NORMALIZACIJA TEKSTUALNIH DOKUMENATA NA SPRSKOM JEZIKU U CILJU EFIKASNIJEG PRETRAŽIVANJA U SISTEMIMA E-UPRAVE

Ejub Kajan, *Državni Univerzitet u Novom Pazaru,* ekajan@ieee.org
Aldina Pljasković, *Državni Univerzitet u Novom Pazaru,* apljaskovic@np.ac.rs
Adela Crnišanin, *Državni Univerzitet u Novom Pazaru,* acrnisanin@np.ac.rs

Sadržaj – *Ovaj rad se bavi proučavanjem osobenosti srpskog jezika koje predstavaljaju ključni izazov kod izrade aplikacija čiji je cilj efikasno pretraživanje tekstualnih dokumenata. Akcenat je dat na metode i algoritame za normalizaciju u cilju pripremanja podataka za računarsku obradu, u ovom slučaju operaciju pretraživanja i grupisanja.*

1. UVOD

Napredni sistemi e-Uprave imaju zadatak da pomognu gradjanima u donošenju odluka. Građani postavljaju pitanja sistemu, a sistem nudi odgovarajući odgovor. U e-Upravama na pitanja odgovaraju stručnjaci. Pitanja i odgovori se vremenom akumuliraju i čine neku vrstu baze znanja. Interakcija između građana i sistema odvija se u tri faze: (1) sistem nudi grupu ključnih reči; (2) sistem nudi set pitanja koja su usko povezana sa izabranom ključnom reči i mogućnost postavljanja novog pitanja; (3) sistem nudi odgovor na postavljeno pitanje, pronalazi najsličnije pitanje i odgovor ili formira novi odgovor.

U gore opisanom scenariju interakcije izmedju gradjanina i sistema e-Uprave, centralno mesto zauzima pronalaženje što tačnijeg odgovora na postavljeno pitanje. U klasičnim sistemima e-Uprave odgovore na postavljena pitanja daju eksperti nakon izvesnog vremena koje može varirati od nekoliko sati do nekoliko dana, pri čemu postoji mogućnost davanja polovičnog odgovora i upućivanje na drugi „šalter“ ili drugi web sajt e-Uprave.

Grupa istraživača sa nekoliko visokoškolskih institucija u Srbiji bavi se razvojem jednog inteligentnog sistema za podršku e-Upravi baziranom na sociološkim, biološkim, ekonomskim i drugim observacijama u cilju efikasnije komunikacije izmedju vladinih institucija, bilo po horizonatali, bilo po vertikali (G2G, Goverment-to-Government), kao i komunikacije izmedju vladinih institucija i gradjana (G2C, Government-to-Citizens).

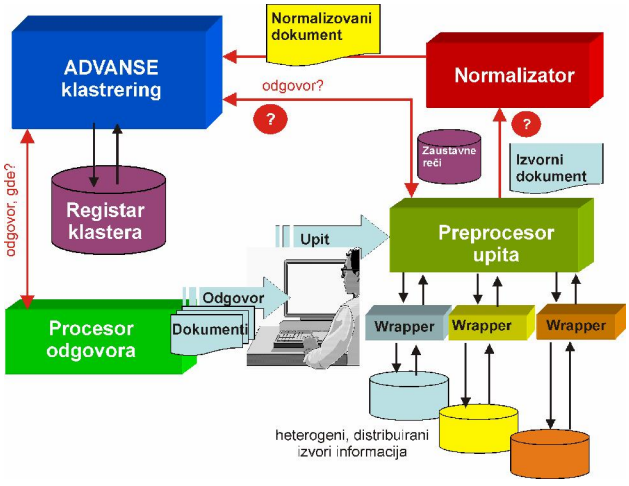
U takvom jednom okruženju postoji veliki broj relevantnih dokumenata, kao i zastarelih dokumenata, koji mogu da utiču na kompletnost i kompetentnost odgovora. U ovom radu prikazani su odredjeni rezultati u toku razvoja G2C podsistema, koji se odnose na postupak normalizacije tekstova na srpskom jeziku za potrebe grupisanja sličnih dokumenata iz heterogenih i distribuiranih izvora e-Uprave.

