MODELOVANJE I PRETRAŽIVANJE NAD NESTRUKTUIRANIM PODACIMA I DOKUMENTIMA U E-UPRAVI REPUBLIKE SRBIJE

dr Vojkan Nikolić

Sadržaj

1. Uvod
2. e-Uprava Republike Srbije
3. Question answering sistemi
4. Apache Lucene
5. Modelovanje sistema za dobijanje brzih odgovora za servise e-Uprave Republike Srbije u oblasti Krivičnog zakonika
6. Analiza eksperimentalnih rezultata
7. Zaključak

1. Uvod

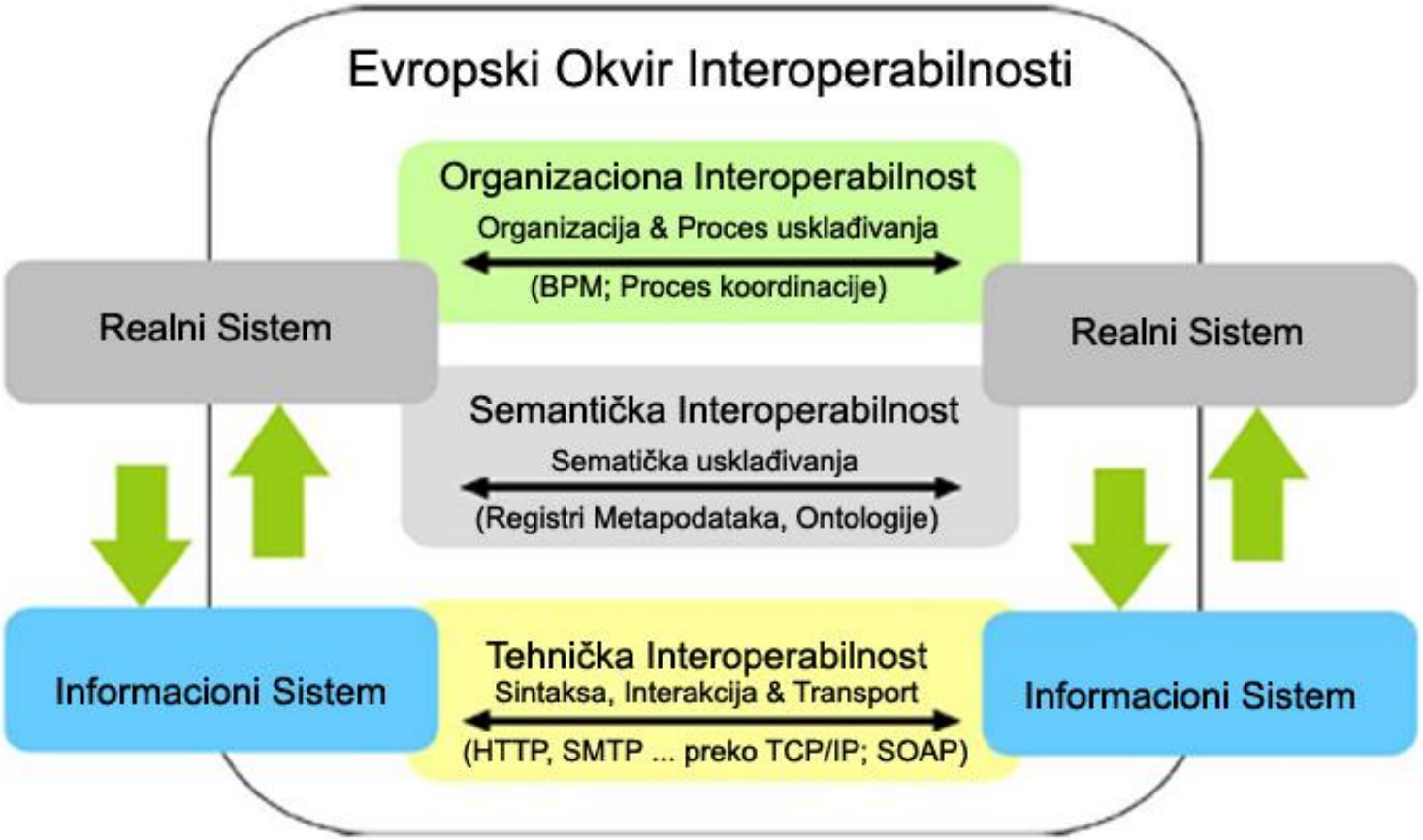
- Veb servisi e-Uprave
- Nestruktuirani podaci
- Dokumena na srpskom jeziku
- Question answer sistemi
- Bag of Words
- Bag of Concept
- Stvaranju koncepta zasnovanog na tekstualnoj prezantaciji i primeni kategorizacije teksta

2. e-Uprava Republike Srbije

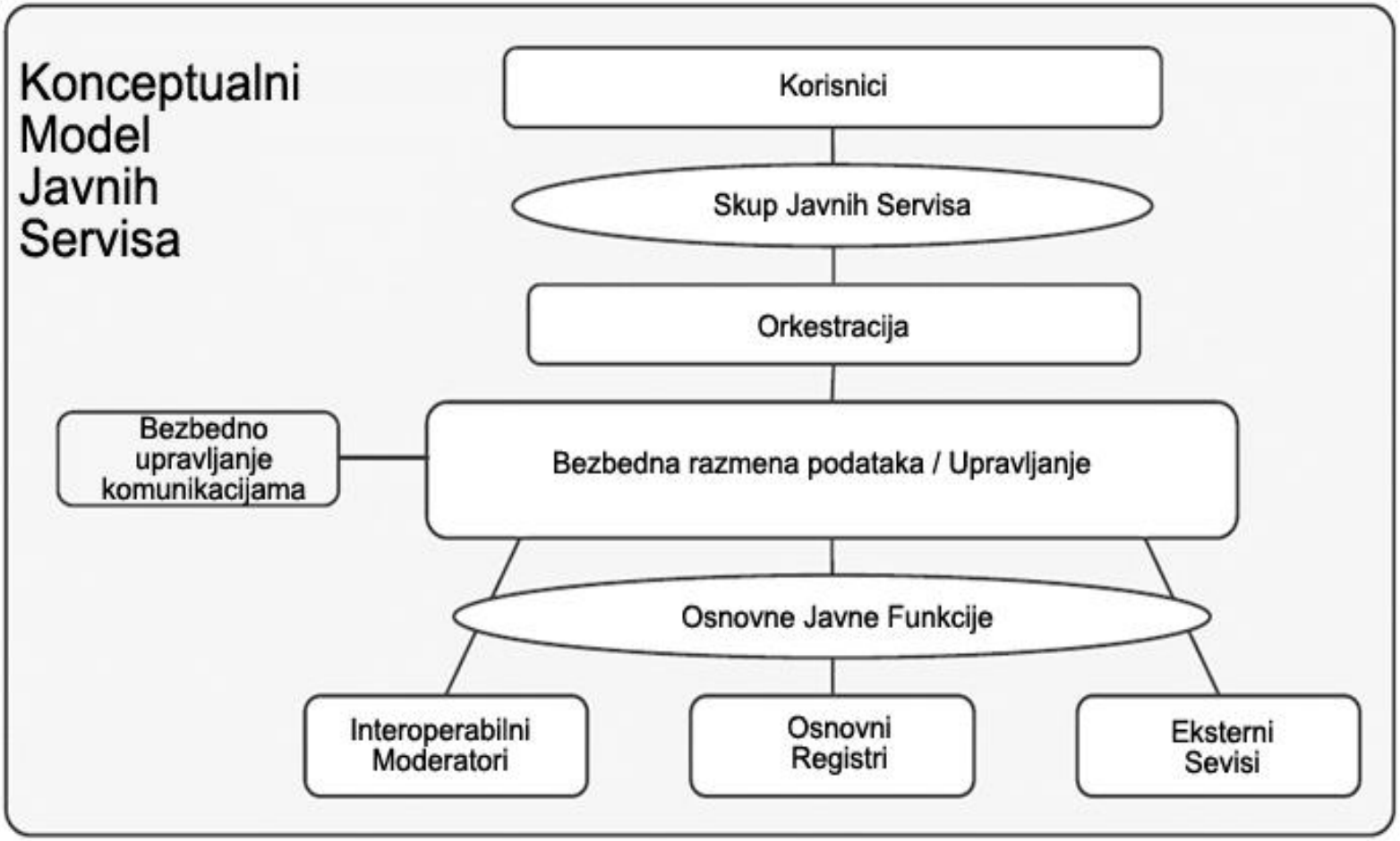
Elektronska uprava se zasniva na digitalnim interakcijama između:

- Vlade određene zemlje i građana (*G2C*),
- Vlade i privrednih subjekata (*G2B)*,
- Vlade i zaposlenih u javnoj upravi (*G2E*).
- Vlade i Vlada i agencija drugih zemalja (*G2G*)

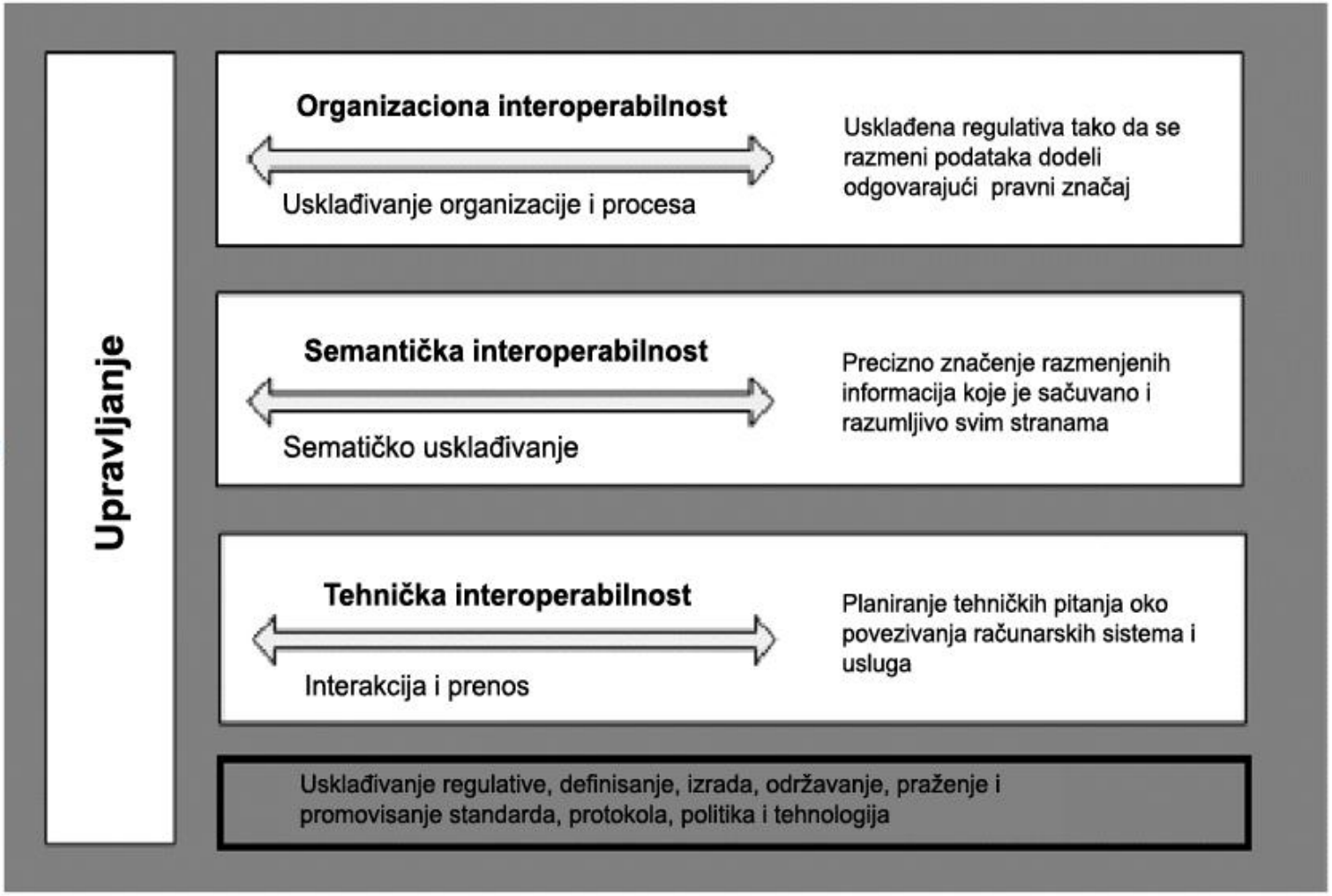
Interoperabilnost e-Uprave u kontekstu Evropskog okvira interoperabilnosti



