Srpski web-korpus PDRS

koncepcija, izgradnja, karakteristika i perspektive

Philipp Wasserscheidt
Circle U.
Humboldt-Universität zu Berlin

Circle U.
European University Alliance

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Koncepcija: mogućnosti

Veb korpus …

- Sa svim sadržajem jednog top-level-domena (.rs)
- Sa svim sadržajem na srpskom
- Sa svim sadržajem jedne mreže/ jednog izvora
- Otkud URL?

Fond za nauku Republike Srpbije 7750183
“Javni Diskurs u Republici Srbiji"

Koncepcija: Realizacija

PDRS:

- Isključivo domena .rs
- Izvori koji koriste srpsku latinicu/ćirilicu
- Korišćenje ključnih reči – pretraga preko pretraživač (Bing)

Izgradnja

Izgradnja: Koraci

- Priprema spiska reči
- Pretraga i sačuvanje sajtova
- Transliteracija
- Čišćenje podataka
- Anotacija
- Validacija
- Post-obrada
- Objava

Ključni reči

- Izvor: SrWaC
- Apsolutna frekventnost između 5.000 i 500.000
- Latinica
- Filtiranje
  - Najmanje 4 slova
  - Strane reči
  - Imena
- Ostaju:
  - ekavske/ijekavske forme
  - Forme bez dijakritika

Ključni reči

- Randomizirano 2.800 reči

(U BootCat-u …)

- Spisci od 20 pojmova
- Skupovi od 3 reči
- Pretraga preko Bing
- Preuzimanje od prvih 50 veb stranica

Transliteracija

- Tager ne čita ćirilicu
- Pajton paket *cyrtranslit*

Čišćenje

- Deduplikacija URL / dokumenata
- Dokumente bez [čćšđž]
- Paragrafi kao
  - ----
  - ….
  - | | |
  - Tagovi (<s>, <p>,
, …)
- Linije iz cifara, jedne krakte reči, velikih slova

Metapodaci

- ID (dokument, paragrafi, rečenica)
- URL
- Area
- Collected Date

Anotacija –“POS“

- CLASSLA Stanza annotation pipeline 1.2
- Serbian, nonstandard
- Samo tokenizacija, lematizacija, POS-anotacija
- POS:
  - UD POS (=UPOS)
  - MTE (=XPOS)
  - UD features
- Output je u CoNLL-U formatu

```
# newdoc id = PDRS.123243
# url="http://uzmiracun.rs/blog-razmisljajte-o-velikim-i-smelim-idejama-i-raunajte-na-tehnologiju-1365"
# area =
# collected = 09/2022
# newpar id = 1
# sent_id = 1.1
# text = Razmišljajte o velikim i smelim idejama i računajte na tehnologiju  Artur Turemka - Generalni direktor za tržišta Balkana, Mastercard
1   Razmišljajte    razmišljati VERB    Vmm2p   Mood=Imp|Number=Plur|Person=2|VerbForm=Fin  _   _   _   _
2   o   o   ADP Sl  Case=Loc    _   _   _   _
3   velikim velik   ADJ Agpfply Case=Loc|Definite=Def|Degree=Pos|Gender=Fem|Number=Plur _   _   _   _
4   i   i   CCONJ   Cc  _   _   _   _   _
5   smelim  smeo    ADJ Agpfply Case=Loc|Definite=Def|Degree=Pos|Gender=Fem|Number=Plur _   _   _   _
6   idejama ideja   NOUN    Ncfpl   Case=Loc|Gender=Fem|Number=Plur _   _   _   _
7   i   i   CCONJ   Cc  _   _   _   _   _
8   računajte   računati    VERB    Vmm2p   Mood=Imp|Number=Plur|Person=2|VerbForm=Fin  _   _   _   _
9   na  na  ADP Sa  Case=Acc    _   _   _   _
10  tehnologiju tehnologija NOUN    Ncfsa   Case=Acc|Gender=Fem|Number=Sing _   _   _   _
11  Artur   Artur   PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
12  Turemka Turemka PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
13  -   -   PUNCT   Z   _   _   _   _   _
14  Generalni   generalan   ADJ Agpmsny Case=Nom|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing    _   _   _   _
15  direktor    direktor    NOUN    Ncmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
16  za  za  ADP Sa  Case=Acc    _   _   _   _
17  tržišta tržište NOUN    Ncnpa   Case=Acc|Gender=Neut|Number=Plur    _   _   _   _
18  Balkana Balkan  PROPN   Npmsg   Case=Gen|Gender=Masc|Number=Sing    _   _   _   SpaceAfter=No
19  ,   ,   PUNCT   Z   _   _   _   _   _
20  Mastercard  MasterCard  PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing    _   _   _   _
21  Razmišljajte    razmišljati VERB    Vmm2p   Mood=Imp|Number=Plur|Person=2|VerbForm=Fin  _   _   _   _
22  o   o   ADP Sl  Case=Loc    _   _   _   _
23  velikim velik   ADJ Agpfply Case=Loc|Definite=Def|Degree=Pos|Gender=Fem|Number=Plur _   _   _   _
24  i   i   CCONJ   Cc  _   _   _   _   _
25  smelim  smeo    ADJ Agpfply Case=Loc|Definite=Def|Degree=Pos|Gender=Fem|Number=Plur _   _   _   _
26  idejama ideja   NOUN    Ncfpl   Case=Loc|Gender=Fem|Number=Plur _   _   _   _
27  i   i   CCONJ   Cc  _   _   _   _   _
```