Trenutni participanti projekta su jedna vladina institucija na nacionalnom nivou i dve lokalne samouprave. Zbog rešavanja administrativnih problema u smislu ko, zašto i kako može da ima pristup odredjenim podacima, prva istraživanja su radjena nad podacima iz akademske sredine, konkretno na forumu Policijske Akademije, koja je takodje učesnik istraživanja.

Naglasak u ovom radu je dat na pojave koje utiču na promenu oblika reči od važnosti, postupke normalizacije i algoritme za normalizaciju rečenice. Ostatak rada je organizovan na sledeći način. Druga sekcija daje kratak osvrt na rad sistema, dok se treća sekcija bavi svojstvima promenljivosti reči i problemima normalizacije. Naredna sekcija sadrži bazične algoritme za normalizaciju rečenice. U cilju provere ispravnosti algoritma, korišćeni su podaci foruma policijske akademije, jer se radi o podacima koja sadrže pitanja i odgovore koji se sastoje od reči koje pripadaju terminologiji jedne oblasti, što je slučaj i u sistemima e-Uprave. Na kraju, izvedeni su zaključci na osnovu dobijenih rezultata i prikazani pravci daljeg istraživanja.

2. KRATAK PREGLED RADA SISTEMA

Grupe ključnih reči su prvi korak i generišu se automatski procesom klasterovanja. Pitanja i dokumenti pripadaju klasteru na osnovu sličnosti termina koje sadrže [1]. Kada korisnik postavi novo pitanje računa se njegova sličnost sa pitanjima u izabranom klasteru. Pojednostavljena arhitektura sistema prikazana je na slici 1. Zbog ograničenog prostora prikazane su samo komponente koje imaju uticaj na normalizaciju ili čiji rezultati zavise od iste.



Sl. 1 *Pojednostavljena arhitektura sistema*

Upit gradjanina preuzima preprocesor upita. Ukoliko se ne može naći odgovor u već postojećem skupu pitanja, odgovor se traži u dokumentima koji se preuzimaju iz dostupnih, bilo lokalnih bilo distribuiranih podataka, preko odgovarajućih wrapper komponenti. U sistemu postoje tri vrste sadržaja: pitanja, formalni dokumenti i odgovori stručnjaka. Ovi sadržaji prestavljeni su u tri sloja: sloj pitanja, sloj odgovora i sloj dokumenata [2]. Da bi se izvršilo

klasterovanje dokumenata i upoređivanje niski mora se analizirati sadržaj dokumenata (kodiranje (obeležavanje) teksta) i prilikom te analize primeniti sve specifičnosti jezika na kome su dokumenti pisani. U ovom slučaju to je srpski jezik.

Kodiranje teksta odnosi na strukturalno (odlomci, rečenice, naslovi itd.) i analitičko (gramatičke kategorije, sintaksne kategorije, itd.) zapisivanje teksta. Osnovne vrste obeležavanja teksta su: segmentacija na rečenice, tokenizacija, označavanje vrsta reči, lematizacija, parsing i semantičko obeležavanje [3]. U semantičkom pogledu, osnovna pretpostavka od koje se polazi je da su tri vrste reči u rečenici nosioci značenja rečenice (u daljem tekstu, **reči od važnosti**). To su:

- Imenice – vrsta reči koja se koristi za označavanje bića, stvari i pojava. Imenice obično imaju glavnu ulogu u strukturi rečenice, i to u svojstvu subjekata (entiteta koji vrše radnju) i objekata (entiteta koji trpe radnju) [4]. Primer: ormar.

- Glagoli – vrsta reči koja označava radnju, stanje ili zbivanje. Obično se u rečenici nalaze u službi predikata, opisujući dinamiku odnosa među imenicama. Primer: hodati.