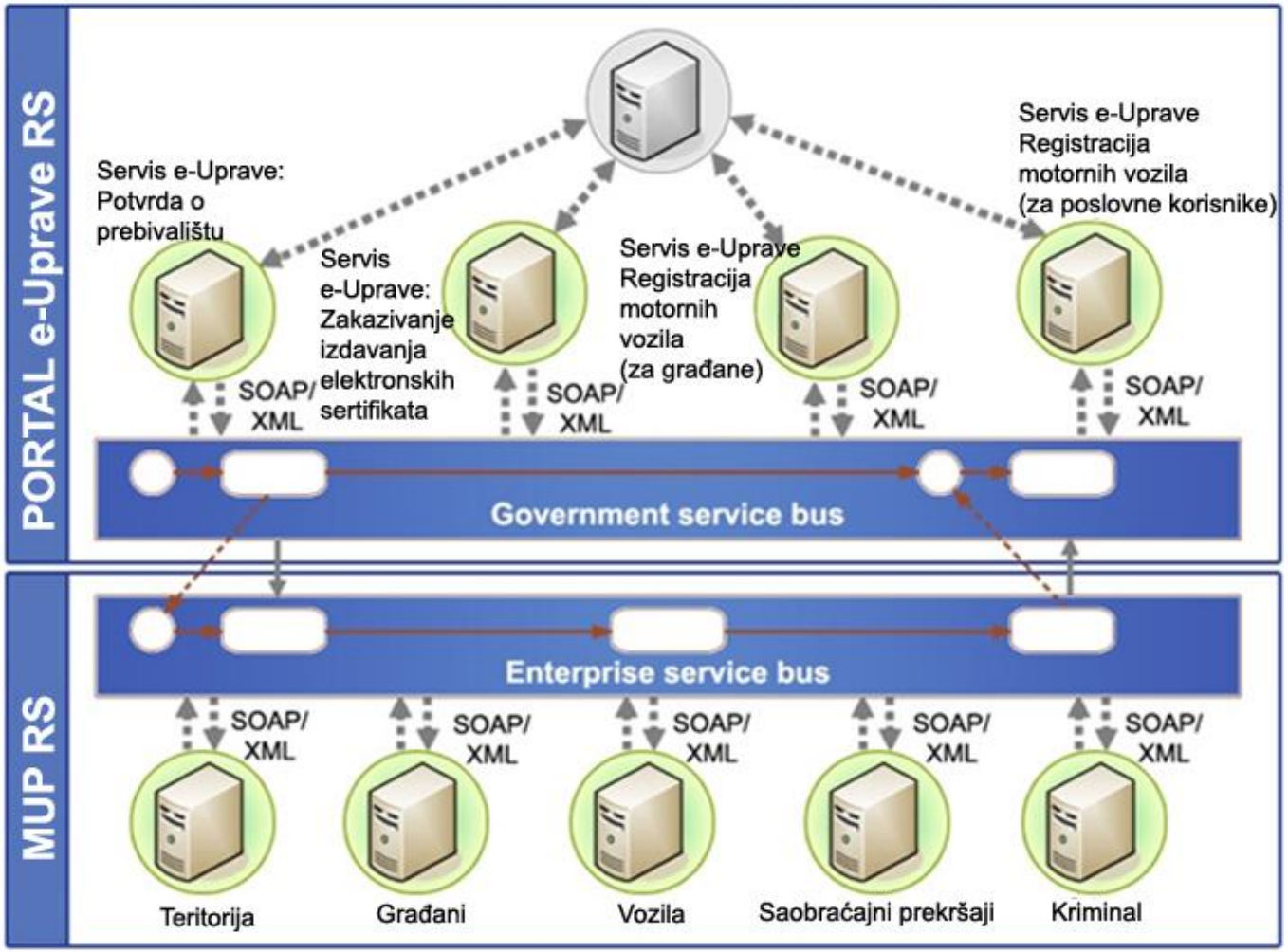
Konceptualni model EIF



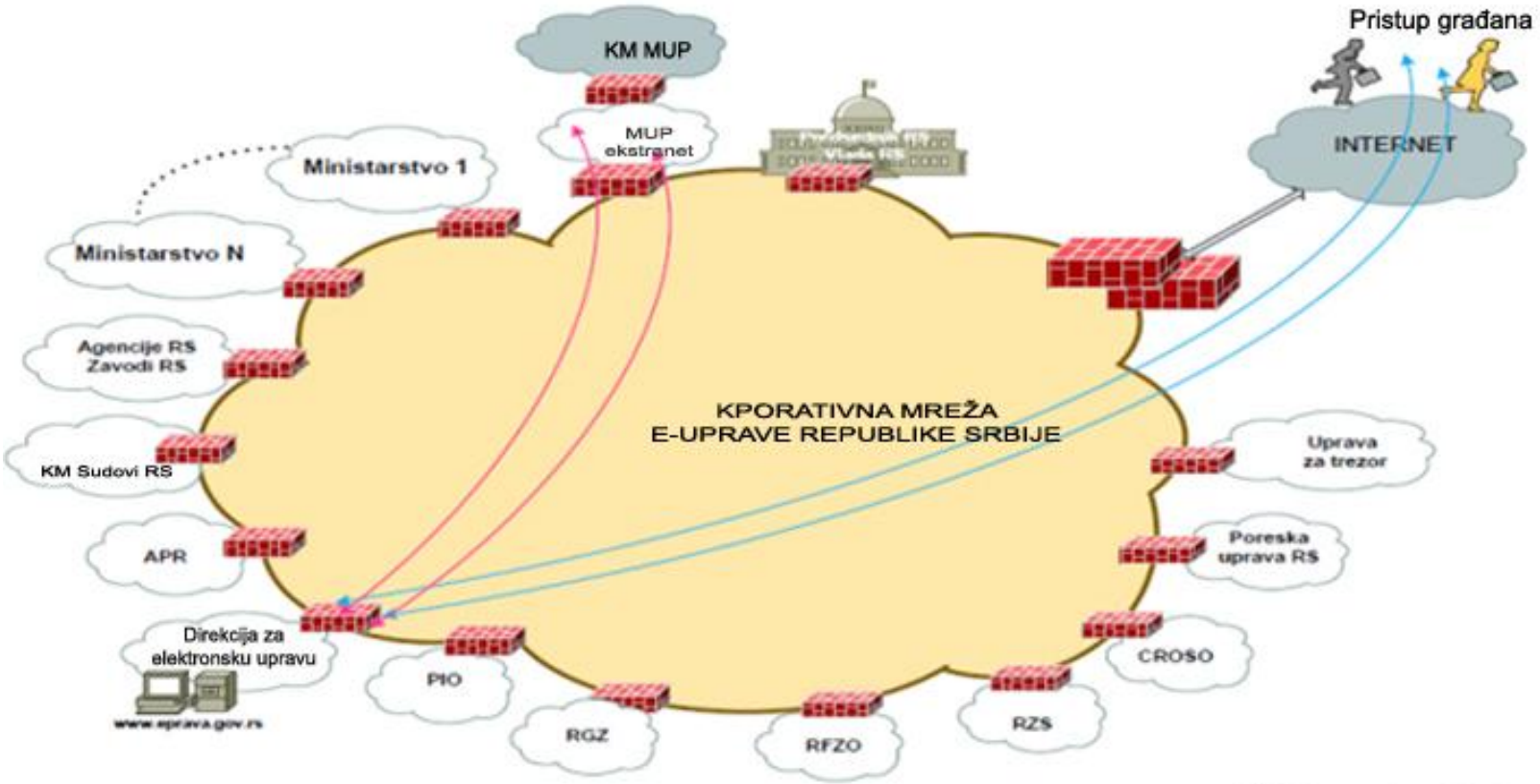
Nacionalni okvir interoperabilnosti Republike Srbije i Servisno orijentisana arhitektura