Anotacija – „Area“

- Area: 10 vrsta veb sajtova
  - media (media outlets with several posts daily)
  - inform (topic-centered sites with infrequent posts - maximum 3 per day)
  - company (presentations of companies)
  - state (websites of government bodies on national, regional and local level)
  - forum (forum posts)
  - portal (topic-centered portals without daily coverage)
  - science (scientific publications)
  - shop (with descriptions of products)
  - database (knowledge bases, dictionaries, databases and similar)
  - community (NGOs, fan clubs, associations and other)
  - education (schools, universities)

Anotacija – „Area“

- Ručna anotacija na osnovi domena
- Bottom-up kategorizacija

Validiranje

- ConLL-U files
- UD-tools
- Level 1

Objavljenje

- Na CLARIN repozitoriji
- Licenca CC BY 4.0

Karakteristika

Dostupnost

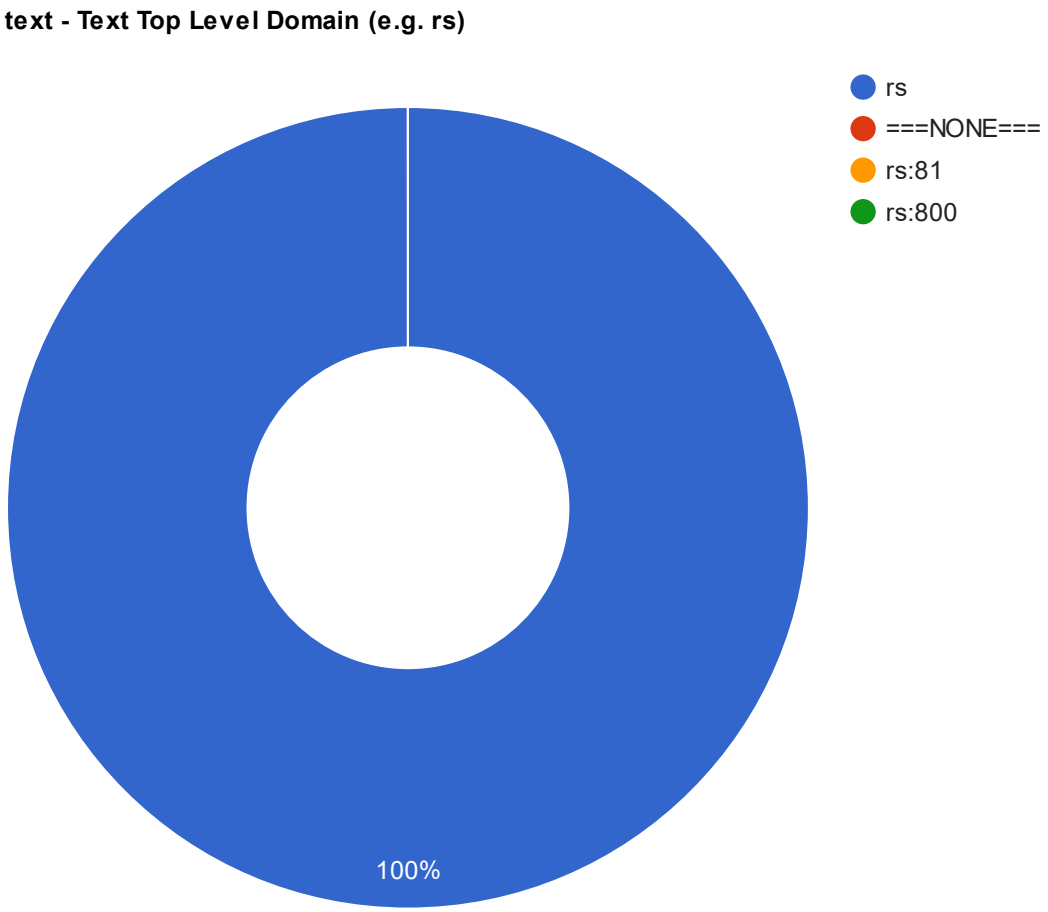
- http://hdl.handle.net/11356/1752
- https://www.clarin.si/ske/#dashboard?corpname=pdrs10