- Pridevi – vrsta reči koja detaljnije opisuje imenice, i odgovara na pitanje koji, čiji i kakav. U rečenici mogu imati ulogu atributa (kada stoje uz imenice i bliže ih opisuju) i ulogu imenskog dela predikata (kada uz pomoćni glagol iskazuju stanje subjekta). Primer: crven.

Ostale vrste reči nisu glavni nosioci značenja rečenice, i tu spadaju zamenice, brojevi, prilozi, predlozi, veznici, uzvici i rečce. Ove grupe reči nazivamo **zaustavnim** ili **stop rečima**. Za razliku od većine njih, sve vrste reči koje spadaju u reči od važnosti su **promenljive** reči. Ova osobenost srpskog jezika predstavlja usko grlo u izradi softverskih rešenja za pretragu dokumenata na srpskom jeziku, jer se jedna reč može naći u mnoštvu različitih oblika, a efikasan pretraživač treba da prepozna da se radi o jednoj reči (jednom značenju) [4]. Da bi se to postiglo, pitanje koje se postavlja mora proći kroz postupak **normalizacije** (svođenja rečenice na skup reči od važnosti u osnovnom obliku).

Zbog toga prepocesor upita prosledjuje i pitanje i izvorni dokumenat (skup dokumenata) **normalizatoru**, čiji je rad opisan u sekciji 4. Rezultat normalizacije su normalizovani dokument (dokumenti) i normalizovano pitanje koji se predaju **mašini za klasterovanje** relevantnih dokumenata. ADVANSE (Advanced Answering Engine) ima zadatak da formira klasterovane resurse, koji se sastoje od dokumenata, akumiliranih pitanja i odgovora, i da na osnovu toga pruži odgovarajući odgovor gradjaninu. Rad ADVANSE je baziran na hibridnom algoritmu koji kombinuje Fuzzy C i kosinusnu sličnost i detaljno je opisan u radovima [2] i [3]. Za davanje odgovora korisniku zadužen je procesor odgovora, u principu komponenta koja vodi računa o davanju odgovora na način koji odgovora gradjaninu i koja vodi računa o ličnim preferencama istog, tako da može da filtrira rezultate ADVANSE mašine, ako isti ne odgovaraju području intersovanja gradjanina, kao i da sugeriše novo pitanje na istu temu.

3. NORMALIZACIJA, ZAŠTO I KAKO?

Kao što je već ranije rečeno normalizacija, u kontekstu ovog rada, je postupak svodjenja teksta na skup reči od važnosti u osnovnom obliku. Normalizacija se upotrebljava kod rešavanja mnogih aplikativnih problema, kao što su klasterovanje i kategorizacija tekstualnih dokumenata [5], [6], obrada računarom generisanog teksta u govor [7], uredjeno pretraživanje Web-a [8], kao i u mnogim drugim aplikativnim područjima gde sadržaji dokumenata utiču na rezultate aplikacije, kao što su digitalne biblioteke, npr.

Bez obzira na područje primene, korišćene metode i algoritmi vrlo često su inspirisani složenošću jezika u kojem su dokumenti napisani. Konsultavanjem GoogleScholar servisa na upit "*text normalization*", medju preko 630.000 pronadjenih rezultata, mogu se pronaći radovi koji se odnose na metode i algoritme koji se koriste u grupama jezika, na vrlo specifične jezike, kao što su Vijetnamski ili Hindu, i sl., kao i na raščlanjivanje složenica u cilju normalizacije, npr. u germanskim jezicima. Broj radova koji bi mogli da se referenciraju na ovu temu prevazilazi obim i svrhu ovog rada, te ovde dajemo pregled samo nekih najvažnijih istraživanja, sa posebnim osvrtom na istraživanja normalizacije srpskog jezika.

Prvi radovi uopšte na temu automatskog pretraživanja teksta datiraju iz kasnih 1950-tih [9], [10]. Ovi radovi postavljaju osnove pretraživanja i odredjivanja korpusa dokumenata na osnovu frekvencije pojavljivanja ključnih reči, pri čemu se rečenica pojavljuje kao prvi nivo hijerarhije komunikacije ideja, kako je to Luhn tada nazvao. Kasnih 1980-tih Salton i Buckley [11] definišu metriku za odredjivanje nekog korpusa dokumenta na osnovu merenja tf (term frequency) i idf (inverse document frequency), uključujući i normalizaciju kao preproces za brže i tačnije odredjivanje frekventnih termina, istovremeno ne narušavajući strukturu i sadržaj izvornih dokumenata. Pregled tehnika za automatsko pretraživanje teksta može se naći u [12], [13].