Realizacija interoperabilnosti pomoću GSB u e-Upravi Republike Srbije

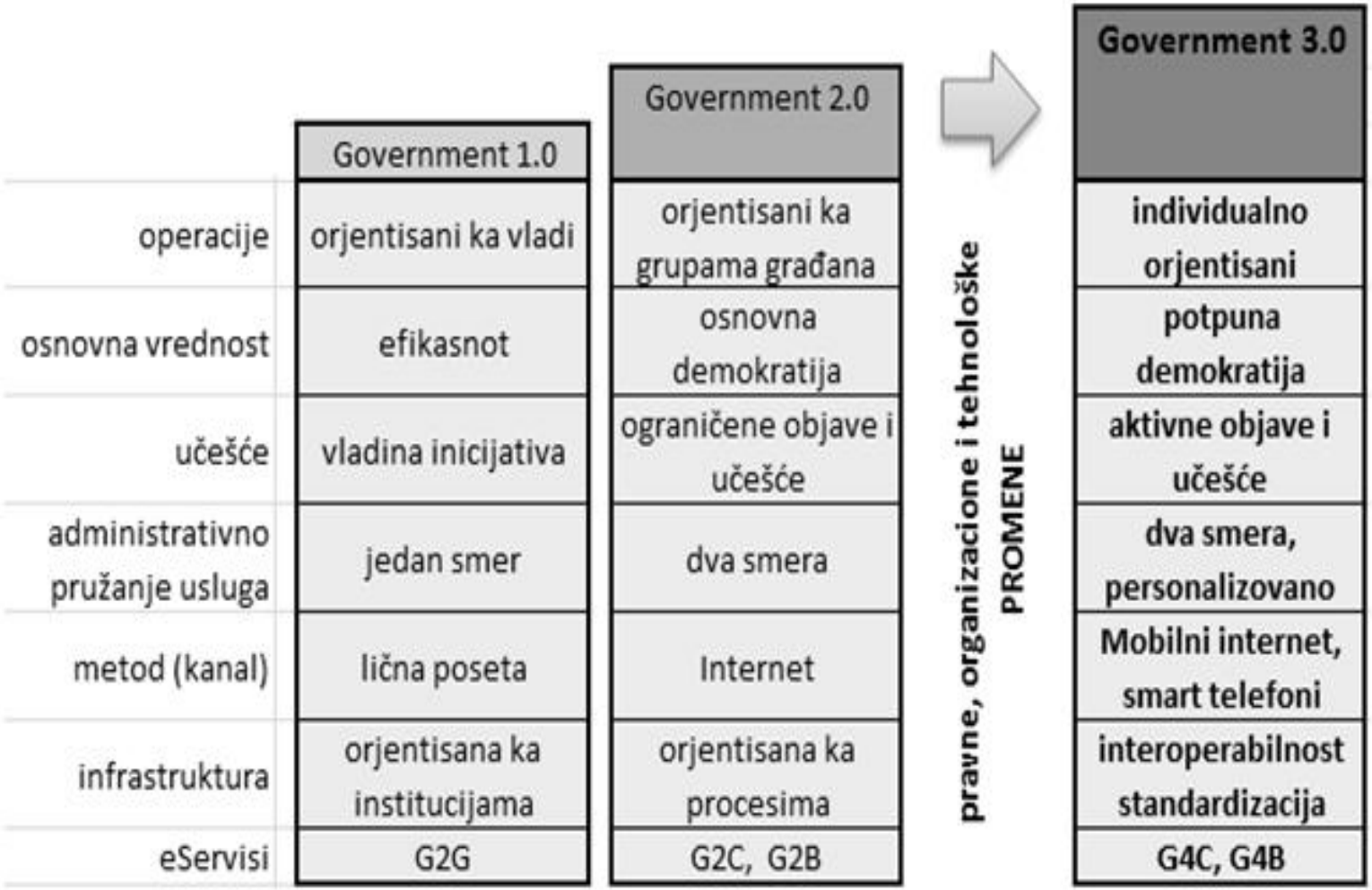


Portal e-Uprava u okviru MDO



KM - Korporativna Mreža

Evolucija koncepta e-Uprave



3. Question answering sistemi

- Obrađuju upite postavljene na prirodnom jeziku i vraćaju ili ekstrahuju odgovor iz struktuiranih (baze podataka) ili nestrukturiranih (tekstualnih) izvora;
- Sistem treba da prepozna tip odgovora koji korisnik očekuje;
- Složenost ovih sistema leži u uspostavljanju podrazumevanih (implicitnih) odnosa između upita i odgovora;
- Fokusirani na davanje kratkih odgovora u vidu polu-informacije, definicije ili vremenske odrednice na postavljeni upit.

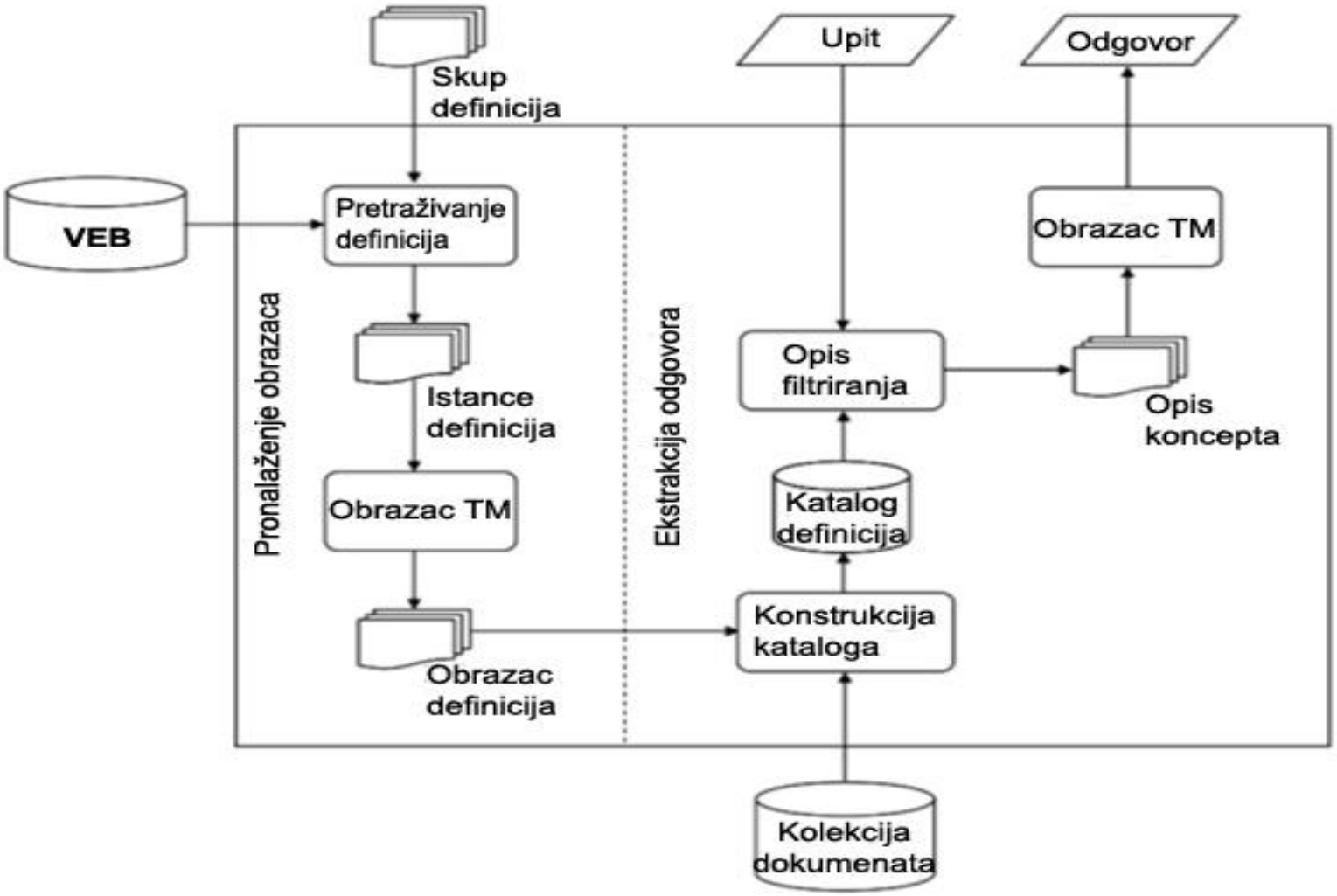
Prvi pristup u arhitekturi razvoja QA sistema

- **Klasifikator upita** - identifikovanje vrste upita (npr. šta, gde);
- **Dokument-Odgovor** - odgovoran za traženje i identifikaciju relevantnih dokumenata;
- **Ekstrahovani kandidat odgovora** - identifikuje potencialni odgovor pronađen u relevantnim izvoru;
- **Selektovanje odgovora** - generiše odgovor na upit .

Drugi pristup u razvoju QA sistema

- **Komponenta za obradu upita** - vrši tokenizaciju i tagovanje, identifikaciju ključnih reči, gramatičku analizu upita, identifikaciju dvosmislenih reči, identifikaciju tipa očekivanih odgovora i proširivanje značenja ključnih reči;
- **Komponenta za pretragu** generiše upite i vrši pretraživanje izvora podataka dostupnih na veb-u;
- **Komponenta za ekstrakciju odgovora** filtrira podatke koje je pronašla komponenta za pretragu, prepoznaje entitet, identifikikuje odgovor i vrši proveru ispravnosti.

Generalna šema metode



U ovoj formuli uvedena su sledeće oznake radi jednostavnosti:

Rangiranje rezultata pretrage (*eng. ranking score*)

Rezultat rangiranja *R* za sekvencu reči ukazuje na njegovu relevantnu frekvenciju i zračunava se na sledeći način:

$$R_{p(n)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n-i+1}\frac{f_{p_j(i)}}{\sum_{\forall q \in S_i} f_{q(i)}}$$

Si - skup sekvenci veličine *i,*

qi - predstavlja sekvencu *q* veličine *i,*

pj(i) - *j*-ta podsekvence veličine *i* koja je uključena u sekvencu,

fq(i) - učestalost pojava sekvence *q* u skupu opisa koncepta, tj. termina

Paketi otvorenog koda za TM analizu

(Sistem otvorenog koda)	Opis
Carrot2 http://project.carrot2.org	Carrot2 je internet pretraživač otvorenog koda baziran na mašinama za grupisanje rezultata pretrage. Može automatski da organizuje male zbirke dokumenata, npr. rezultate pretrage u tematske kategorije. Carrot2 nudi gotove komponente za pronalaženje rezultata pretraga iz različitih izvora, uključujući Google API, Bing API, eTools Meta Search, Lucene, SOLR, Google Desktop itd.
GATE http://gate.ac.uk	Softver otvorenog koda koji može da reši skoro svaki problem obrade teksta, sve vrste jezičkih obrada i aplikacija, uključujući glas korisnika: problem istraživanja raka, istraživanja droge, podršku pri odlučivanju, veb-mining, izdvajanje informacija, semantičke napomene. Mnoge obrazovne ustanove su već uključile GATE u svoje TM tehnike.
Natural Language Toolkit (NLTK) http://www.nltk.org	Skup biblioteka i programa za simboličku i statističku obradu NLP pomoću programskog jezika Python. NLTK je praćen brojnim struktuiranim tekstovima, pojednostavljenom gramatikom, obučenim modelima, itd. NLTK je pogodan za kurseve u mnogim oblastima, uključujući obradu prirodnog jezika, računarsku lingvistiku, empirijsku lingvistiku, kognitivne nauke, veštačku inteligenciju, pronalaženje informacija i mašinsko učenje.
RapidMiner http://rapid-i.com/content/view/181/190	Formalno *Yet Another Learning Environment* (YALE) je okruženje za mašinsko učenje, DM, TM, prediktivnu i poslovnu analitiku. *Plug-in* komponenta je specijalno dizajnirana da pripremi tekstualni dokument za analizu, kroz proces tokenizacije, izbacivanje stop-reči i strimovanje. Dodatne komponente *RapidMiner* su *Java* biblioteke koje treba dodatno instalirani u *lib\plugins* direktorijume.
Arhitektura za upravljanje nestruktuiranim informacijama (UIMA) http://uima.apache.org	Prvobitno razvijena od strane IBM-a. To je otvorena, industrijski snažna prilagodljiva i proširiva platforma za kreiranje, integrisanje i primenu rešenja za upravljanje nestrukturiranim informacijama kombinovanjem semantičke analize i komponenata pretrage. Cilj UIMA-a je da obezbedi temelj zajedničke saradnje između industrijske i akademske zajednice širom sveta i da ubrza razvoj onih tehnologija koje su ključne za otkrivanje vitalnog znanja prisutnog u sve obimnijim izvorima informacija.
Text Mining paket http://cran.r-project.org/web/packages/tm/index.html	Ovaj paket nudi funkcionalnost u upravljanju tekstualnim dokumentima, skraćuje proces upravljanja dokumentom i olakšava korišćenje heterogenih tekstualnih formata. Ovaj paket ima pozadinsku podršku zasnovanu na integrisanim podacima kako bi se minimazirali zahtevi za memorisanjem. Unapređeno upravljanje metapodacima se koristi za prikupljanje tekstualnih dokumenata kako bi se lakše koristili veliki (obogaćeni sa metapodacima) skupovi dokumenata.