- Potpuno slobodno dalje korišćenje

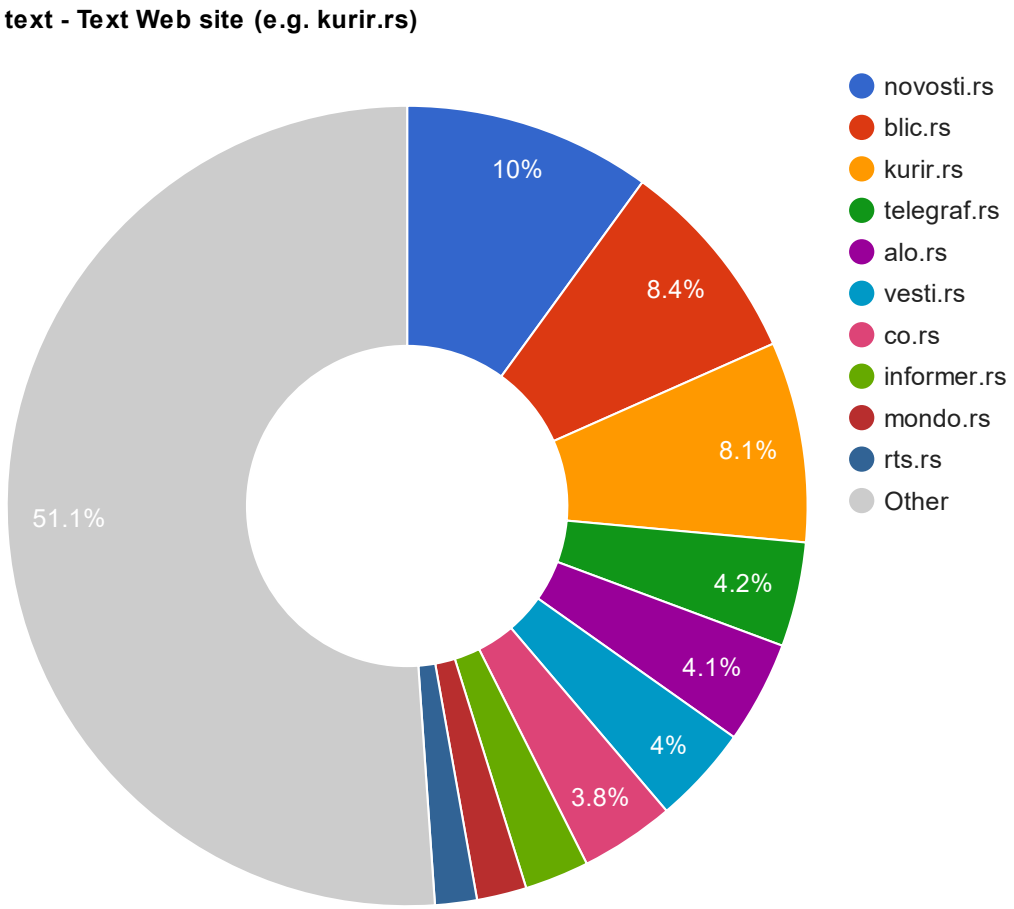
Opseg

- 454.187 tekstova
- 31.401.284 rečenica
- 715.419.977 tokena
- 602.431.307 reči

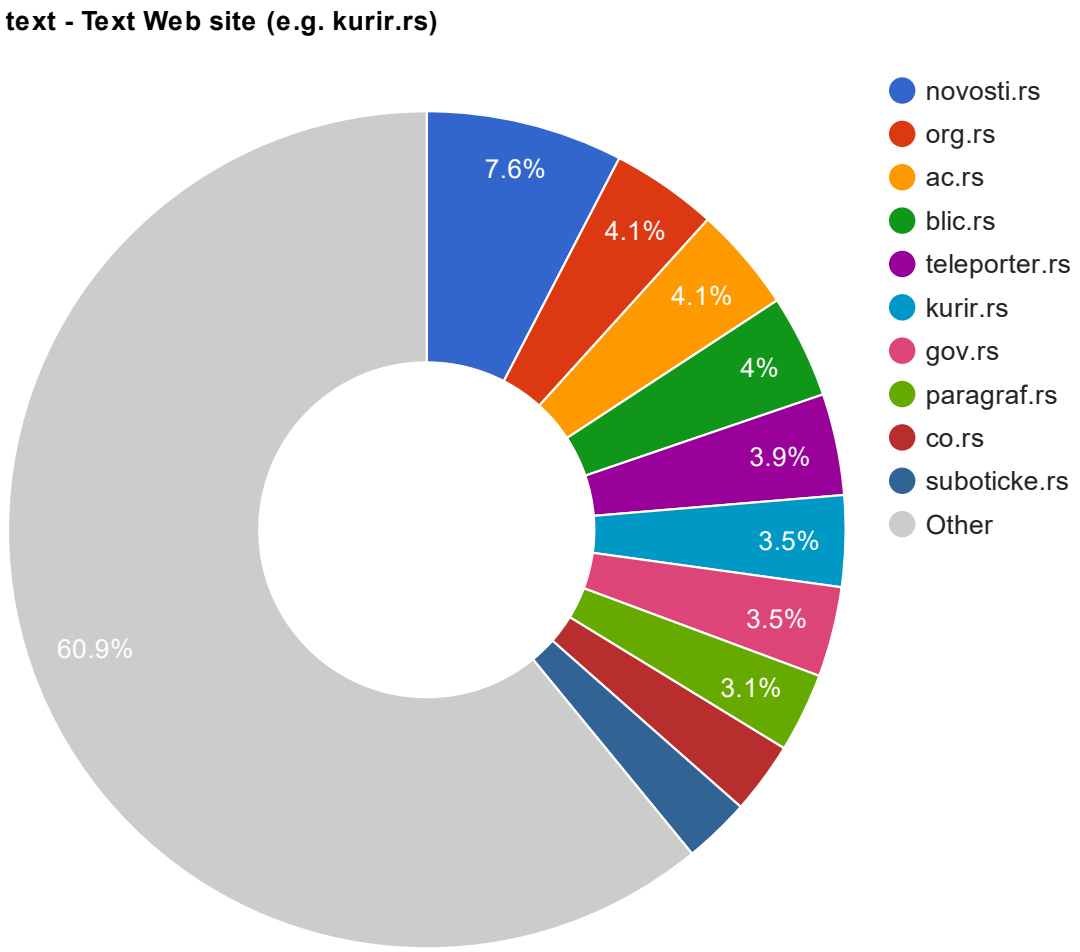