Ako se skoncentrišemo samo na probleme, pravila i tehnike normalizacije mogu se uočiti neke zajedničke odrednice, pre svega izuzetan uticaj konkretnog jezika na same algoritme normalizicije i organizaciju odgovarajućih struktura podataka. Na primer, ako se posmatra elizija (izbacivanje nepotrebnih samoglasnika), jedno interesantno istraživanje na temu normalizacije a da prethodno nije poznat jezik dokumenata [14], pokazuje da u se u francuskom i italijanskom jeziku to obično radi kada su u pitanju zamenice ili članovi ispred reči, npr. *l'avion*, dok se u engleskom jeziku član obično zamenjuje početak druge reči, npr. *they'are*.

Ako se vratimo na Srpski jezik možemo zaključiti da isti ima složen glagolski i imenski sistem. U morfologiji srpskog jezika razlikujemo tri roda, sedam padeža kao i jedninu i množinu. Imenice (i pridevi) se dele u tri klase [15] po deklinaciji (promeni po padežima) u zavisnosti od toga kojeg je roda imenica, kojeg je broja, da li se nominativ imenice završava na –o, -e, -a ili se završava suglasnikom. Shodno tome imamo nastavke za različite oblike (-a, -u, -om, -e, -i, -o, -ima, -ama). Ovi i ostali oblici nastavaka su korišćeni za pronalaženje osnove reči u našem algoritmu [16]. Npr. žena, žene, ženi, ženu, ženo, ženom.

Glagoli se menjaju po sledećim gramatičkim kategorijama: lice (zavisno od toga da li se radnja pripisuje govornom ili nekom drugom biću), npr. radim, radiš, radi; broj (jednina i množina), npr. radim, rade; rod (muški, ženski, srednji), npr. radio, radila, radilo; vreme (prezent, perfekat, futur, aorist...), npr. radim, radio (sam), radiću, radih. Pored ovih „otežavajućih okolnosti", u srpskom jeziku postoje i glasovne promene, koje se čak javljaju i prilikom deklinacije imenica. Rezultat glasovnih promena je pojava nekog drugog glasa u rečima na mestu nekog glasa koji se logički očekuje da se tu pojavi. Npr. učenik – učenici (umesto učeniki). Glasovnim promenama bavićemo se u daljim istraživanjima.

Normalizacija teksta je dosta širok pojam koji govori o tome da se tekst pretvori u neki drugi oblik koji je pogodan za neku vrstu računarske obrade. U ovom slučaju oblik treba biti pogodan za potrebe pretraživanja. Kad je srpski jezik u pitanju mogu se istaći dve grupe istraživanja. Jedna se odnosi na AlfaNum sistem za pretvaranje teksta u govor [17], dok se druga grupa istraživanja odnosi na izradu srpskog WordNet-a, morfološkog rečnika srpskog jezika [18], [19].

Upravo zbog napred navedenih osobenosti našeg jezika, normalizacija teksta na srpskom jeziku predstavlja pravi izazov kada je u pitanju njena složenost. U engleskom jeziku padežni i glagolski oblici se najčešće dobijaju dodavanjem druge reči (predloga, pomoćnog glagola), i eventualno postoje nepravilni glagoli, dok je broj sufiksa (nastavak koji se dodaje na kraju reči) neuporedivo manji nego u srpskom jeziku. Ukoliko ne bismo posvetili pažnju normalizaciji tekstova napisanih na srpskom jeziku, desio bi se sledeći scenario. Pretpostavimo da je korisnik našeg naprednog sistema e-Uprave potražio odgovor na pitanje:

„Koja su mi dokumenta potrebna za vozačku dozvolu?“

Potpuni apsurd bi bio da se kao odgovor samo razmatraju dokumenta u kojima se ključni nosilac značenja ovog pitanja **vožačka dozvola** nalazi samo u akuzativu jednine, ili da se daje mnogo manji faktor značenja obliku **vožačka dozvola**, u odnosu na **vozačku dozvolu**, kada imaju identično semantičko značenje.

