjaci)
```

Алгоритам 5: Аутоматска анотација корпуса у складу са XML синтаксом.

```
import os
from os import listdir
from os.path import isfile, join
import html
from yattag import Doc, indent

def replace_char_entities(s):
  return html.escape(s)

myroot = '/Users/ljudmilapetkovic/Documents/LyricsMaster'
doc, tag, text = Doc().tagtext()
authors = [dir for dir in listdir(myroot) if not isfile(join(myroot, dir))
   ]

with tag('exYuPesme'):
  for author in authors:
    curr_path = '{}/{}'.format(myroot, author)
```

```
    albums = [dir for dir in listdir(curr_path) if not isfile(join(curr_
   path, dir))]
    with tag('autor', ime=author, brojAlbuma=len(albums), pol=""):
      for album in albums:
        with tag('album', naziv=album, godina=""):
          album_path = '{}/{}'.format(curr_path, album)
          songs = [f for f in listdir(album_path) if isfile(join(album_
   path, f))]
          for song in songs:
            if not song.startswith('.DS_S'):
              with tag('pesma', naslovPesme=song[:-4]):
                song_path = '{}/{}'.format(album_path, song)
                for stih in open(song_path, encoding="utf8", errors='
   ignore').read().split('\n')[:-1]:
                  with tag('li'):
                    text('{}'.format(replace_char_entities(stih)))

result = indent(
doc.getvalue(),
indentation = '␣'*4,
newline = '\r\n'
)

print(result)
with open('exyuxml␣(bez␣.DS-Store).xml', 'w') as x:
  x.write(result)
```

Алгоритам 6: XSL стилски лист за трансформацију XML корпуса у HTML.

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0"
  xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:template match="/">
    <html>
      <head>
        <style>
          body {
          background-color: white;
          font-family: Geneva;
          }
          h1 {
          color: white;
          text-align: center;
          }
          table {
          margin: 10px -1 330px;
          border: 1;
          background: cornflowerblue
          }
          th {
          background-color: yellow;
          }
          tr {
          background-color: white;
          }
          #th0 {
          text-align: center;
          color: red;
          background-color: #E6E6FA;
          font-size:14px;
          font-style: italic;
          }
        </style>
      </head>
      <body>
        <title>Statistika</title>
        <h2>Statistika</h2>
        <table>
          <tr>
            <td id="th0">Ukupno: <xsl:value-of select="count(distinct-
  values(//autor/@ime))" /></td>
            <td id="th0">Ukupno: <xsl:value-of select="count(//pesma/
  @naslovPesme)" /></td>
            <td id="th0">Ukupno: <xsl:value-of select="count(distinct-
  values(//album/@naziv))" /></td>
          </tr>
          <tr>
            <th>Autori</th>
```

```
            <th>Pesme</th>
            <th>Albumi</th>
          </tr>
          <xsl:for-each-group select="//autor" group-by="@ime">
            <xsl:sort select="count(.//pesma/@naslovPesme)" data-type="
    number" order="descending"/>
            <tr>
              <td><xsl:value-of select="@ime" /></td>
              <td><xsl:number value="count(.//pesma/@naslovPesme)"/></td>
              <td><xsl:number value="count(.//album/@naziv)"/></td>
            </tr>
          </xsl:for-each-group>
        </table>
        <table>
          <tr>
            <th>Muskarci</th>
            <th>Zene</th>
            <th>Sastavi</th>
          </tr>
          <tr>
            <td><xsl:value-of select="count(//autor[@pol='Muski'])" /></td
    >
            <td><xsl:value-of select="count(//autor[@pol='Zenski'])" /></
    td>
            <td><xsl:value-of select="count(//autor[@pol='Grupa'])" /></td
    >
          </tr>
        </table>
      </body>
    </html>
  </xsl:template>
</xsl:stylesheet>
```