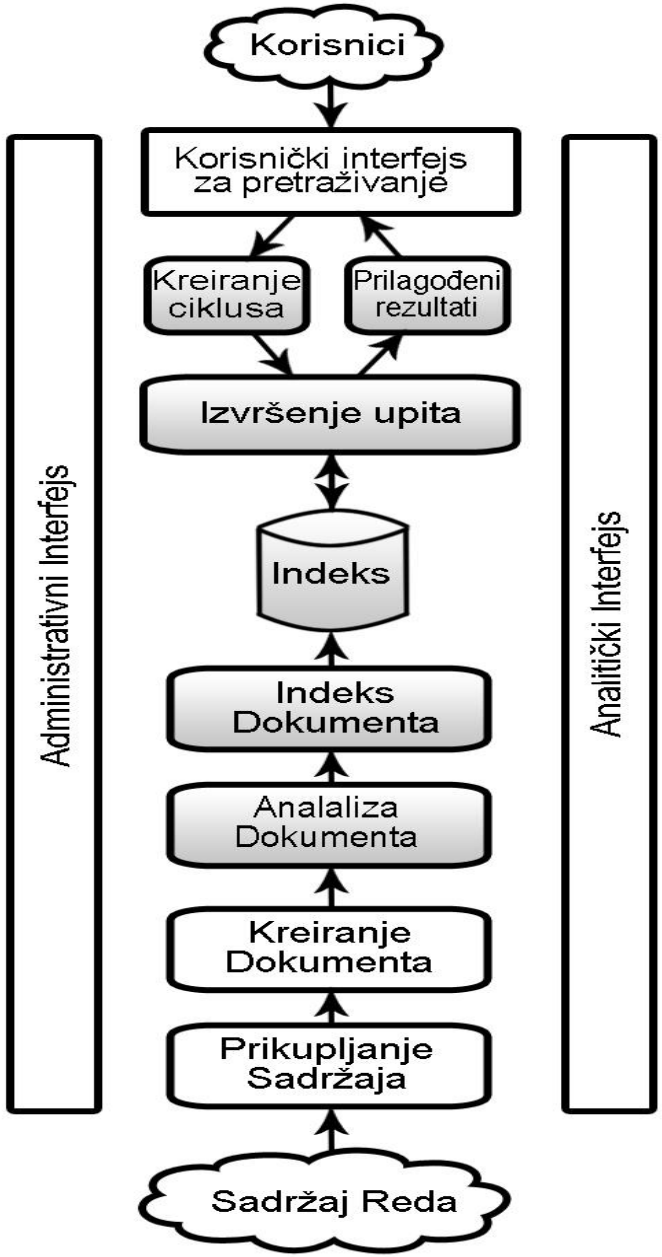
4. Apache Lucene

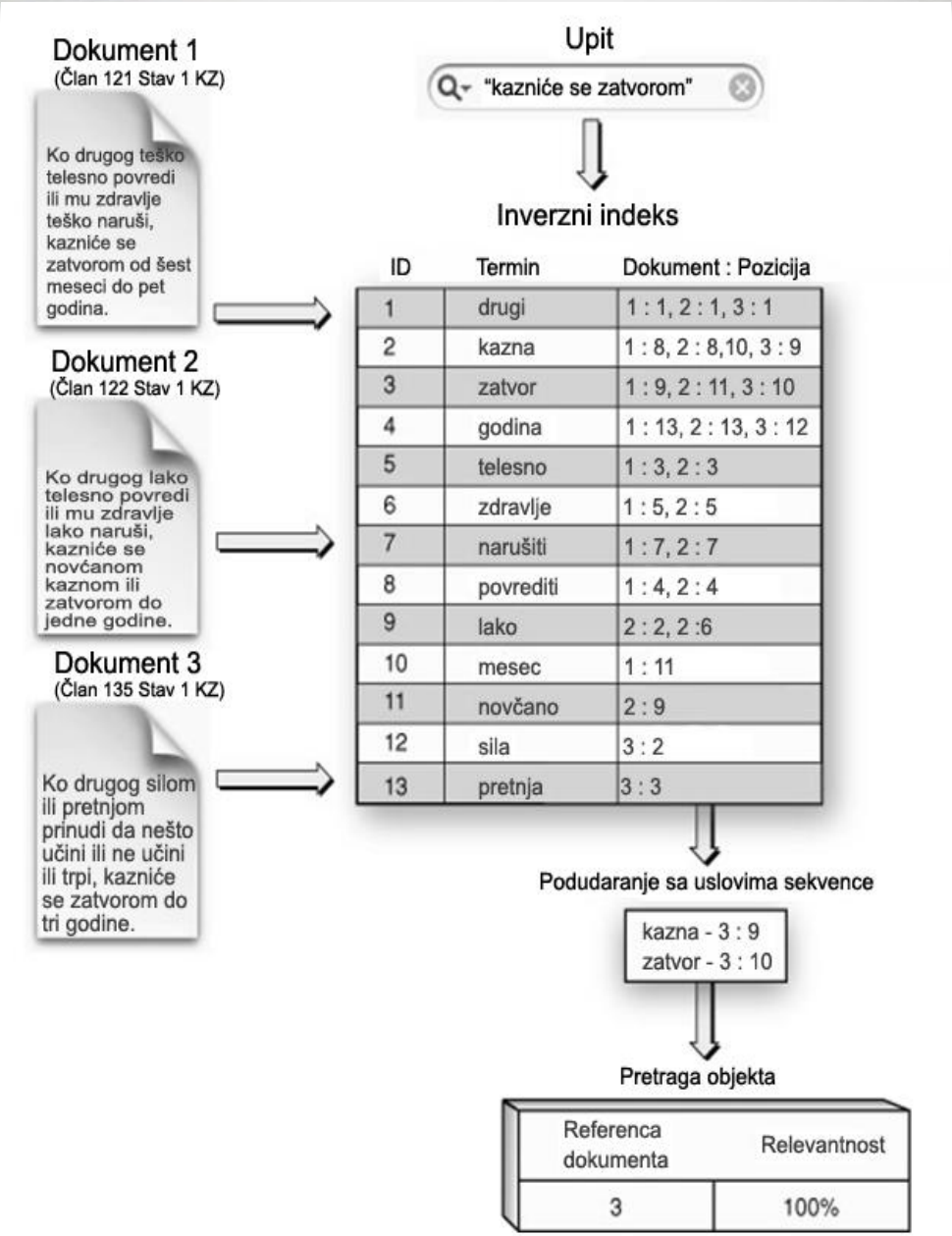
- Specijalizovana biblioteka za složena pretraživanja i sastoji se od veoma kompleksnih funkcija;
- Algoritmi i mehanizmi na kojima su funkcije bazirane su veoma složeni;
- *Lucene* mehanizam za indeksiranje
- *Lucene* mehanizam za pretraživanje

Arhitektura Apache Lucene

- Document
- Field
- Analyzer
- IndexWriter
- IndexSearcher
- QueryParser
- Query
- Hits

Glavne komponente aplikacije za pretraživanje





Koncept Apache Lucene

TF-IDF

TF:

$$TF(t, d_i) = \frac{n_{t,i}}{\sum_{k=1}^{|T|} n_{k,i}}$$

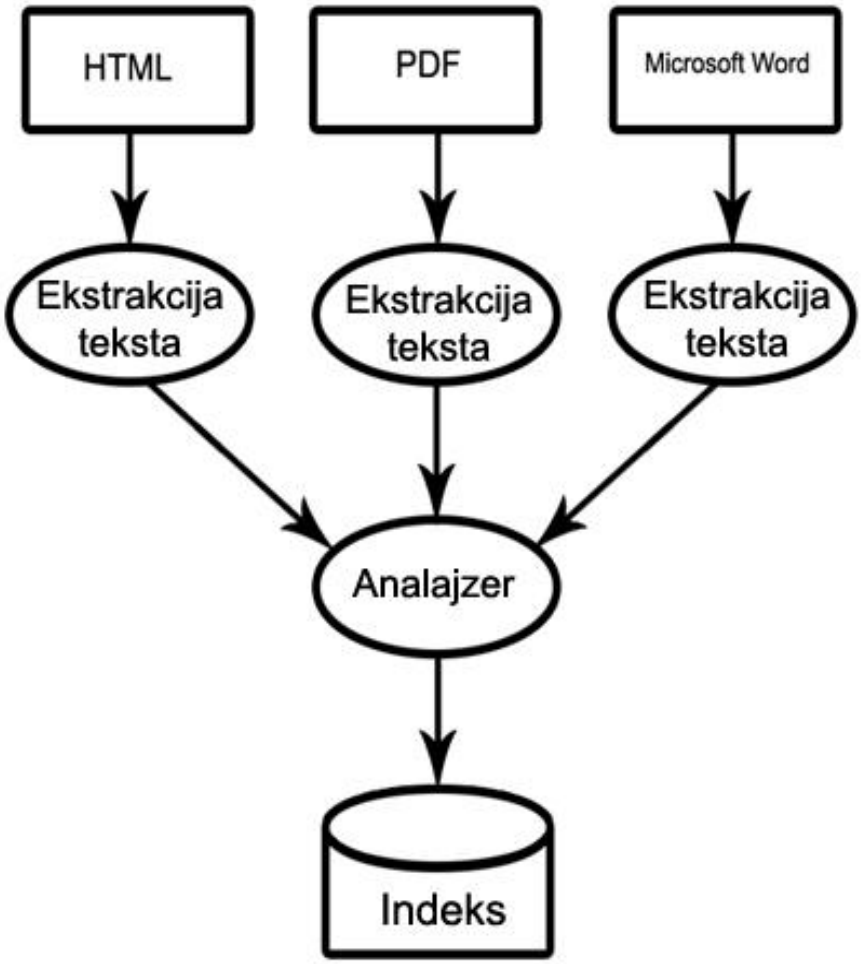
IDF:

$$IDF_t = \log \frac{M}{m_t + 0.01}$$

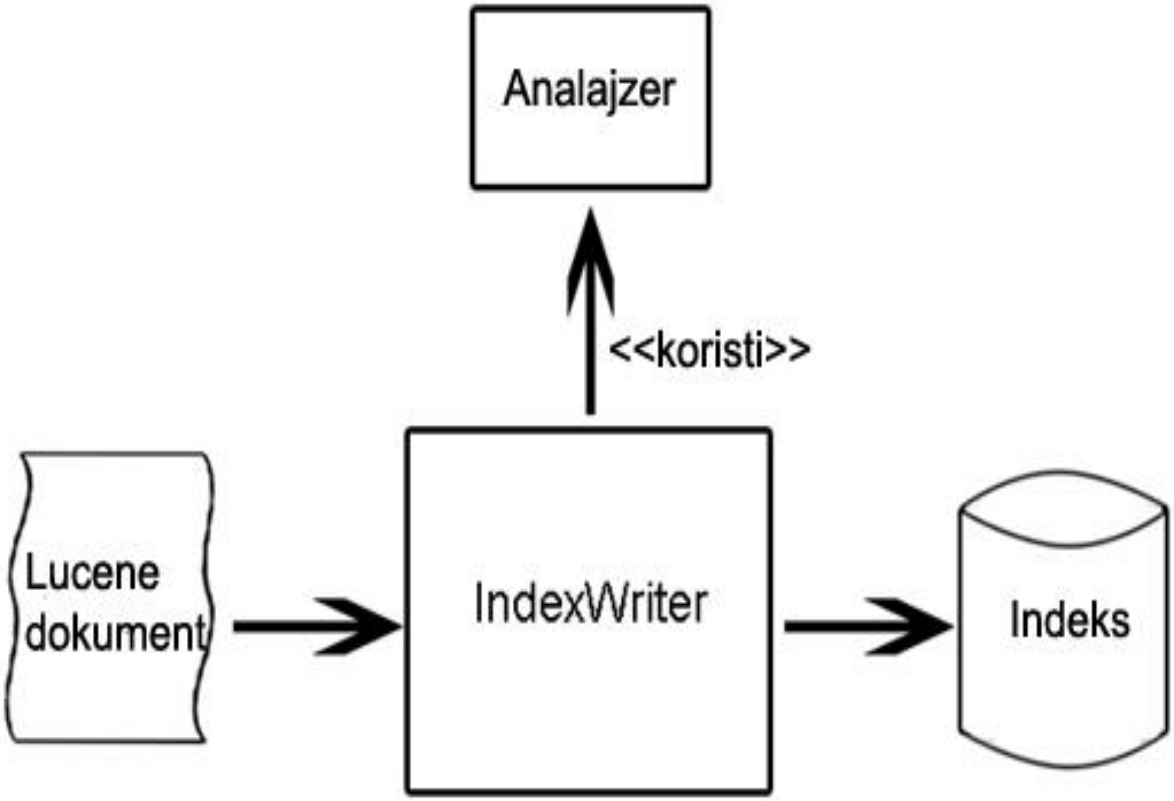
TF-IDF:

$$w(t, d_i) = TF(t, d_i) \times IDF_t$$

Proces indeksiranja



Lucene indeksiranje

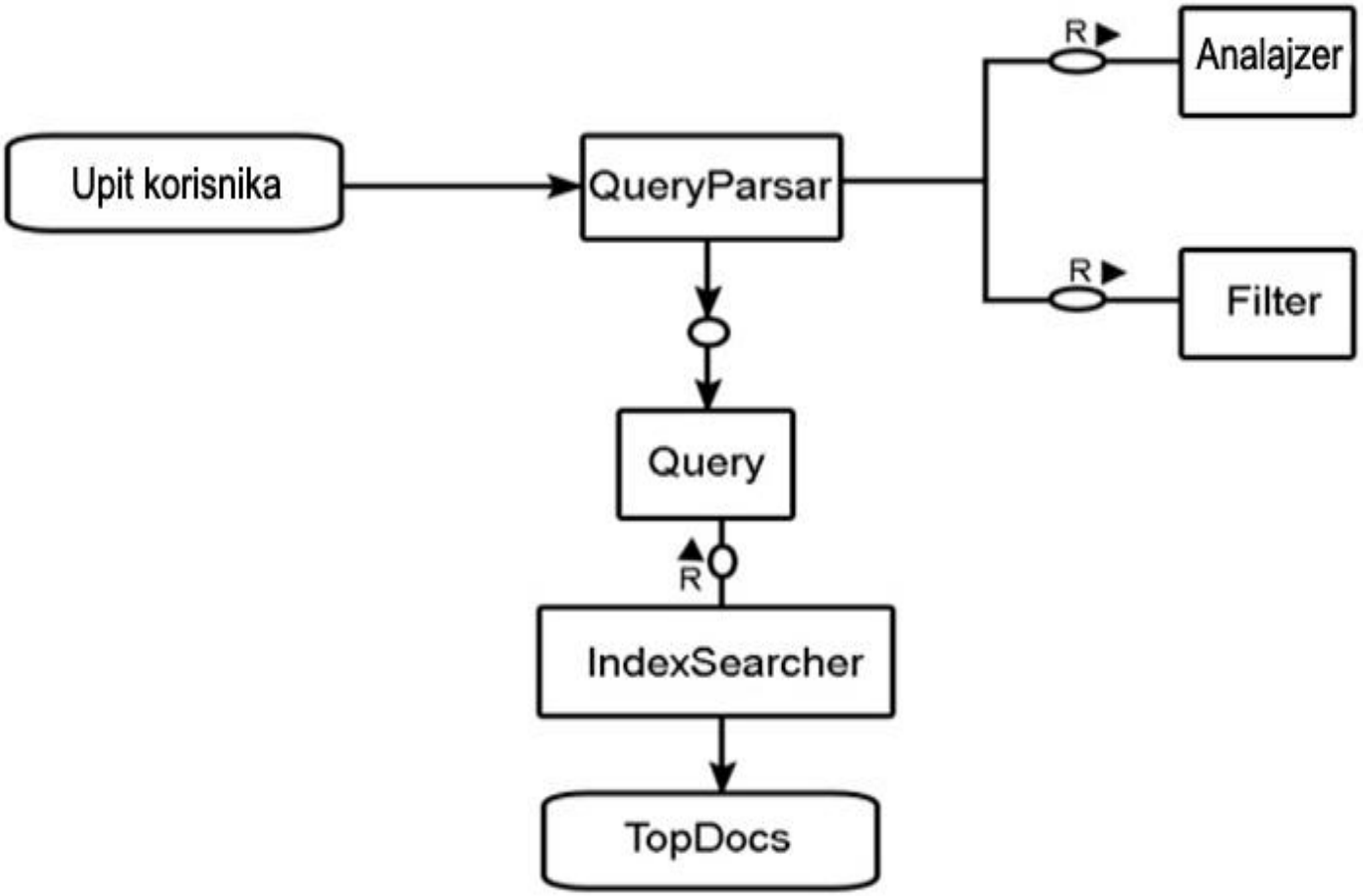


Analizer

public final class **StandardAnalyzer**extends
StopwordAnalyzerBase

public class SerbianNormalizationFilterFactory
extends TokenFilterFactory
implements MultiTermAwareComponent
Factory for SerbianNormalizationFilter

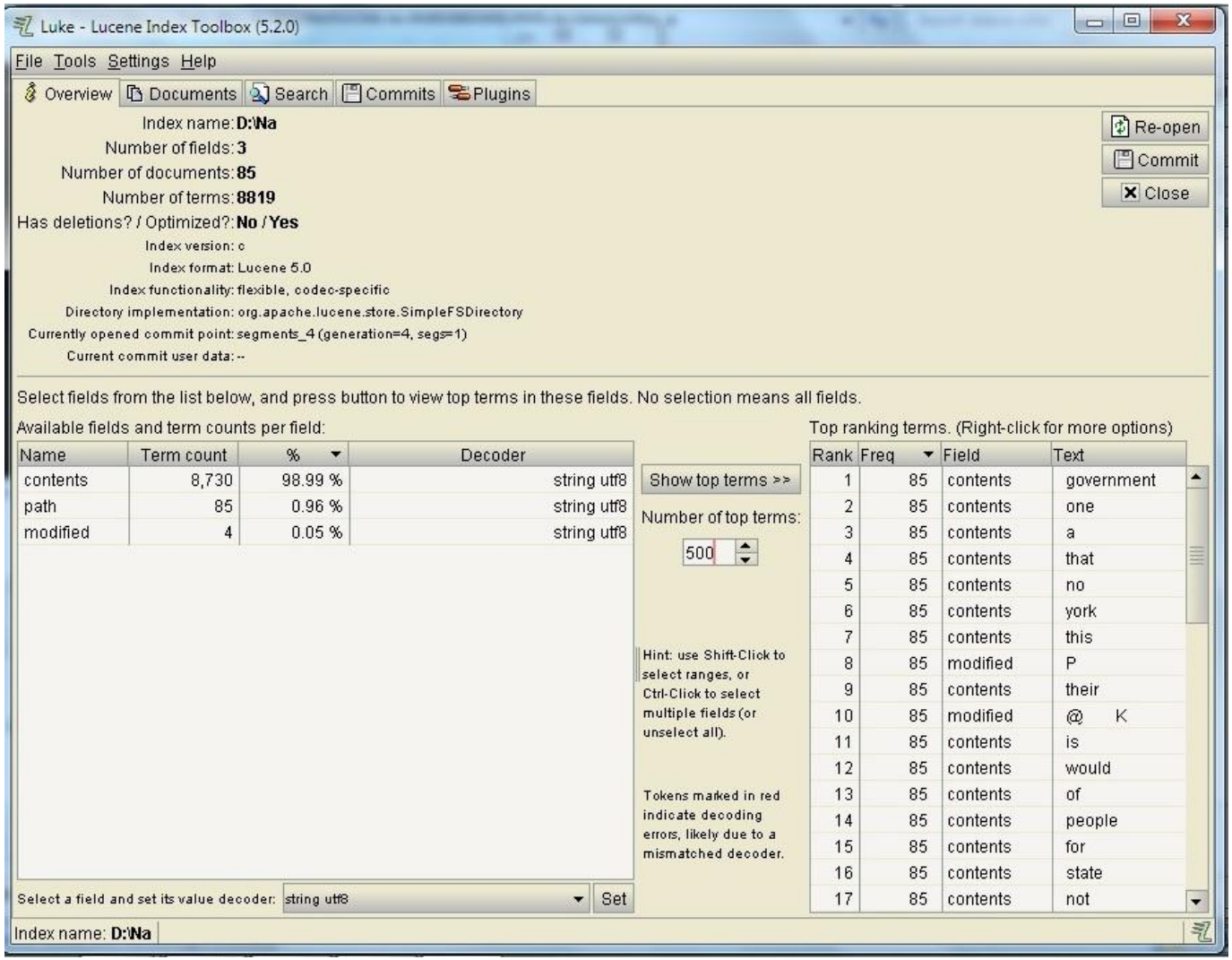
Lucene pretraživanje



LUKE

- pretraživanje po broju dokumenata ili termina,
- prikazivanje dokumenta,
- preuzimanje rang liste najčešćih termina,
- pretraživanje i omogućava pregled rezultata,
- analiziranje rezultata pretrage,
- selektivno brisanje dokumenata iz indeksa,
- rekonstrukciju originalnih polja u dokumentu, menja ih i ponovo upisuje u indeks,
- optimizaciju indeksa.