Sadržaj – top-level domen



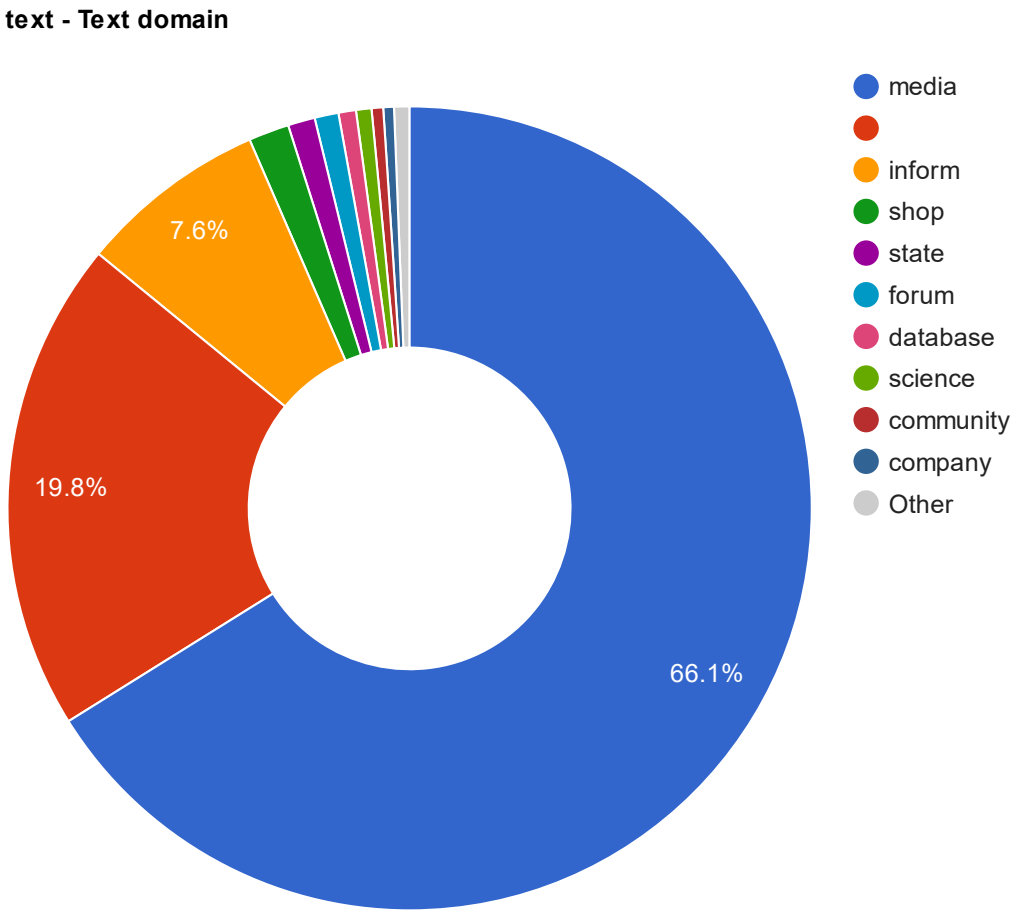
Sadržaj – domeni (broj dokumenata)



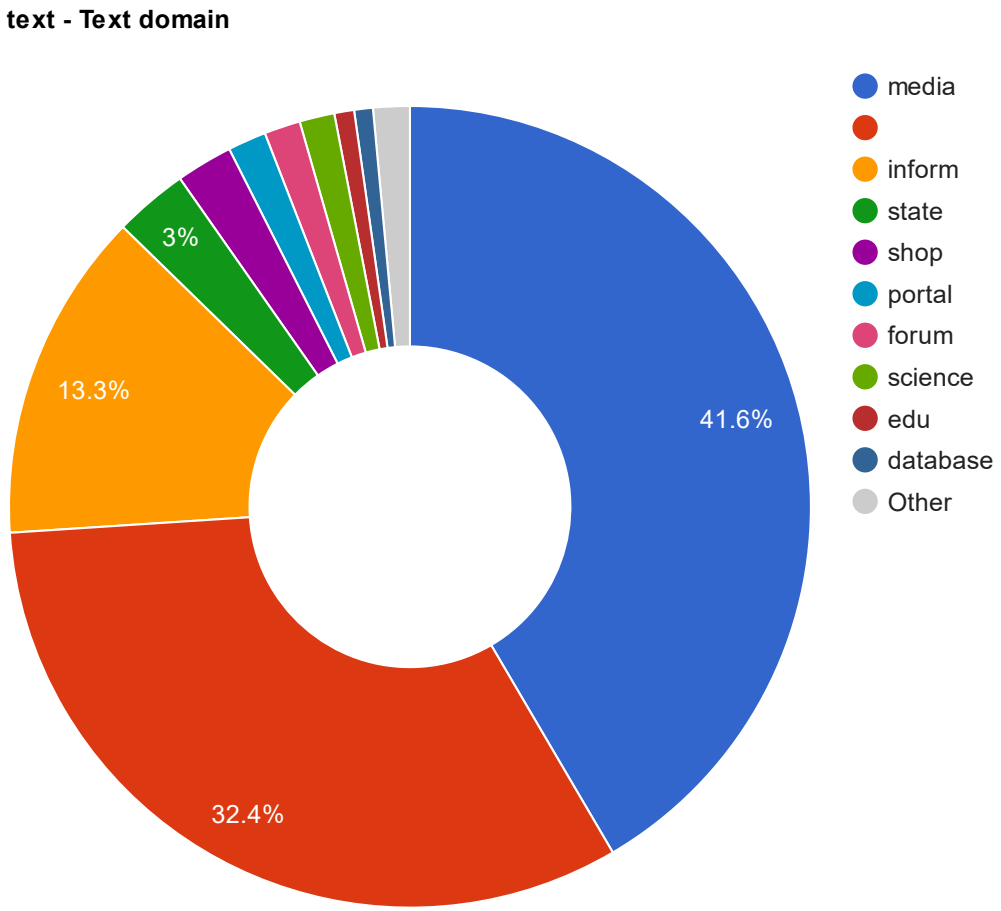
Sadržaj – domeni (broj tokena)