Proces normalizacije za potrebe pretraživanja ima sledeće zadatke: (1) izbacivanje reči bez veće sadržajne vrednosti – zaustavnih reči; (2) svodjenje različitih oblika iste reči na zajednički oblik. Ovako pripremljen tekst je optimalan za pretraživanje. Prva iteracija ovog istraživanja na temu problema normalizacije pitanja na srpskom jeziku predstavljala je pokušaje nalaženja resursa u elektronskom obliku koji bi olakšali rešavanje ovog problema. Jedan od takvih resursa bio bi morfološki rečnik, koji bi oba zadatka normalizacije znatno olakšao. Kako takav rečnik nismo uspeli da pronađemo u elektronskom onbliku, bili smo primorani da sve reči od značaja posmatramo na isti način, bez obzira kojoj vrsti reči pripadaju.

Takođe, ne postoji ni zabeležen skup zaustavnih reči u srpskom jeziku, pa je, za potrebe ovog istraživanja, napravljen sopstveni skup zaustavnih reči. To je urađeno prevođenjem skupa zaustavnih reči sa engleskog jezika [15], pri čemo su promenljive reči izmenjane po padežima i dodati sinonimi, te pomoćni i modalni glagoli u odgovarajućim glagolskim vremenima. Takodje, napravljen je skup nastavaka za padeže u jednini i množini, kao i za glagolska vremena, koristeći literaturu datu u referencama [16].

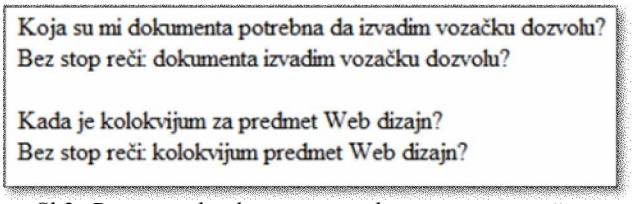
U nastavku rada opisani su algoritmi za normalizaciju, u slučaju ne razmatranja glasovnih promena. Kada se govori o drugom zadatku normalizacije, svođenju na zajednički oblik, pod tim smo podrazumevali osnovu imenice (prideva) i infinitivnu osnovu glagola. Osnova imenice se dobija odbijanjem nastavka genitiva jednine, a infinitivna osnova glagola odbijanjem nastavka –ti, ukoliko se ispred njega nalazi samoglasnik, ili odbijanjem nastavka za 1. lice aorista jednine u ostalim slučajevima. Npr. vožačka – vozačk, raditi – radi. Kako mi ne znamo o kojoj se vrsti reči radi kada imamo kao ulaz samo tekstualni dokument, osnovna ideja odnosi se na poređenje kraja reči sa sufiksima sakupljenim u bazi podataka, i nalaženja što većeg poklapanja. Da bismo proverili tačnost algoritma, ručno smo našli osnove na skupu od 1000 različitih reči (odgovora na teme foruma policijske akademije) i na isti skup primenili algoritam i našli procenat tačnosti.

4. OPIS ALGORITMA

Prvi algoritam, Slika 2, uzima pitanje i vraća pročišćeno pitanje izbacujući stop reči. Izlaz koji se dobije se sastoji od reči od važnosti u izvornom obliku (u odgovarajućem padežnom ili glagolskom obliku, onako kako se nalaze u pitanju). Složenost algoritma je kvadratna: O(nw * nsw), gde su nw i nsw broj reči u pitanju i broj stop reči u pitanju, respektivno.

```
connect to Db
select question asked
q= question= array of words
nw= number_of_words(q)
select stop_words from db
sw=array of stop_words
nsw= number_of_words(sw)
for i= 1 to nw
{
  for j= 1 to nsw
    {
        if q[i] != sw[j]
        }
clean= clean + q[i]
}
return clean
```