Luke 5.2.0

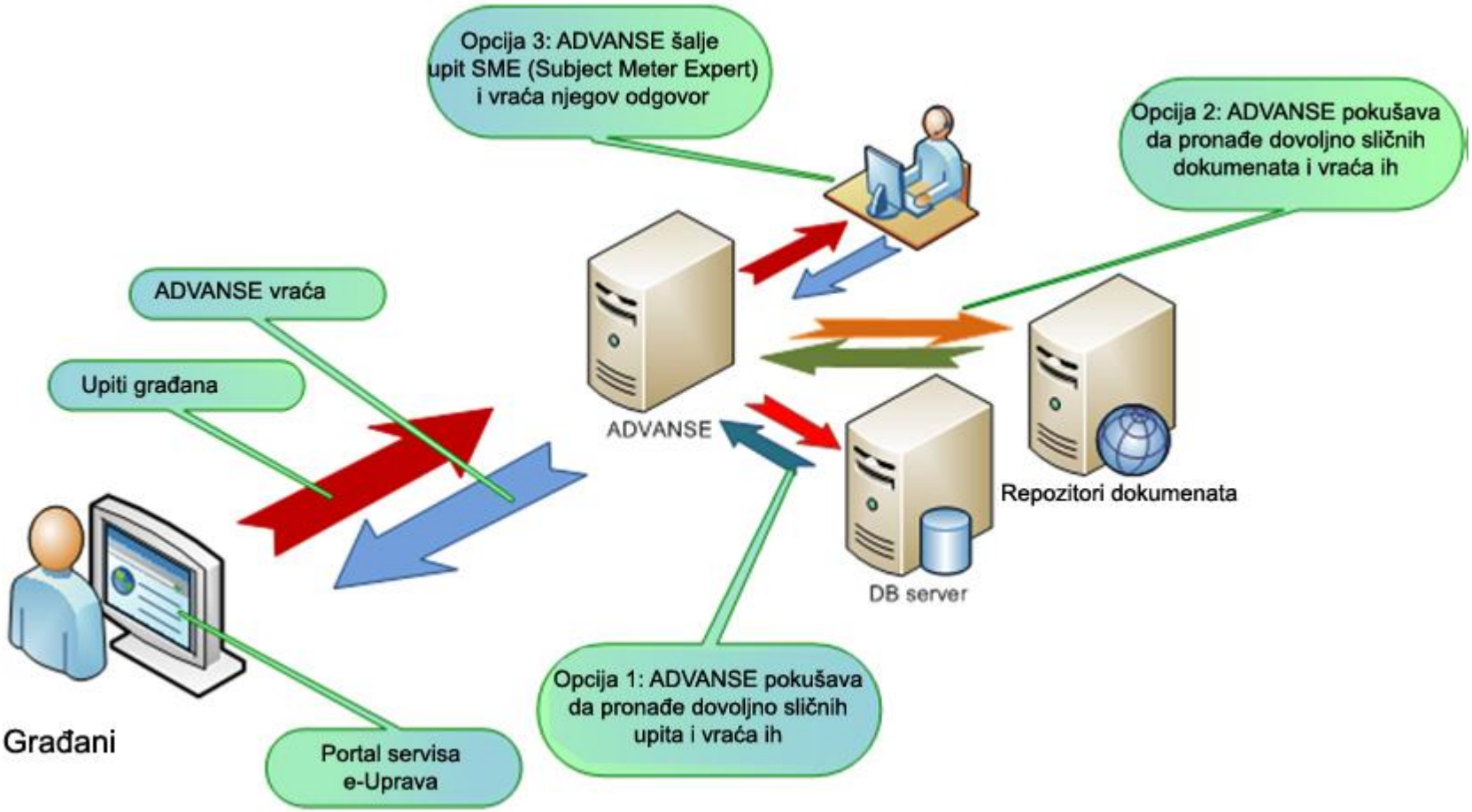

Luke - Lucene Index Toolbox (5.2.0)
File Tools Settings Help
Overview Documents Search Commits Plugins
Index name: D:\Na
Number of fields: 3
Number of documents: 85
Number of terms: 8819
Has deletions? / Optimized?: No / Yes
Index version: c
Index format: Lucene 5.0
Index functionality: flexible, codec-specific
Directory implementation: org.apache.lucene.store.SimpleFSDirectory
Currently opened commit point: segments_4 (generation=4, segs=1)
Current commit user data: --
Re-open
Commit
Close
Select fields from the list below, and press button to view top terms in these fields. No selection means all fields.
Available fields and term counts per field:
Name | Term count | % | Decoder
contents | 8,730 | 98.99 % | string utf8
path | 85 | 0.96 % | string utf8
modified | 4 | 0.05 % | string utf8
Show top terms >>
Number of top terms: 500
Hint: use Shift-Click to select ranges, or Ctrl-Click to select multiple fields (or unselect all).
Tokens marked in red indicate decoding errors, likely due to a mismatched decoder.
Top ranking terms. (Right-click for more options)
Rank | Freq | Field | Text
1 | 85 | contents | government
2 | 85 | contents | one
3 | 85 | contents | a
4 | 85 | contents | that
5 | 85 | contents | no
6 | 85 | contents | york
7 | 85 | contents | this
8 | 85 | modified | P
9 | 85 | contents | their
10 | 85 | modified | @ K
11 | 85 | contents | is
12 | 85 | contents | would
13 | 85 | contents | of
14 | 85 | contents | people
15 | 85 | contents | for
16 | 85 | contents | state
17 | 85 | contents | not
Select a field and set its value decoder: string utf8 Set
Index name: D:\Na

5. Modelovanje sistema za dobijanje brzih odgovora za servise e-Uprave Republike Srbije u oblasti Krivičnog zakonika

- Sistem za dobijanje brzih odgovora;
- Question Answer, Klasterovanje, Pristup BoW, Pristup BoC;
- Natural language processing (NLP), N-gram analiza;

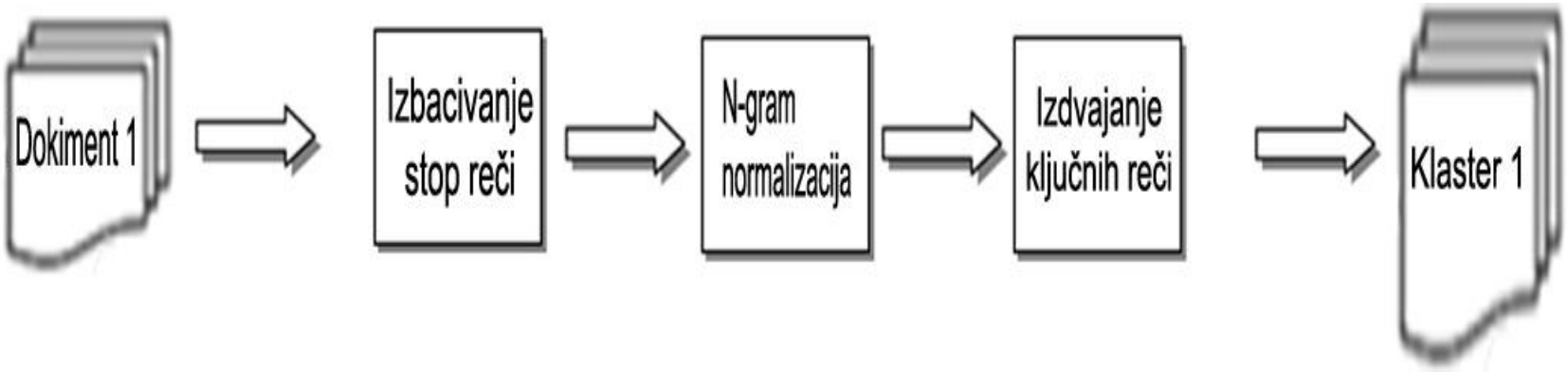
- Domen: Krivični zakonik Republike Srbije

Interakcija između građana i sistema predstavljena ADVANSE sistemom



Savremeni trendovi u razvoju e-Uprave

- Natural language processing (NLP)
- N-gram analiza
- Klasterovanje



Sistem brzih odgovora

- Question Answer (QA);
- TM u QA (Klasterovanje, Klasifikaciju);
- Pristup BoW;
- Pristup BoC;

Bag of Words (BoW)

- Uspostavlja niz ključnih reči i rečenica na osnovu faktora statističke analize (učestalost pojavljivanja termina i distribucija).
- Stop reči

Metode za automatsko izdvajanje ključnih reči iz tekstualnih dokumenata

- jednostavan statistički pristup (statističke informacije o rečima zasnovane na učestalosti pojavljivanja reči, TF-IDF, matrici uzajamnog pojavljivanja itd.),
- lingvistički pristup (leksička analiza, sintaksna analiza),
- pristup baziran na mašinskom učenju (model se generiše na bazi skupa dokumenata, gde su izdvojene ključne reči i koristi se za pronalaženje istih u novim dokumentima),
- poziciono-težinski pristup (reči na različitim pozicijama imaju različit stepen značajnosti).