Sadržaj – oblasti (broj dokumenata)



Sadržaj – oblasti (broj tokena)



Kvalitet podataka

personalmag.rs ...	lnih jezika. (OPJ – oblast veštačke inteligencije i	**lingvistike**	koja se bavi proučavanjem problema automatsk
koreni.rs ● med...	ɛdno novo otkriće na polju arhe-o-logije, isto-rije,	**lingvistike**	ili etnografije koje bi mog-lo da po-služi kao osno
org.rs ●	Bugarski, Ranko: Govorite li zajednički (filologija,	**lingvistika**	) 06.52 h; čita: Mirjana Tomić Bugarski, Ranko :
org.rs ●	aćimović, DAISY Ivić, Milka: Pravci u lingvistici (	**lingvistika**	) 11.17 h; čita: Predrag Mihailović Ivkov, Boško:
org.rs ●	, Miloš ; Čudo jezika : Razgovori sa lingvistima (	**lingvistika**	) 14.33 h; čita : Vico Dardić Jeger, Vernar: Paidei
org.rs ●	: Šta se krije iza etimologije (popularna nauka-	**lingvistika**	) 07.28 h; čita : Dušica Popović Ševalić Tempo, S
org.rs ●	ɛktronska knjiga) Bugarski, Ranko: Uvod u opštu	**lingvistiku**	(lingvistika – vaspitanje, obrazovanje, nastava) 1
org.rs ●	jiga) Bugarski, Ranko: Uvod u opštu lingvistiku (	**lingvistika**	– vaspitanje, obrazovanje, nastava) 10.09 h; čita
org.rs ●	ıolingvistička istraživanja, odsek za srpski jezik i	**lingvistiku**	, Filozofski fakultet, Univerzitet u Novom Sadu. N
salonknjiga.rs ...	ca koja je jezikom baratala kao da je doktorirala	**lingvistiku**	, govorila je „da se nešto shvati, mora da se zahv
edu.rs ●	ije – medicina i farmacija, prosveta, psihologija i	**lingvistika**	, studenti, koji su pridavali značaj važnosti shvat
edu.rs ●	zmeđu ispitanika profesije prosveta, psihologija i	**lingvistika**	i ispitanika profesije umetnost i kultura (Sign.=0,(
mom.rs ●	ao obrazovanje iz klasičnih nauka i proučavao je	**lingvistiku**	, pa je počeo da se bavi filozofijom, čitajući Rase

Kvalitet podataka - dublete

ac.rs ●	hotomija ozbiljno neobavezujuće slobodno vreme	**veštačka** i da dokoličarsko iskustvo bolje opisuje kontinuum
rts.rs ● media	e Valjavčeve i Žajine ulice. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	e Valjavčeve i Žajine ulice. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	znaci da je sused zasticen. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
multiradio.rs ●...	virusa otkako se pojavio u... Getty ImagesMože li	**veštačka** inteligencija da pronađe lek za virus korona?Čini s
rts.rs ● media	esno prepustila protivniku. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
glas-javnosti.r...	ode – pre pet godina su proces unapredili tako da	**veštačka** inteligencija prepoznaje biljke s fotografija. iNatura
zubarolog.rs ●	uba kako bi izgledali ravno i dobili savršen osmeh.	**Veštačka** krunica je proteska nadoknada koja buhvata ceo o
rts.rs ● media	u salatu, za početak, sami. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	rbije koji je sve dogovarao. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	nteresa građana u regionu. komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	o što to ide ovim tempom." komentari Inteligencija	**Veštačka** inteligencija menja svet, na dobro ili loše zavisi od
rts.rs ● media	adrenalinski sport, odgovara: "Rekla bih da jeste."	**Veštačka** inteligencija se, kaže, koristi prilično, ali obično već

