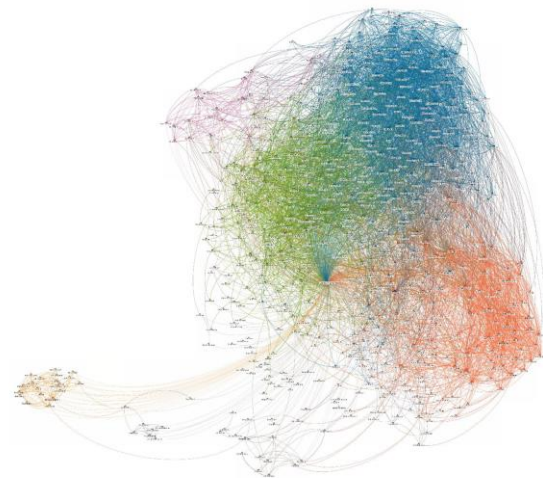


Vizualizacija strukture velikih tekstualnih kolekcija srpskog jezika

Anđelka Zečević
andjelkaz@matf.bg.ac.rs



Pregled tema

- računarska lingvistika
 - klasični zadaci
 - jezički resursi i alati
 - reprezentacija dokumenata
- primeri primene
 - oblaci reči
 - imenovani entiteti
 - analiza sentimenata

Računarska lingvistika

- mašinska obrada jezika, njegovo razumevanje i generisanje
- discipline:
 - lingvistika
 - statistika
 - mašinsko učenje

Računarska lingvistika

- pretraživanje informacija
- klasifikacija pošte
- pravopisni korektori
- predlozi reči prilikom kucanja

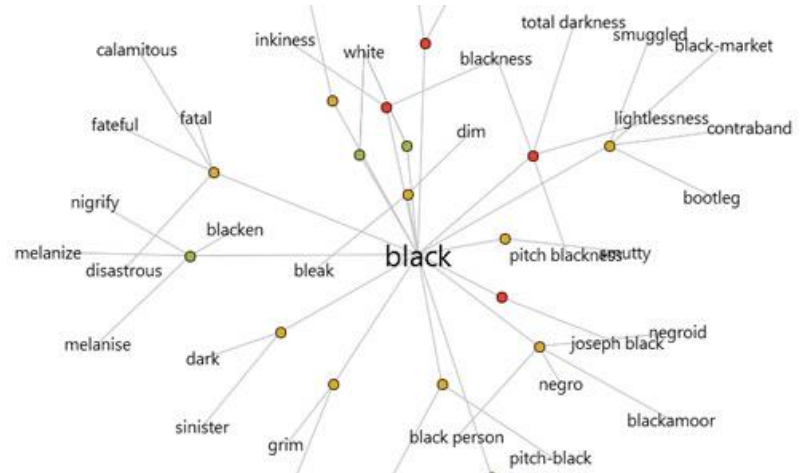
Računarska lingvistika

- superkompjuter koji može da odgovori na proizvoljno pitanje otvorenog tipa
- IBM Watson
 - “Želite li da postanete milioner?” za 2011. godinu
 - 4TB memorije
enciklopedije, literalna dela, vesti, rečnike,
 - obrada 500GB podataka u sekundi



Jezički resursi

- elektronski rečnici
 - *generaciju.generacija.N600:fs4q*
- jezički korpusi
 - jednojezični
 - višejezični
- semantičke i leksičke mreže



Jezički alati

- Osnovni:
 - tokenizeri - izdvajanje tokena
 - tageri - pridruživanje vrste reči i leme tokenima
 - parseri - sintaksna anotacija
 -
- Izvedeni:
 - klasifikatori
 - za ekstrakciju informacija
 - za OCR
 - ...

Zašto je obrada jezika teška?

- višeznačnost
 - *računarstvo i informatika* ali i skup *Informatika*
 - *peta glava* (knjige)
- složene konstrukcije
 - provlačenje subjekta kroz više rečenica
- subjektivna procena važnosti informacije
 - zavisi od ekspertize, iskustva ili zadatka koji korisnik obavlja

Obrada srpskog jezika

JERTEH grupa Univerziteta u Beogradu

- elektronski rečnik
- korpus savremenog srpskog jezika SrpKor2013
- EN-SR i FR-SR paralelni korpusi
- WordNet
- razni alati za anotaciju, klasifikaciju, ekstrakciju informacija, pretraživanje...



<http://www.korpusi.jer.teh.ac.rs>

Zašto vizuelizacija?

