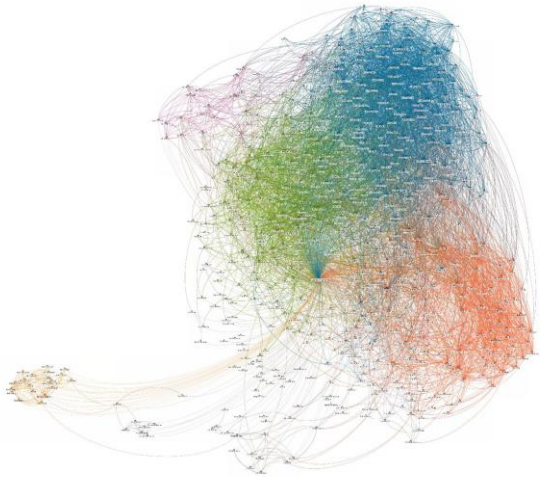
Vizualizacija strukture velikih tekstualnih kolekcija srpskog jezika

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

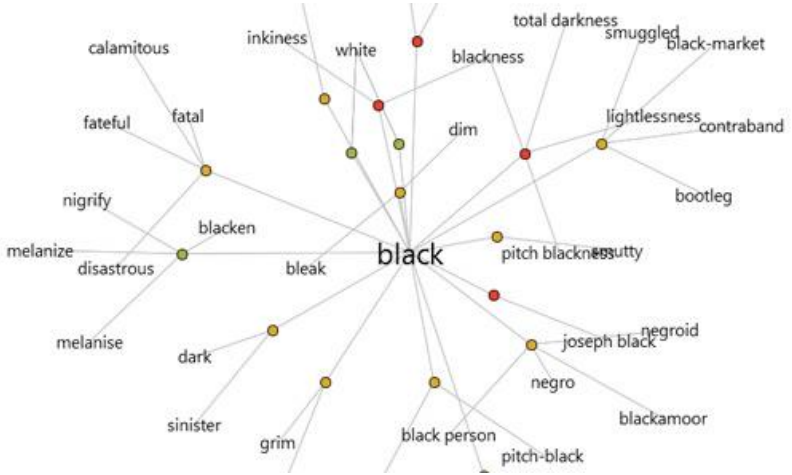
- pretraživanje informacija
- klasifikacija pošte
- pravopisni korektori
- predlozi reči prilikom kucanja

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.korpus.[illegible].ac.rs

Zašto vizuelizacija?

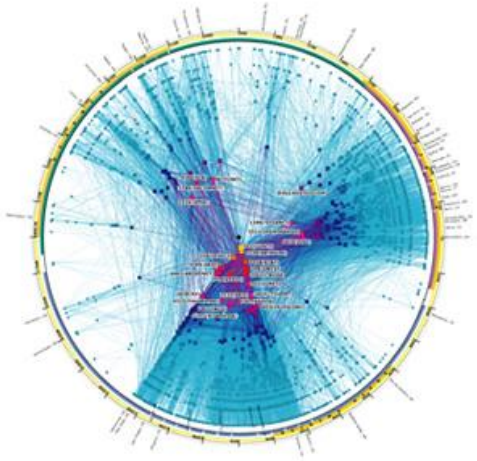
- puno informacija koje treba pročitati i obraditi
- sumarizacija
  - generisanje sažetaka
  - agregacija informacija iz više izvora
- vizuelizacija - brz grafički uvid u strukturu

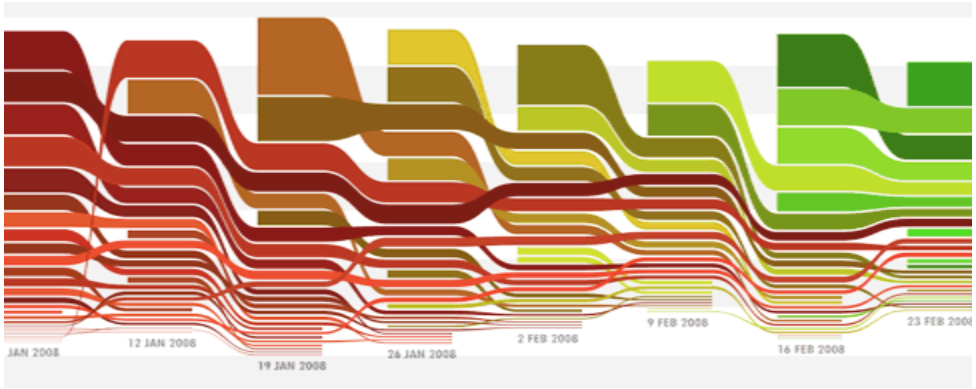
Kako predstaviti dokument?