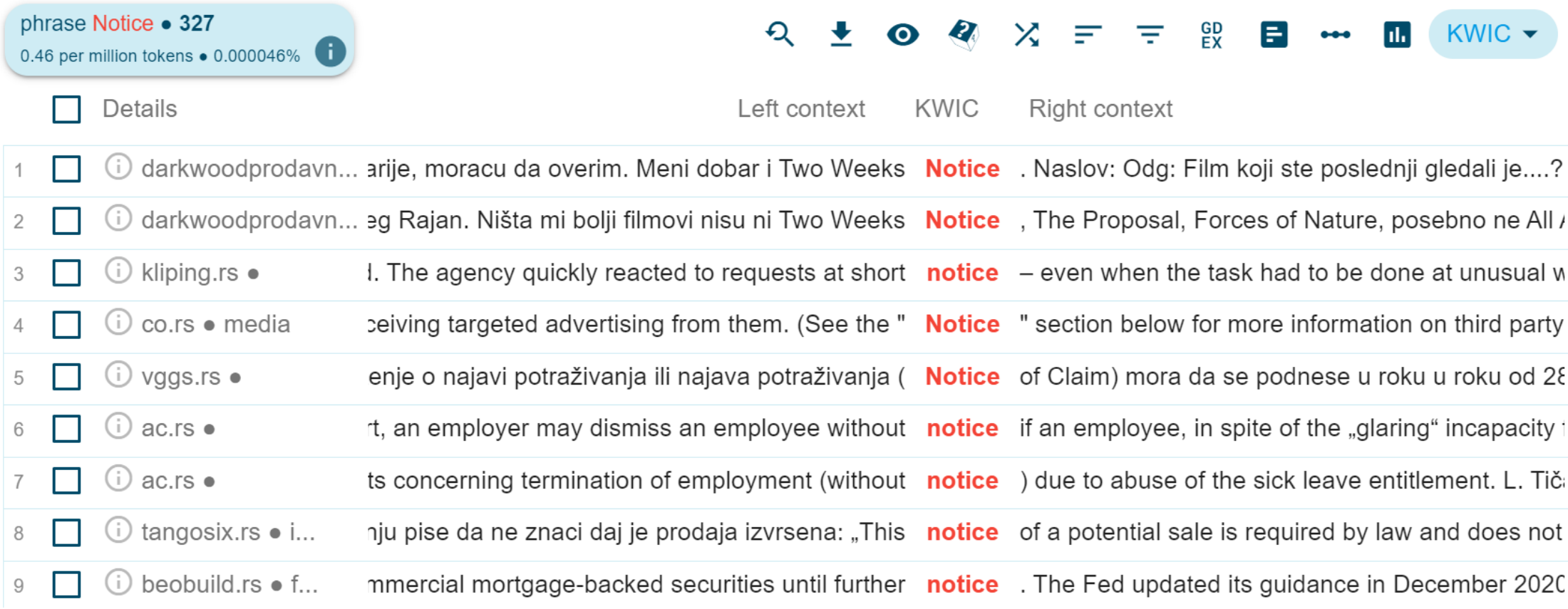
Kvalitet podataka – „tokenizacija“

```
# newpar id = 6
# sent_id = 6.1
# text = Ko je njega ovlastio..
1   Ko  tko PRON    Pq3m-n  Case=Nom|Gender=Masc|PronType=Int,Rel   _   _   _   _
2   je  biti    AUX Var3s   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   _   _   _   _
3   njega   on  PRON    Pp3msa  Case=Acc|Gender=Masc|Number=Sing|Person=3|PronType=Prs  _   _   _   _
4   ovlastio    ovlastiti   VERB    Vmp-sm  Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Act  _   _   _   SpaceAfter=No
5   ..  ..  PUNCT   Z   _   _   _   _   _

# newpar id = 7
# sent_id = 7.1
# text = ...
1   ... ... PUNCT   Z   _   _   _   _   SpaceAfter=No

# sent_id = 7.2
# text = da priča u ime srpskih klubova.
1   da  da  SCONJ   Cs  _   _   _   _   _
2   priča   pričati VERB    Vmr3s   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   _   _   _   _
3   u   u   ADP Sa  Case=Acc    _   _   _   _
4   ime ime NOUN    Ncnsa   Case=Acc|Gender=Neut|Number=Sing    _   _   _   _
5   srpskih srpski  ADJ Agpmpgy Case=Gen|Definite=Def|Degree=Pos|Gender=Masc|Number=Plur    _   _   _   _
6   klubova klub    NOUN    Ncmpg   Case=Gen|Gender=Masc|Number=Plur    _   _   _   SpaceAfter=No
7   .   .   PUNCT   Z   _   _   _   _   _
```

Kvalitet podataka – kodovi

```
# sent_id = 5.8
# text = Popisna komisija Bela Palanka   Notice: Undefined index: loginModal in /var/www/vhosts/belapalanka.org.rs/httpdocs/content/themes/saCity/
1   Popisna popisni ADJ Agpfsny Case=Nom|Definite=Def|Degree=Pos|Gender=Fem|Number=Sing _   _   _   _
2   komisija    komisija    NOUN    Ncfsn   Case=Nom|Gender=Fem|Number=Sing _   _   _   _
3   Bela    beo ADJ Agpfsny Case=Nom|Definite=Def|Degree=Pos|Gender=Fem|Number=Sing _   _   _   _
4   Palanka Palanka PROPN   Npfsn   Case=Nom|Gender=Fem|Number=Sing _   _   _   _
5   Notice  Notice  PROPN   Npfsn   Case=Nom|Gender=Fem|Number=Sing _   _   _   SpaceAfter=No
6   :   :   PUNCT   Z   _   _   _   _
7   Undefined   Undefined   X   Xf  _   _   _   _   _
8   index   Index   X   Xf  _   _   _   _   SpaceAfter=No
9   :   :   PUNCT   Z   _   _   _   _   _
10  loginModal  loginModal  X   Xf  Foreign=Yes _   _   _   _
11  in  in  X   Xf  _   _   _   _   _
12  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
13  var var X   Xf  _   _   _   _   SpaceAfter=No
14  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
15  www www X   Xf  _   _   _   _   SpaceAfter=No
16  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
17  vhosts  vhosts  X   Xf  _   _   _   _   SpaceAfter=No
18  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
19  belapalanka.org.rs/ belapalanka.org.rs/ X   X   _   _   _   _   SpaceAfter=No
20  httpdocs    httpdocs    X   Xf  _   _   _   _   SpaceAfter=No
21  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
22  content Content X   Xf  _   _   _   _   SpaceAfter=No
23  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
24  themes  themes  X   Xf  _   _   _   _   SpaceAfter=No
25  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
26  saCity  saCity  X   Xf  _   _   _   _   SpaceAfter=No
27  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
28  views   views   X   Xf  _   _   _   _   SpaceAfter=No
29  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
30  includes    includes    X   Xf  _   _   _   _   SpaceAfter=No
31  /   /   PUNCT   Z   _   _   _   _   SpaceAfter=No
```

