

Sustav za upravljanje i pretraživanje baze PDF dokumenata

Završni rad br. 5672

Luka Čupić

Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Zagreb, 5. srpnja 2018.

Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

1 Uvod

2 Prikaz dokumenata

3 Vizualizacija dokumenata

4 Dodaci

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

Predobrada
dokumenata

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k -srednjih
vrijednosti

Demo 2

Dodaci

- Digitalni zapis informacija danas je sveprisutan
- Potrebno je pronaći način za obradu takvih informacija

Opis problema

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

Predobrada
dokumenata

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k-srednjih
vrijednosti

Demo 2

Dodaci

- Digitalna knjižnica sastavljena od velikog broja dokumenata
- Postojeći dokumenti moraju se moći pretraživati
- Moraju se moći pronaći slični dokumenti novododanim dokumentima
- Dobivene sličnosti bilo bi prikladno vizualno prikazati krajnjem korisniku

Pitanje:

- Kako učinkovito uspoređivati i pretraživati dokumente?

Model dokumenata

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

Predobrada
dokumenata

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k -srednjih
vrijednosti

Demo 2

Dodaci

- Vektorski zapis dokumenata u višedimenzijском prostoru
- Model vreće riječi za zapis riječi dokumenata

Predobrada dokumenata

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

**Predobrada
dokumenata**

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k -srednjih
vrijednosti

Demo 2

Dodaci

Prije prikaza dokumenata, potrebno je iste obraditi:

- Uklanjanje zaustavnih riječi
- Stematizacija riječi
- Pronalazak sličnih riječi, sinonima, antonima, ...

Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

- TF komponenta - riječ je važnija za semantiku dokumenta što se češće u njemu pojavljuje:

$$\text{tf}(w, d) = f_{w,d}$$

- IDF komponenta - riječ je manje važna za semantiku dokumenta što se češće pojavljuje u drugim dokumentima:

$$\text{idf}(w, D) = \log \frac{N}{|\{d \in D : w \in d\}|}$$

Sličnost dokumenata

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

Predobrada
dokumenata

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k -srednjih
vrijednosti

Demo 2

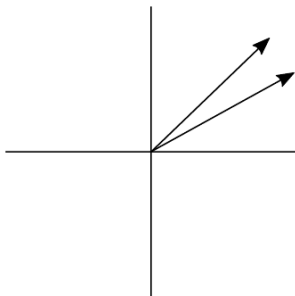
Dodaci

- Mjera sličnosti dokumenata: što dokumenti dijele više riječi, to su sličniji
- Primjer: Ako se u zbirci nalazi dokument o Zvezdanim ratovima, a kao ulazni vektor dovede se fraza poput "May the Force be with you", taj ulazni vektor i taj dokument imati će relativno visoku mjeru sličnosti.

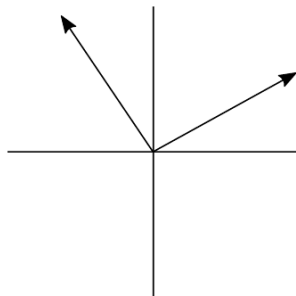
Kosinusna sličnost

- Sličnost dokumenata d_i i d_j :

$$\text{similarity}(d_i, d_j) = \cos(\angle(d_i, d_j)) = \frac{v_{d_i} \cdot v_{d_j}}{\|v_{d_i}\| \cdot \|v_{d_j}\|}$$



(a) Primjer sličnih vektora.



(b) Primjer različitih vektora.

Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

Demonstracija 1: uspoređivanje dokumenata

Kategorije zbirke dokumenata

Uvod

Motivacija

Opis problema

Prikaz dokumenata

Model dokumenata

Predobrada
dokumenata

Sličnost dokumenata

Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova

Grupiranje k -srednjih
vrijednosti

Demo 2

Dodaci

- Arhitektura
- Astronomija
- Biologija
- Filozofija
- Kemija
- Računarska znanost

Vizualizacija dokumenata

Pitanje:

- Kako vizualizirati dobivene rezultate sličnosti dokumenata?

Problem vizualizacije dokumenata

Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

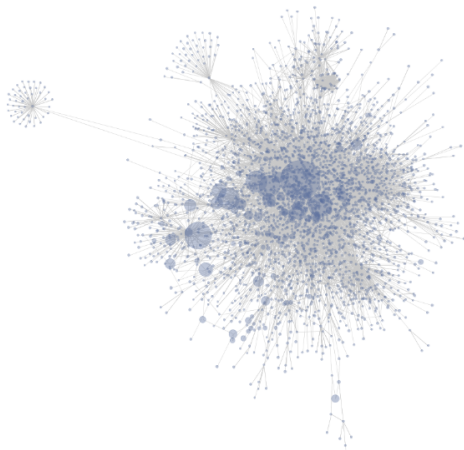
Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

- Problem prevelike dimenzionalnosti
- Problem broja dokumenata

Silom usmjereno crtanje grafova

- Simuliranje privlačnih i odbojnih sila među čvorovima grafa
- Iterativno ponavljanje s ciljem minimizacije energije



Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

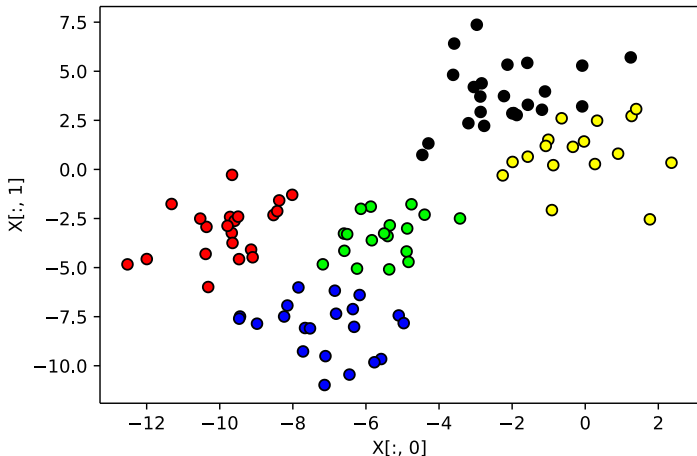
Vizualizacija dokumenata

**Silom usmjereno
crtanje grafova**
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

Grupiranje k -srednjih vrijednosti

- Iterativna dodjela grupa podacima (dokumentima)
- Ponavljanje do ispunjenja (nekog od) uvjeta algoritma



Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti

Demo 2

Dodaci

Demonstracija 2: vizualizacija dokumenata

Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

Dodaci

Slijede dodaci...

Prikaz rezultata

Sustav za upravljanje i pretraživanje baze PDF dokumenata

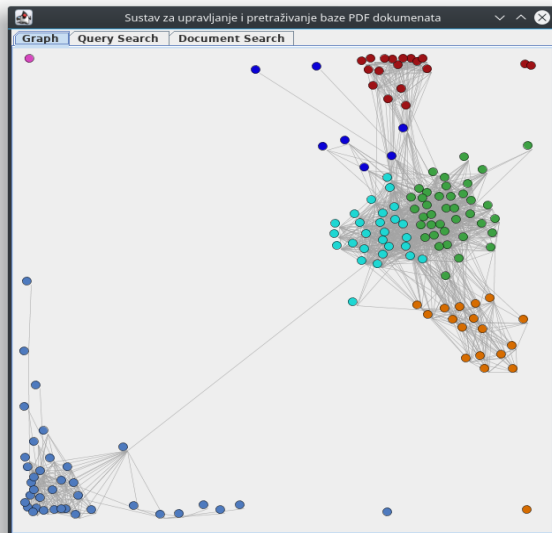
Graph Query Search Document Search

Enter Query

Query Stars are made of hot plasma Search

Document	Similarity
Lectures on Stellar Statistics.pdf	0.10719264822367666
An Illustrated Guide for Amateur Astronomers and a P...	0.10155446312176562
Side-lights on Astronomy and Kindred Fields of Popula...	0.09142932666411072
A Field Book of the Stars.pdf	0.08899235415041451
The Astronomy of Milton's 'Paradise Lost'.pdf	0.0763400322296148
Reactions in Astronomy.pdf	0.06956529098250436
A Text-Book of Astronomy.pdf	0.06789905181877794
Astronomy for Amateurs.pdf	0.05925561582171638
The Story of the Heavens.pdf	0.058468775544588
The Future of Astronomy.pdf	0.049829490599896345
History of Astronomy.pdf	0.04644507379146968
Myths and Marvels of Astronomy.pdf	0.04617056167050128
The Astronomy of the Bible - An Elementary Comment...	0.03814096356132948
A Popular History of Astronomy During the Nineteent...	0.038041233225760744
Astronomy of To-day: A Popular Introduction in Non-Te...	0.03740534054379687
Are the Planets Inhabited?.pdf	0.03155348719300043
Pioneers of Science.pdf	0.02713584524471014
Watchers of the Sky.pdf	0.02634531029396533
Their Nature, Possibilities and Habitability in the Light...	0.015387806115096807
The Uses of Astronomy - An Oration.pdf	0.014042625548950083
The Martyrs of Science, or, The Lives of Galileo, Tycho ...	0.011584497716355677
The gradual acceptance of the Copernican theory of t...	0.009741319806179583
The Chemistry, Properties and Tests of Precious Stone...	0.008226804647355541
On Laboratory Arts.pdf	0.008095323810916072
Darwin and Modern Science.pdf	0.007616842731289242
A Theory of Creation: A Review of 'Vestiges of the Natu...	0.007276031665122302

Vizualizacija dokumenata



Uvod

Motivacija
Opis problema

Prikaz dokumenata

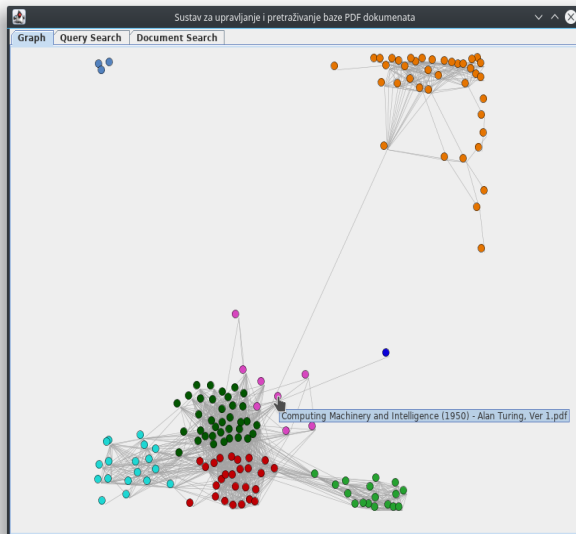
Model dokumenata
Predobrada dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno crtanje grafova
Grupiranje k -srednjih vrijednosti
Demo 2

Dodaci

Zanimljivost s Turingom



Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci



Uvod

Motivacija
Opis problema

Prikaz dokumenata

Model dokumenata
Predobrada
dokumenata
Sličnost dokumenata
Demo 1

Vizualizacija dokumenata

Silom usmjereno
crtanje grafova
Grupiranje k -srednjih
vrijednosti
Demo 2

Dodaci

Hvala na pažnji! Pitanja?