Ključne reči i normalizacija

R.br.	Termin	Frekvencija
1.	delo	3
2.	učinilac	3
3.	povreda	2
4.	nastupila	2
5.	zdravlje	2
6.	telesna	1
7.	teška	1

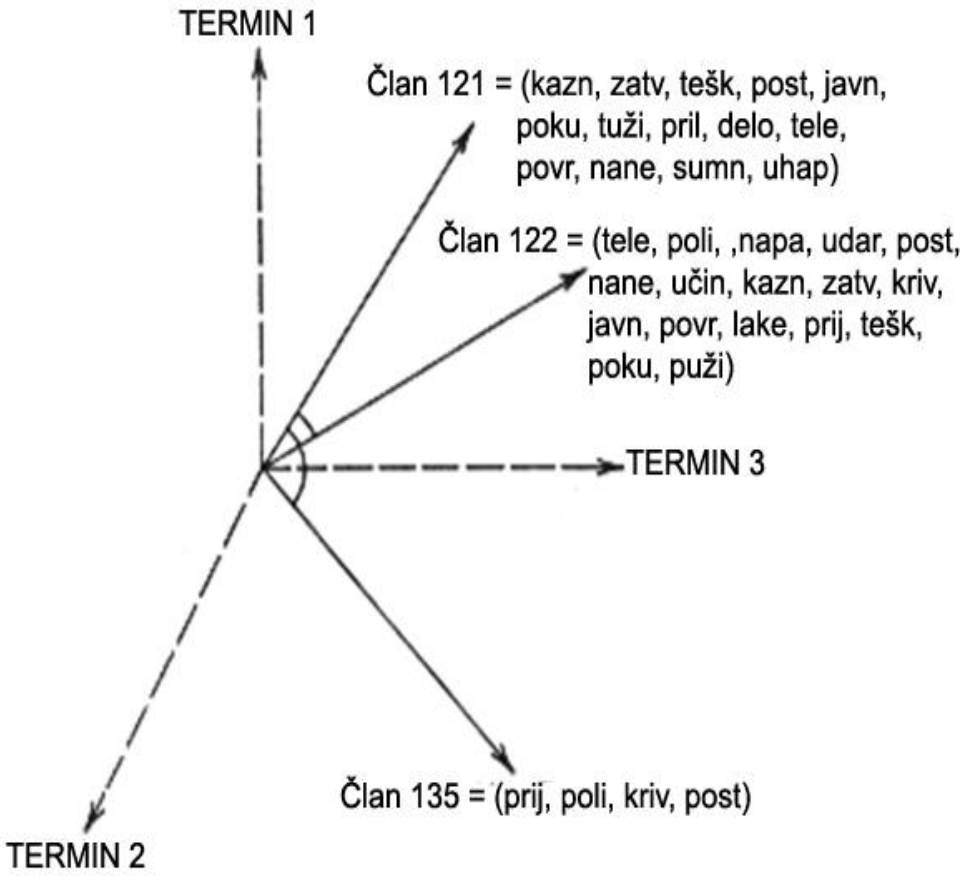
BoC

- Konceptualizacija kratkog teksta ima za cilj da izdvoji skup najreprezentativnijih termina koji ga najbolje opisuju;
- N-gram.

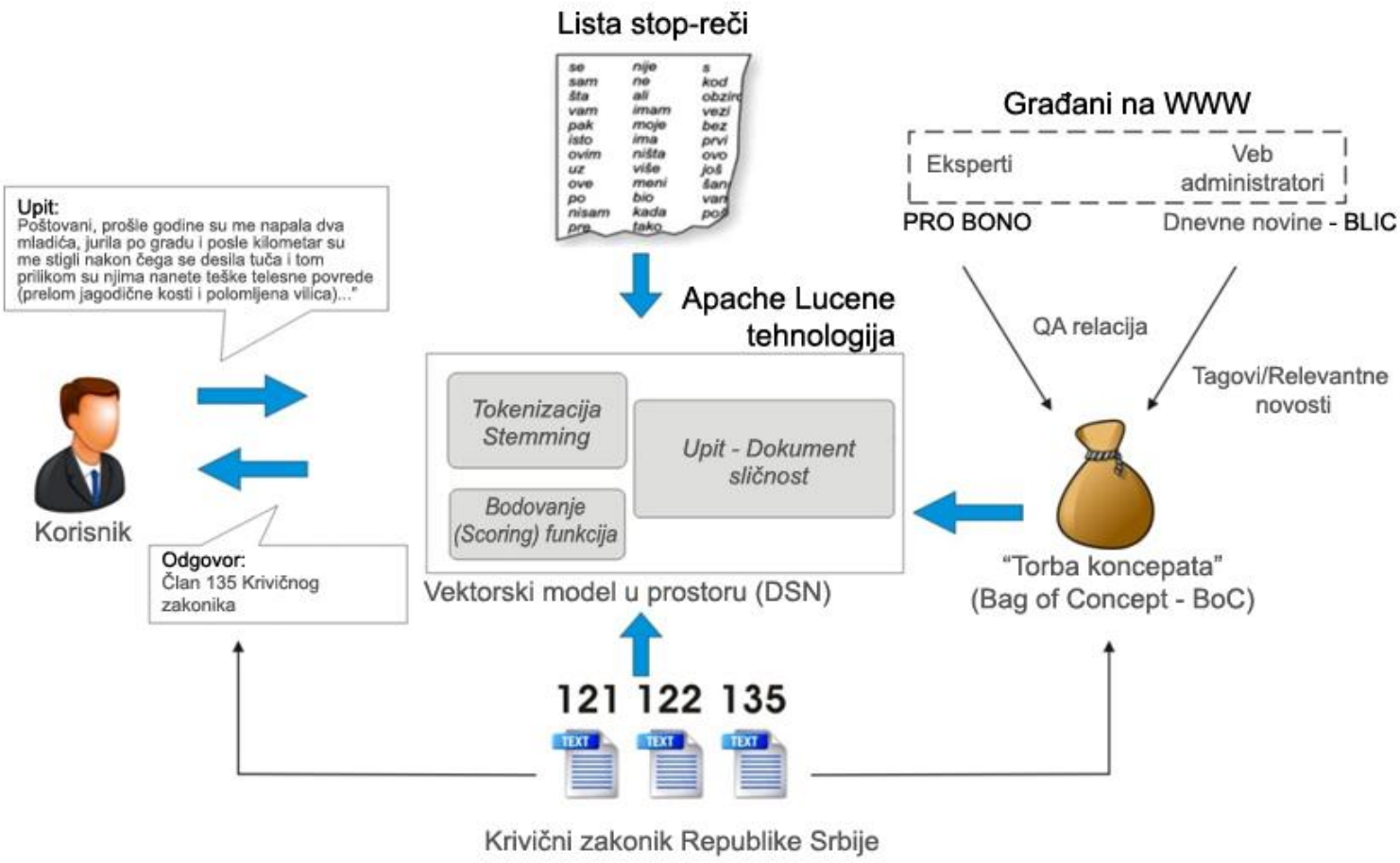
N-gram (n=4)

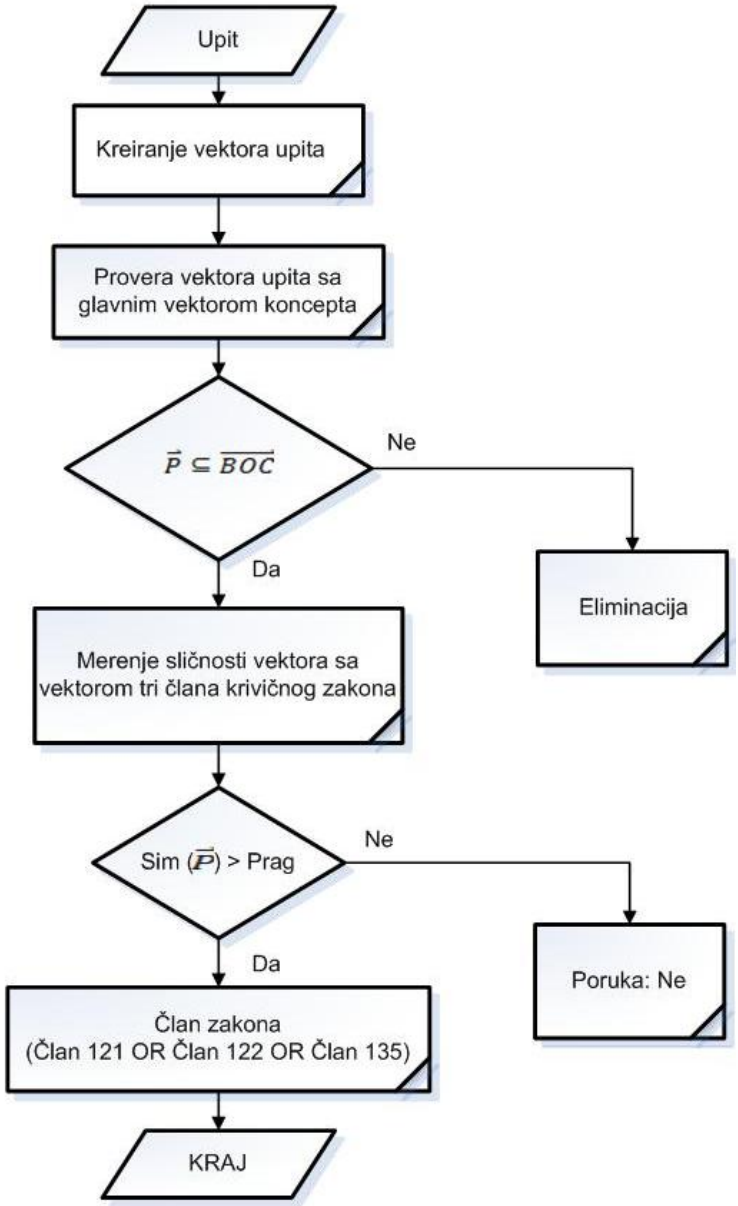
R.br.	Termin	Frekvencija
1.	delo	3
2.	učin*	3
3.	kazn*	3
4.	zatv*	3
5.	tele*	3
6.	tešk*	3
7.	povr*	3

Vektori dokumenata (tri člana Krivičnog zakonika) u prostoru



QA sistem baziran na BoC modelu





Algoritam za dela sistema za klasifikaciju upita za članove krivičnih dela Krivičnog zakonika

6. Analiza eksperimentalnih rezultata

- Tri člana Krivičnog zakonika Republike Srbije:
  - Član 121
  - Član 122
  - Član 135
- 540 upita

31 reč

R.br.	Reč
1	delo
2	učin*
3	kazn*
4	zatv*
5	tele*
6	tešk*
7	povr*
8	prot*
9	prij*
10	poli*
11	kriv*
12	napa*
13	podn*
14	udar*
15	post*
16	nane*
17	sumn*
18	lake
19	osno*
20	javn*
21	osum*
22	pril*
23	sasl*
24	poku*
25	uhap*
26	zadr*
27	prin*
28	saop*
29	upra*
30	tuži*
31	dogo*

Grupisanje reči po članovima

a) Prikupljen je skup od 10 inicijalnih pitanja sa portala PRO BONO čiji su odgovori direktno vezani sa 3 pomenuta člana Krivičnog zakonika Republike Srbije. Za svako pitanje formiran je poseban skup reči iz definisanog skupa reči.

b) Prikupljen je skup od 45 pitanja sa istog portala radi uočavanja pojavljivanja definisanih reči, a koja su kao rezultat imala ponuđene članove Krivičnog zakonika Republike Srbije.

c) Prikupljen je skup tekstova koje je administrator veb-sajta Blica označio kao telesne povrede.

d) Klasterovanje je urađeno prebrojavanjem pojavljivanja reči u skupu a), skupu b) i skupu c) i dobijen je reprezentativan skup za date članove.