Kvalitet podataka – drugi jezici



	Details	Left context	KWIC	Right context
1	darkwoodprodavn...	arije, moracu da overim. Meni dobar i Two Weeks	Notice	. Naslov: Odg: Film koji ste poslednji gledali je....?
2	darkwoodprodavn...	eg Rajan. Ništa mi bolji filmovi nisu ni Two Weeks	Notice	, The Proposal, Forces of Nature, posebno ne All A
3	kliping.rs •	l. The agency quickly reacted to requests at short	notice	– even when the task had to be done at unusual w
4	co.rs • media	ceiving targeted advertising from them. (See the "	Notice	" section below for more information on third party
5	vggs.rs •	enje o najavi potraživanja ili najava potraživanja (	Notice	of Claim) mora da se podnese u roku u roku od 28
6	ac.rs •	t, an employer may dismiss an employee without	notice	if an employee, in spite of the „glaring“ incapacity
7	ac.rs •	ts concerning termination of employment (without	notice	) due to abuse of the sick leave entitlement. L. Tič
8	tangosix.rs • i...	nju pise da ne znaci daj je prodaja izvrsena: „This	notice	of a potential sale is required by law and does not
9	beobuild.rs • f...	mmercial mortgage-backed securities until further	notice	. The Fed updated its guidance in December 2020

Kvalitet anotacije

- *Sto dinara* (phrase) - PDRS

	MULTEXT-East MSD	Frequency
1	Pi3n-a Ncmpg	61
2	Cs Ncmpg	26
3	Mlc Ncmpg	21
4	Pi3n-n Ncmpg	6
5	Pi3n-a Ncmsg	3
6	Qo Ncmpg	2

Kvalitet

- *Sto dinara* (phrase) - SrWaC

		Tag	Frequency
1	☐	Mlc Ncmpg	237
2	☐	Ncmsan Ncmpg	37
3	☐	Ncmsan Ncmsg	5
4	☐	Pi3n-a Ncmpg	3
5	☐	Pi3n-n Ncmsg	2
6	☐	Pi3n-a Ncmsg	2
7	☐	Ncmsn Ncmsg	1
8	☐	Pi3n-a Npfsn	1
9	☐	Cs Ncmsg	1
10	☐	Cs Ncmpg	1

Kvalitet

- *Sto dinara* (phrase) - Classla-web.sr

		MULTEXT-East MSD	Frequency
1	☐	Mlc Ncmpg	893
2	☐	Pi3n-n Ncmpg	43
3	☐	Cs Ncmpg	16
4	☐	Pi3n-a Ncmpg	14
5	☐	Pi3n-n Ncmsg	8
6	☐	Mlc Ncmsg	6
7	☐	Pi3n-a Ncmsg	2

Kvalitet

- *Hiljada dinara* (phrase) - PDRS

		MULTEXT-East MSD	Frequency
1	☐	Ncfpg Ncmpg	6,619
2	☐	Ncfsn Ncmpg	9

Kvalitet

- *Hiljada dinara* (phrase) - SrWaC

		Tag	Frequency
1	☐	Ncfpg Ncmpg	4,794
2	☐	Ncfsn Ncmpg	796
3	☐	Ncfsn Ncmsg	590
4	☐	Ncfpg Ncmsg	361
5	☐	Npmsg Npmsg	4
6	☐	Ncfpg Npmsg	2

Kvalitet

- *Hiljada dinara* (phrase) - Classla-web.sr

		MULTEXT-East MSD	Frequency
1	☐	Ncfpg Ncmpg	19,944
2	☐	Ncfsn Ncmpg	18

Jezičke pojave

Jezičke pojave – standard/nestandard

Forma	Tpm	lema	Tpm
otišo	0,215	otišao	33,986
došo	0,200	došao	68,954
išo	0,064	išao	15,595
našo	0,063	našao	32,155
prošo	0,046	prošao	19,648
izašo	0,035	izašao	17,468
pošo	0,025	pošao	2,637
ušo	0,024	ušao	24,055
snašo	0,013	snašao	2,077

Jezičke pojave – standard/nestandard

Forma	Frekv.	Forma	Tpm
otišo	154	otići	151
došo	143	doći	143
išo	46	ići	46
našo	45	naći	41
prošo	33	proći	33
izašo	25	izaći	25
pošo	18	poći	18
ušo	17	ući	16
snašo	9	*snašti*	9

Jezičke pojave - paradigme

- M.* + evr.*

Forma	Tokens	TpM
evra	114279	159,74021
evro	3196	4,4674
evrica	127	0,17752
evrića	95	0,13279
evradi	12	0,01677
evri	11	0,01538
evrova	9	0,01258
evre	6	0,00839
evrai	4	0,00559
evrov	3	0,00419

Jezičke pojave - paradigme

- M.* + evr.*

Forma	Tokens	TpM	Lema	Frekv.
evra	114279	159,74021	evro	114,279
evro	3196	4,4674	evro	3,196
evrica	127	0,17752	evrica	127
evrića	95	0,13279	evrić	95
evradi	12	0,01677	evrada	12
evri	11	0,01538	evro	11
evrova	9	0,01258	evrov	9
evre	6	0,00839	evro	3
evrai	4	0,00559	evra	4
evrov	3	0,00419	evrov	3