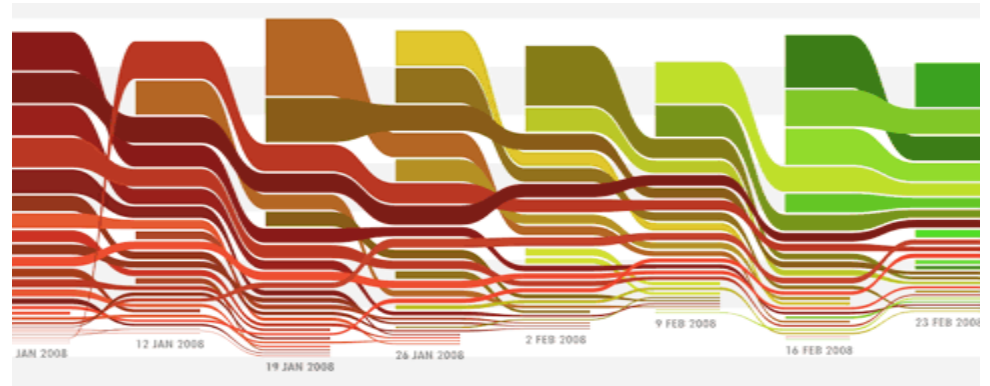
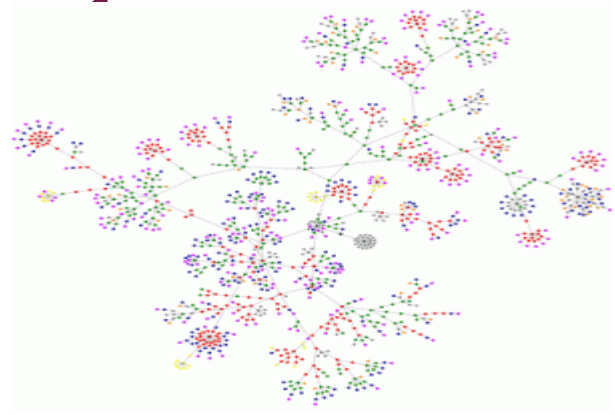
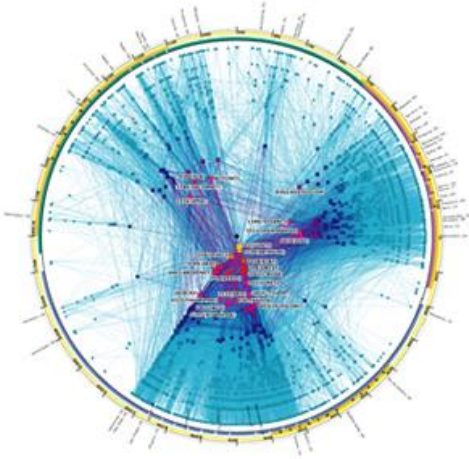
- puno informacija koje treba pročitati i obraditi
- sumarizacija
 - generisanje sažetaka
 - agregacija informacija iz više izvora
- vizuelizacija - brz grafički uvid u strukturu

Kako predstaviti dokument?

- uređeni vektori reči
 - dimenzija je jednaka broju reči jezika
 - izbor odgovarajućih karakteristika
 - gubi se redosled reči i kontekst za koji su vezane
- distribuirani vektori reči
 - realni vektori koji kompozitno oslikavaju morfološke, sintaksne i semantičke karakteristike sadržaja

Mogući izbori vizuelizacije

umetnost u nauci



Vizuelizacija tema

- poseban zadatak klasifikacije dokumenata: odrediti temu/teme zastupljene u tekstu
- pristup:
 - praćenje ključnih pojmova vezanih za neku temu na osnovu frekvencija i statističkih modela jezika
 - algoritmi klasifikacije: logistička regresija, nekoliko najbližih suseda, Naïve Bayes, ...

Vizuelizacija tema

advokat akcija Beograd BiH bit brod broj dan drzava EU Gaza
glumac godina grad gradanin groblje grupa hiljada Izrael
Kosovo ljudi mesto minut nega odnos opstina oruzje pitanje podatak pomoc posao
poslanik pravo predsednik prilika problem put rad rada rec sad sat
Srbija sud svet voda vreme zakon zemlja zivot

Vizuelizacija imenovanih entiteta

- zajednički naziv za imena osoba, gradova, organizacija, događaja, vremenske odrednice, mere...
- zadatak je prepoznati i klasifikovati imena
- pristup:
 - različiti spiskovi ličnih imena, gradova, ...
 - ontologije
 - algoritmi zasnovani na kaskadama transduktora
 - statistički modeli obeležavanja sekvenci: modeli maksimalne entropije i CRF modeli