Reprezentativni skupovi za tri člana krivičnog zakonika

Član 121		Član 122		Član 135
kazn*		tele*		prij*
zatv*		poli*		poli*
tešk*		napa*		kriv*
post*		udar*		post*
javn*		post*		
poku*		nane*		
tuži*		učin*		
pril*		kazn*		
delo*		zatv*		
tele*		kriv*		
povr*		javn*		
nane*		povr*		
sumn*		lake		
uhap*		prij*		
		tešk*		
		poku*		
		tuži*		

Reprezentativni skupovi reči u vektorskom obliku, za svaki član kao BoC:

- Član 121 [kazn,zatv,tešk,post,javn,poku,tuži, pril,delo,tele,povr, nane,sumn,uhap]
- Član 122 [tele,poli,napa,udar,post,nane,učin, kazn,zatv,kriv,javn,povr,lake,prij,tešk,poku,tuži]
- Član 135 [prij,poli,kriv,post]

Mere sličnosti (distance):

- Kosinusna sličnost

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|^2 \times |\vec{t_b}|^2}$$

- Džakard korelacioni koeficijent

$$SIM_J(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a}.\vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a}.\vec{t_b}}$$

- Euklidova distanca

$$D_E(\vec{t_a}, \vec{t_b}) = \left(\sum_{t=1}^{n} |w_{t,a} - w_{t,b}|^2\right)^{\frac{1}{2}}$$

Referentna mera sličnosti

Primer 1:

Upit KP 1: „*Poštovani, zanima me kada i kako zastareva delo urađeno pre više od 10 godina? U pitanju je teška telesna povreda. Ja sam se potukao sa jednim momkom. Pošto sam ga udario vema nezgodno on je pao i onda sam ga ja u besu polomio. Još se vodi postupak i nije rešeno ništa. Koliko je vremena potrebno da sve to zastari i kako?*“

Vektorski oblik upita KP 1: {delo, tele, tešk, povr, udar, post}

Član 121 {kazn, zatv, tešk, post, javn, poku, tuži, pril, delo, tele, povr, nane, sumn, uhap}

	delo	tele	tešk	povr	udar	post	kazn	zatv	javn	poku	tuži	pril	nane	sumn	uhap
v1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
v2	8	2	6	6	0	0	6	10	1	0	0	0	0	0	0

Član 122 {tele, poli, napa, udar, post, nane, učin, kazn, zatv, kriv, javn, povr, lake, prij, tešk, poku, tuži}

	delo	tele	tešk	povr	udar	post	poli	napa	nane	učin	kazn	zatv	kriv	javn	lake	prij	poku	tuži
v1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
v2	1	1	2	3	0	0	0	0	1	3	3	2	0	0	0	0	0	0

Član 135 { prij, poli, kriv, post}

	delo	tele	tešk	povr	udar	post	prij	poli	kriv
v1	1	1	1	1	1	1	0	0	0
v2	3	2	3	2	0	0	0	0	0

Zbirna tabela

	Član 121	Član 122	Član 135	Ekspert
Upit KP 1	sim(Cos)= 0.539644 sim(Jacc.)= **1.000000** sim(Eucl.)= 15.459625	sim(Cos)= 0.463586 sim(Jacc.)= 0.800000 sim(Eucl.)= 5.477226	sim(Cos)= **0.800641** sim(Jacc.)= **1.000000** sim(Eucl.)= **3.464102**	Član 121
Upit KP 2	sim(Cos)= 0.692688 sim(Jacc.)= 0.750000 sim(Eucl.)= 14.071247	sim(Cos)= 0.717109 sim(Jacc.)= **0.800000** sim(Eucl.)= 4.358899	sim(Cos)= **0.891133** sim(Jacc.)= 0.666667 sim(Eucl.)= **3.872983**	Član 121
Upit KP 3	sim(Cos)= 0.438231 sim(Jacc.)= 0.833333 sim(Eucl.)= 15.297059	sim(Cos)= 0.438599 sim(Jacc.)= **0.909091** sim(Eucl.)= 5.656854	sim(Cos)= **0.685994** sim(Jacc.)= 0. 750000 sim(Eucl.)= **3.000000**	Član 122
Upit KP 4	sim(Cos)= 0.269892 sim(Jacc.)= 0.900000 sim(Eucl.)= 16.248077	sim(Cos)= **0.081111** sim(Jacc.)= 0.909091 sim(Eucl.)= 6.324555	sim(Cos)= 0.500000 sim(Jacc.)= **1.000000** sim(Eucl.)= **2.645751**	Član 135
Upit KP 5	sim(Cos)= 0.660926 sim(Jacc.)= **1.000000** sim(Eucl.)= 15.394804	sim(Cos)= 0.507833 sim(Jacc.)= 0.777778 sim(Eucl.)= 5.385165	sim(Cos)= **0.877058** sim(Jacc.)= **1.000000** sim(Eucl.)= **3.316625**	Član 121
Upit KP 6	sim(Cos)= 0.367277 sim(Jacc.)= 0.900000 sim(Eucl.)= 15.937377	sim(Cos)= 0.469809 sim(Jacc.)= 0.900000 sim(Eucl.)= 5.385165	sim(Cos)= **0.639602** sim(Jacc.)= **1.000000** sim(Eucl.)= **4.582576**	Član 122
Upit KP 7	sim(Cos)= 0.413528 sim(Jacc.)= **1.000000** sim(Eucl.)= 15.362291	sim(Cos)= 0.452589 sim(Jacc.)= 0.818182 sim(Eucl.)= **5.656854**	sim(Cos)= **0.476469** sim(Jacc.)= **1.000000** sim(Eucl.)= 7.211103	Član 121
Upit KP 8	sim(Cos)= 0.622992 sim(Jacc.)= **0.916667** sim(Eucl.)= 15.297059	sim(Cos)= **0.790875** sim(Jacc.)= 0.900000 sim(Eucl.)= **3.872983**	sim(Cos)= 0.744279 sim(Jacc.)= 0.875000 sim(Eucl.)= 7.348469	Član 122
Upit KP 9	sim(Cos)= 0.368166 sim(Jacc.)= **1.000000** sim(Eucl.)= 15.968719	sim(Cos)= 0.497468 sim(Jacc.)= 0.777778 sim(Eucl.)= 5.385165	sim(Cos)= **0.693103** sim(Jacc.)= **1.000000** sim(Eucl.)= **3.000000**	Član 121
Upit KP 10	sim(Cos)= 0.576214 sim(Jacc.)= **1.000000** sim(Eucl.)= 13.96424	sim(Cos)= 0.485574 sim(Jacc.)= 0.909091 sim(Eucl.)= 8.944272	sim(Cos)= **0.72175** sim(Jacc.)= **1.000000** sim(Eucl.)= **5.567764**	Član 135

Provera preciznosti predloženog sistema

- Prikupljeno je 540 upita iz oblasti Krivičnog zakonika Republike Srbije;
- Sistem je odmah eliminisao 130 pitanja koja se ne odnose date članove;
- Za dalju analizu i proveru sistema korišćeno je 410 upita.

„Zlatni“ standard

$$Preciznost\ (eng.Precision) = \frac{Relevantno\ Pronađeno}{Pronađeno}$$

$$Odziv\ (eng.Recall) = \frac{Relevantno\ Pronađeno}{Relevantno}$$

$$F_{i\ (i=1,n)} = \frac{2 * preciznost_i * odziv_i}{preciznost_i + odziv_i}$$

$$F_{Prosečno} = \frac{F_1 + F_2 + \cdots + F_n}{N}$$

Tačnost:

$$Tačnost\ (eng.Accuracy) = \frac{TP + TN}{TP + TN + FP + FN}$$

- gde je:

  TP – Tačno Pozitivno (eng. True Positive);

  TN – Tačno Negativno (eng. True Negative);

  FP – Netačno Pozitivno (eng. False Positive);

  FN – Netačno Negativno (eng. False Negative).

Rezultati

Preciznost	75,71 %
Odziv	57,00 %
$F_{Prosečno}$	0.6936
TP	287
FP	205
FN	123
Tačnost	66,66 %

7. Zaključak

Na osnovu prethodno izloženog, u doktorskoj disertaciji su ostvareni sledeći rezultati:

- Analizirana je mogućnost i ograničenja primene postojećih analitičkih metoda u predloženom sistemu. Posebna pažnja u disertaciji je data na TF-IDF meri kojom se izdvaja najbolji skup onih atributa koji su frekventni u individualnim dokumentima ali se retko pojavljuju u ostalom delu kolekcije.
- Pregled tehnologija modelovanja u servisima e-Uprave brojnih zemalja sveta, kao i unapređenje značenja pronađenih informacija kroz algorime bazirane na takozvanom BoW modelu. Zakon ili njegovi delovi su preko ovog modela predstavljeni kao neuređeni skup reči, zanemarujući pri tome njihov redosled.
- Konkretno opisan način kreiranja korpusa za modelovanje znanja kroz primer jedne vrste krivičnih dela. Predstavljen je proces ekstrakcije i prevođenja suštine iz teksta napisanog prirodnim jezikom u jasno definisan format. Pitanja koja su ovde modelovana napisana su ljudski razumljivim jezikom, tzv. prirodnim jezikom. Takođe, prezentovan je sistem za obradu teksta na prirodnom jeziku, pronalaženje i vizuelno predstavljanje konteksta i koncepata. Na taj način izvršeno je unapređenje postojećih tehnika označavanja koje mogu da obezbede odgovor sa ekstrahovanog dela (delova) dokumenta umesto celog tela dokumenta.

- Kako bi se obezbedio adekvatan odgovor iz ekstrahovanih delova dokumenata umesto iz celih tekstova dokumenata u disertaciji su ispitivane različite tehnike merenja sličnosti u tekstualnim dokumentima. Ispitivane su razne funkcije sličnosti, kao što je Džakard korelacioni koeficijent, Euklidova distanca i kosinusna sličnost između dokumenta i kratkih tekstova koji predstavljaju pitanja građana. Džakard korelacioni koeficijent ukazuje na veliku sličnost klaster analize i faktorske analize u kratkim tekstualnim dokumentima.
- Izvršena eksperimentalna provera efikasnosti predloženog sistema primenom najčešće osnovnih mere za efektivnost pronalaženja podataka - preciznost i odziv. Eksperiment je urađen na skupu podataka iz postojećih elektronskih verzija Krivičnog zakonika Republike Srbije s jedne strane i skupom pitanja izdvojenih sa internet portala za besplatnu pravnu pomoć – PRO BONE, s druge strane.

KRAJ