Jezičke pojave – nove reči

- „drona“
  - SrWaC 23 – PDRS 1.224 tokena

	Lemma (lowercase)	Frequency
1	drona	13
2	dron	10

	Lemma (lowercase)	Frequency
1	dron	1,121
2	drona	96
3	dronati	7

Jezičke pojave – nove reči

- JerTEH

1	ac.rs •	du teksta Društva za jezičke resurse i tehnologije	**Jerteh** Npmsn/Jerteh	66 (Stanković et al. 2020), koji omogućava
2	novosti.rs • me...	ngvistiku Društva za jezičke resurse i tehnologije	**JeRTeH** Npmsn/JeRTeH	iz Beograda u saradnji sa leksikografima iz

Perspektiva

Podaci

- Širenje
- Dodavanje SrWaC-a
  - Nova anotacija sa Classla 2.1
- Dodavanje parlamentskih protokola
  - +/- 3 godina

Priprema

- Deduplikacija na nivou paragrafa
- = Smanjenje sadržaja iz medija

Anotacija

- CLASSLA Stanza pipeline 2.1
- Serbian, tpye ‚web‘ (novo!)
- Puna anotacija
  - Uklj. Dependencije i NER
- Area anotacija
  - Dopunjeno
  - SrWaC

Anotacija

```
# newpar id = 60
# sent_id = 60.1
# text = Veljković i Đorđević ističu da prilikom stvaranja stava o određenom proizvodu, potrošači formiraju
1   Veljković   Veljković   PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing   4   nsubj   _   NER=B-PER
2   i   i   CCONJ   Cc   _   3   cc   _   NER=O
3   Đorđević   Đorđević   PROPN   Npmsn   Case=Nom|Gender=Masc|Number=Sing   1   conj   _   NER=B-PER
4   ističu   isticati   VERB   Vmr3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   0   root
5   da   da   SCONJ   Cs   _   14   mark   _   NER=O
6   prilikom   prilikom   ADP Sg   Case=Gen   7   case   _   NER=O
7   stvaranja   stvaranje   NOUN   Ncnsg   Case=Gen|Gender=Neut|Number=Sing   14   obl _   NER=O
8   stava   stav   NOUN   Ncmsg   Case=Gen|Gender=Masc|Number=Sing   7   obl _   NER=O
9   o   o   ADP Sl   Case=Loc   11   case   _   NER=O
10   određenom   određen ADJ Agpmsly   Case=Loc|Definite=Def|Degree=Pos|Gender=Masc|Number=Sing   11   amod
11   proizvodu   proizvod   NOUN   Ncmsl   Case=Loc|Gender=Masc|Number=Sing   8   nmod   _   NER=O|SpaceA
12   ,   ,   PUNCT   Z   _   7   punct   _   NER=O
13   potrošači   potrošač   NOUN   Ncmpn   Case=Nom|Gender=Masc|Number=Plur   14   nsubj   _   NER=O
14   formiraju   formirati   VERB   Vmr3p   Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin   4   ccom
15   čitav   čitav   ADJ Agpmsann   Animacy=Inan|Case=Acc|Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing
16   set set NOUN   Ncmsan   Animacy=Inan|Case=Acc|Gender=Masc|Number=Sing   14   obj _   NER=O
```

Anotacija

# newpar id = 60										
# sent_id = 60.1										
# text = Veljković i Đorđević ističu da prilikom stvaranja stava o određenom proizvodu, potrošači formiraju čitav set emocionalnih odnosa, što ukazu										
1	Veljković	Veljković	PROPN	Npmsn	Case=Nom|G	4	nsubj	_	NER=B-PER	
2	i	i	CCONJ	Cc	_	3	cc	_	NER=O	
3	Đorđević	Đorđević	PROPN	Npmsn	Case=Nom|G	1	conj	_	NER=B-PER	
4	ističu	isticati	VERB	Vmr3p	Mood=Ind|N	0	root	_	NER=O	
5	da	da	SCONJ	Cs	_	14	mark	_	NER=O	
6	prilikom	prilikom	ADP	Sg	Case=Gen	7	case	_	NER=O	
7	stvaranja	stvaranje	NOUN	Ncnsg	Case=Gen|G	14	obl	_	NER=O	
8	stava	stav	NOUN	Ncmsg	Case=Gen|G	7	obl	_	NER=O	
9	o	o	ADP	Sl	Case=Loc	11	case	_	NER=O	
10	određenom	određen	ADJ	Agpmsly	Case=Loc|De	11	amod	_	NER=O	
11	proizvodu	proizvod	NOUN	Ncmsl	Case=Loc|Ge	8	nmod	_	NER=O|SpaceAfter=No	
12	,	,	PUNCT	Z		7	punct		NER=O	

Metapodaci

Dodatni metapodaci:

- Deo (PDRS, SrWaC, Parlament)
- Dužine dokumenta
- Sekcije (sport, ekonomika, forum, ...)
- Filetype (pdf, doc, html, ...)
- Izvorno pismo (ćirilica, latinica – samo PDRS)
- Žanr (X-GENRE classifier)

Validacija, Licenca

- Validacija
  - UD-tools, level 2
- Licenca
  - CC BY SA 4.0 (zbog SrWaC-a)

Hvala na pažnju!

philipp.wasserscheidt@hu-berlin.de