Алгоритам 7: Проналажење суфикса на -ија који указују на друштвено-политичке теме.

```
import nltk
import cyrtranslit

def sufiksi():
  with open('ex-yu-korpus.txt', 'r') as f:
    data = str(f.readlines())
    data = cyrtranslit.to_latin(data)
    tokens = nltk.word_tokenize(data)
    tokens = [token.lower() for token in tokens]
    tokens = nltk.Text(tokens)
    lista_tokena = sorted(w for w in set(tokens) if w.endswith('ija'))
    print('Lista tokena:', len(lista_tokena), '\n\n', lista_tokena)

sufiksi()
```

Алгоритам 8: Конкорданца.

```
import nltk
from nltk import bigrams
import cyrtranslit

def kolokacije():

  with open('ex-yu-korpus.txt', 'r') as f:
    data = str(f.readlines())
    data = cyrtranslit.to_latin(data)
    tokens = nltk.word_tokenize(data)
    tokens = [token.lower() for token in tokens]
    tokens = nltk.Text(tokens)
    a = tokens.collocations()
    return a
    print(a)

kolokacije()
```

Алгоритам 9: Генерисање појединачних .txt фајлова за све текстове песама аутора.

```
import os
from os import listdir
from os.path import isfile, join
import html

def replace_char_entities(s):
  return html.escape(s)

myroot = '/Users/ljudmilapetkovic/Documents/LyricsMaster'

authors = [dir for dir in listdir(myroot) if not isfile(join(myroot, dir))
   ]
```

```
def autor_pesme(a, datoteka):
  with open(datoteka, 'w') as x:
    for author in authors:
      curr_path = '{}/{}'.format(myroot, author)
      if curr_path.endswith('{}'.format(a)):
        albums = [dir for dir in listdir(curr_path) if not isfile(join(
   curr_path, dir))]
        for album in albums:
          album_path = '{}/{}'.format(curr_path, album)
          songs = [f for f in listdir(album_path) if isfile(join(album_
   path, f))]
          for song in songs:
            if not song.startswith('.DS_S'):
              song_path = '{}/{}'.format(album_path, song)
              x.write(song[:-4] + '\n\n')
              print(song[:-4], '\n\n')
              for stih in open(song_path, encoding="utf8", errors='ignore'
   ).read().split('\n')[:-1]:
                line = '{}'.format(replace_char_entities(stih))
                print(line)
                x.write(line + '\n')
      else:
        pass

a = 'Madame-Piano'
autor_pesme(a, a + '.txt')
```

Алгоритам 10: Генерисање .txt фајлова са текстовима песама на одређеном албуму.

```
import os
from os import listdir
from os.path import isfile, join
import html

def replace_char_entities(s):
  return html.escape(s)

myroot = '/Users/ljudmilapetkovic/Documents/LyricsMaster'

def autor_album(a, al, datoteka):
  with open(datoteka, 'w') as x:
    authors = [dir for dir in listdir(myroot) if not isfile(join(myroot,
   dir))]
    for author in authors:
      curr_path = '{}/{}'.format(myroot, author)
      if curr_path.endswith('{}'.format(a)):
        albums = [dir for dir in listdir(curr_path) if not isfile(join(
   curr_path, dir))]
        for album in albums:
          album_path = '{}/{}'.format(curr_path, album)
```

```
            if album_path.endswith('{}'.format(al)):
              songs = [f for f in listdir(album_path) if isfile(join(album_
    path, f))]
              for song in songs:
                if not song.startswith('.DS_S'):
                  print(song[:-4], '\n')
                  x.write(song[:-4] + '\n')
                  song_path = '{}/{}'.format(album_path, song)
                  for stih in open(song_path, encoding="utf8", errors='
    ignore').read().split('\n')[:-1]:
                    line = '{}'.format(replace_char_entities(stih))
                    x.write(line)
                    print(line)

a = 'Bajaga-Instruktori'
al = 'Zmaj-Od-Nocaja'
autor_album(a, al, a + '_' + al + '.txt')
```