Sl.2. *Algoritam za izbacivanje stop reči iz pitanja*



Sl.3. *Primer rada algoritma za izbacivanje stop reči iz pitanja*

Drugi algoritam, Slika 4, uzima rezultat prvog algoritma (reči iz pitanja u izvornom obliku) i vraća osnovu reči. Za svaku reč traži se poklapanje sa što dužim sufiksom iz spiska sufiksa i ostatak reči se proglašava osnovom. U kombinaciji sa algoritmom koji u nizu ponavljajućih vrednosti uzima onu koja se najviše ponavlja, ovaj algoritam može se iskoristi kada se kao ulaz da skup od više oblika jedne reči. Tada, u slučaju većeg broja poklapanja, ostatak reči sa najvećim brojem pojavljivanja proglašava se osnovom. Složenost ovog algoritma je kubna: O(nw * max * ns), gde su nw, max i ns broj reči iz pročišćenog pitanja, maksimalni broj slova u reči, i broj sufiksa, respektivno.

```
begin
words= select clean_question
suffixes= select all suffixes from db
max=number_of_letters(the_longest_suffix)
nw= number_of_word(words)
ns= number_of_suffixes(suffixes)
a_words = array of words
a_suffixes = array of suffixes
for i=0 to nw-1
 {
        word_base[i]=a_words[i]
  for k=1 to max
   {
   if (k<length(a_words[i]))
    {
        word_suffix[k-1]=last_k_letters(a_words[i])
        for l=1 to ns
     {
         if word_suffix[k-1] == a_suffixes[l]
         word_base[i]= first_(number_of_letters(a_words[i])-k)_letters(a_words[i])
        }
    }
   }
  }
end
```

Sl. 4. *Algoritam za svođenje reči iz pitanja na osnovu reči*

Ako rezultati primene algoritma za izbacivanje stop reči nad pitanjima datim na Slici 2., budu ulaz za algoritam za svođenje reči na osnovu, dobija se rezultat prikazan na Sl.5.

Sl.5. *Primer rada algoritma za svođenje reči iz pitanja na osnovu reči*

Nakon implementacije algoritma, proverena je i njegova tačnost na skupu od 1000 reči. Razlog provere na ovako skromnom skupu reči je zahtevnost i vreme koju iziskuje ručno pronalaženje gramatičkih osnova različitih reči, usled brojnih pravila. Ne radi se o bilo kom skupu reči, već su one uzete iz pitanja i odgovora na forumu policijske akademije, odnosno deo su terminologije jedne oblasti, što će biti slučaj i u sistemu e-Uprave. Tom prilikom dobijeni su sledeći rezultati:

Tab.1. *Rezultati provere tačnosti rada algoritma za svođenje reči iz pitanja na osnovu*

Broj reči	*Broj pogodaka*	*Broj grešaka*	*Procenat pogodaka*	*Procenat grešaka*
1000	887	113	88,7%	11,3%

Visoki procenat tačnosti pripisujemo činjenici da se radi o skupu reči čiji najveći deo čine termini jedne oblasti, obično poreklom iz drugih jezika, pa i ne podležu glasovnim promenama (npr. topografija, kolokvijum, akademija i sl.). Prilikom analize ovih rezultata, uočeno je da su osnovni uzroci greške glasovne promene i poklapanje zadnjeg dela reči u nominativu sa nekim od sufiksa.

5. ZAKLJUČAK

U postavljanju osnova algoritma normalizacije sve reči smo posmatrali na isti način (imenice, glagoli i pridevi). U budućnosti ćemo težiti da razvijemo posebne algoritme za normalizaciju svake vrste reči od važnosti (imenice, glagoli i pridevi). Pratićemo efikasnost algoritama opisanih u ovom radu uporedo sa porastom baze sufiksa i stop reči. Pošto je složenost ovih algoritama polinomna, može se očekivati usporavanje procesa normalizacije sa porastom veličine ove baze, te ćemo raditi na optimizaciji ovih algoritama, kao i na istraživanju novih algoritama. Trenutne strukture ovih podataka, zbog malog uzorka, su serijske tabele uredjene po redosljedu otkrivanja stop reči i sufiksa. Nakon povećanja uzorka istražiće se mogućnost da se iste urede u hash tabelu, a nakon njihove stabilizacije da se poronadje perfektna hash funkcija. Naredni korak je rešavanje problema glasovnih promena.