Vizuelizacija imenovanih entiteta

Dr Zlatko Nikošić	Fidesa Lajoša	Benjamin Netanjahu	Gordana Pop Lašić	Jovanom Jovom	Seka Sabjić	Oskar Davičo	Dragan Kresoja	Miroslav Prokapijević	Novica Kocić	Petar Iad	Filip Milošević	Igor Blažević	Isidora Bjelica	Milena Dragičević	Dominik Moizi	Zoran Musteroš	Dušan Cavić
Milo Đukanović	Angela Merkel	Milan Mišić	Miloslav Petronijević	Zoran Davinić	Stanislav Pešić	Branko Čepić	Artur Milet	Ako Popović	Robin Soderling	Boško Buha	Rade Marjanović	Ivo Andrić	Dušica Vučković	Dimitrije Stanković	Sasa Petrović	Veško Bulajić	
Novak Đoković	Lesli Gelb	Pamela Smok	Jorgovanka Tabaković	Slobodan Šušajgić	Zoran Radmilović	Paja Jovanović	Dobrica Čosić	Nikola Tasić	Nenad Radičević								
Rodžer Federer	Džozef Bajden	Endru Cerlin	Konstantin Arsenović	Branislav Nušić	Sloba Aligrudić	Alfred Nobel	Aleksandra Petrović	Miroslav Zdravković		Mile Dragić	Zorana Markovića	Boris Tadić	Prvoslav Davinić	Mladan Dinkić	Nenad Popović		
Rolan Garos	Kišon Prvi	Ala Gora	Smiša Štamenko	Aleksandra Popovića	Svetislav Gončić	Bernard So	Obren Četković	Rasim Ljajić	Oliver Dučić								
Zorić On	Toma Todorović	Sulejman Tihjić	Elvira Kovač	Draganom Lakovićem	Dragoljub Stevanović	Teoprast Rendo	Mileje Miletić	Miladin Kovačević		Danijel Cvjetičanin							
Toma Zora	Jovan Krivokapić	Ivo Josipović	Aleksandra Janković	Olivera Marković	dr Slobodan	dr Oliver Stojković	Slobodan Jocić	Vladimir Gligorov	Milorad Dorđić	Vuk Jeremić							
Hilari Klinton	Marko Živić	Milan Morničilović	Judiša Popović	Nebojša Kornadina	Aleksandar	Snežana Simić	Dragan Trivun	Mirko Cvetković	Oli Ren		Barak Obama	Rodoljub Gerić	Velimir Bata Živojinović				
P Dyej Krouli	Vlado Georgijev	Vladimir Zaharijev	Srdan Trailović	Mira Lilović	dr Ceda Mihajlović	Robert Goldvoter	Marko Albunović	Dimitrije Ljotić	valentin Incko	Zoran Krašić	Boško Jakić						
Brisa Ortefoa	Milan Tlačinac	Jadranka Božović	Milica Vojić- Marković	Miroslav Belović	Jovan Jovanović	Fernana Lažesa	Miroslavu Labusu	Kir Stefan		Soja Jovanović	Ivo Lola Ribar						
Mari Le Pena	Sulejman Ugljanin	Slavoljub Blagojević	Snežana Stojanović	Milan Milošević	Jova Zmaj	Đermano Celant	Zivadina Jovanovića	Aleksandar Apostolov	Mile Ilić								
Nikola Sarkozi	Dragan Dilas	Stojanka Petković	Branko Ružić	Aleksandra Bakočević	Miroslav Čangalović	Dubravka Lakić	Anica Nikošić	Benjamin Netanjahu									
Mihajl Varga	Martin Mekginis	Zorana Živkovića	Aleksandra Mijalković	Milivoje Ivanišević	Petar Baničević	Srdan Dragojević	Daniilo Suković	Darko Kalezić	Dragan Marković Palma	Božidar Stošić	Dragoljub Đorđević						

Vizuelizacija sentimenata

- odrediti polarisanost, skalu ili dominantnu emociju
- odrediti attribute koji vode do ovih sentimenata
- pristup:
 - afektivni leksikoni (vokabular specifican za neki sentiment npr. tugu, sreću, zadovoljstvo...)
 - specijalne tehnike analize teksta (emotikoni, negacija, ironija, ...)
 - tehnike učenja: metoda potpornih vektora i modeli maksimalne entropije

Vizuelizacija sentimentata



*predsednik u
SrpKor2013
korpusu*

Cilj

- jasniji i pregledniji prikaz podataka
- bolje razumevanje podataka
- širi skup primena

Hvala na pažnji!

andjelkaz@matf.bg.ac.rs