- 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

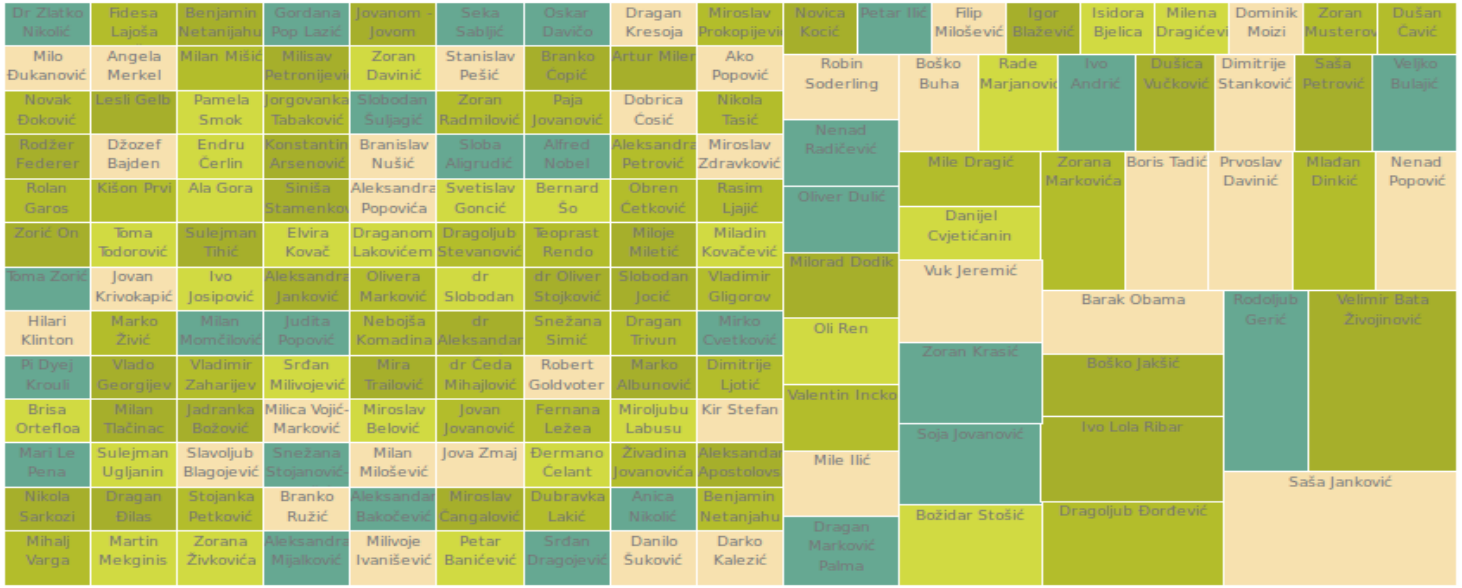
- poseban zadatak klasifikacije dokumenata: odrediti temu/teme zastupljene u tekstu
- pristup:
  - praćenje ključnih pojmova vezanih za neku temu na osnovu frekvencija i statističkih modela jezika
  - algoritmi klasifikacije: logistička regresija, nekoliko najbližih suseda, Naïve Bayes, ...

Vizuelizacija tema

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



Vizuelizacija sentimenata

- odrediti polarisanost, skalu ili dominantnu emociju
- odrediti atribute 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 sentimenata

predsednik u *SrpKor2013* korpusu

Cilj

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

Hvala na pažnji!

andjelkaz@matf.bg.ac.rs