ZAHVALNICA

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

LITERATURA

[1] G. Šimić, Z. Jeremić, E. Kajan, D. Randjelović. "A Framework for Delivering e-Government Support", Submitted for publication, 2012.

[2] G. Šimić, E. Kajan, Z. Jeremić, D. Randjelović. "An Approach to Document Clustering using Hybird Method", IADIS e-society conference, Berlin, March, 9-13, 2012, pp. 153-159.

[3] U. Marovac, A. Pljasković, E.Kajan, "Applying native xml databases in advanced e-government systems", *ICIST*, Kopaonik, March, 1-4, 2012.

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

[5] W. Zhang, T. Yoshida, X. Tang, Q. Wang, "Text clustering using frequent item sets", *Knowledge-Based Systems*, 23 (2010), pp. 379-388.

[6] M. Radovanović and M. Ivanović, "Interaction Between Document Represenatation and Feature Selection in Text Categorization", *DEXA 2006*, LNCS 4080, pp. 489-498, 2006.

[7] M. Adda-Decker, G. Adda and L. Lamel, "Investigating text normalization and pronunciations variants for German broadcast transcription", *ICSLP 2000*, Vol. I, pp. 266-269.

[8] C. Carpineto, S. Osinski, G. Romano, D. Weis, "A Survey of Web Clustering Engines". *ACM Computing Surveys*, Vol. 41, No. 3, Art. 17, 2009.

[9] H. P. Luhn, "A Statistical Approach to the Mechanized Encoding and Searching of Literature Information", *IBM Journal of Research and Development*, 1(4), 1957, pp. 309-317.

[10] H. P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, 2(2), 1985, pp. 159-165.

[11] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval", *Information Processing & Management*, 24(5), 1988, pp. 513-523.

[12] F. Sebastiani, "Machine learning in automated text categorization", *ACM Computing Surveys*, 34(1), pp. 1-47.

[13] A. Stavrianou, P. Andistos, N. Nicoloyanis, "Overview and semantic issues of text mining", *ACM SIGMOD Record*, 36(3), 2007, pp.23-24.

[14] E. Giguet, "The stakes of multilinguality: Multilingual Text Tokenization in Natural Language Diagnosis", *Proceedings of PRICAI Workshop*, 1996.

[15] http://www.ranks.nl/resources/stopwords.html

[16] http://sr.wikisource.org/sr/Српска_граматика_(Ж._Симић)/Поглавље_4.

[17] M. Sečujski, "Akcentski rečnik srpskog jezika namenjen sintezi govora na osnovu teksta", *DOGS*, Bečej, 2002, str. 17-20.

[18] V. Satev, N. Nikolov, "Using the Web as a corpus for extracting abbreviations in the Serbian language", *Proc. of the 6th Language Technology Conference*, October 2008, Ljubljana, Slovenia, pp. 75-79.

[19] C. Krstev, D. Vitas, A. Savary, "Preriquestes for a Comprehensive Dictionary of Serbian Compounds", *LNAI 4139*, 2006, pp. 552-563.

Abstract – This paper analyses challenges that arise with dealing with unique characteristics of Serbian language during application development aiming to efficient searching of text documents. The emphasis has given to normalization methods and algorithms in order to prepare documents for further computer processing, in this specific case, searching and clustering.

NORMALIZATION OF TEXT DOCUMENTS IN SERBIAN LANGUAGE FOR EFICIENT SEARCHING IN E-GOVERNMENT SYSTEMS

Ejub Kajan, Aldina Pljasković, Adela Crnišanin