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.

Pregled

Motivacija

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

Opis problema

- 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 dokumenate?

Model dokumenata

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

Predobrada dokumenata

Prije prikaza dokumenata, potrebno je iste obraditi:

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

TF-IDF

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

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

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

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

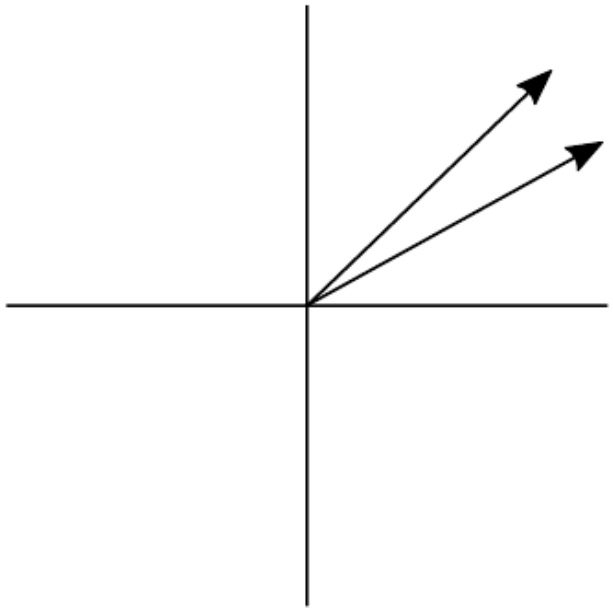
Sličnost dokumenata

- Mjera sličnosti dokumenata: što dokumenti dijele više riječi, to su sličniji
- Primjer: Ako se u zbirci nalazi dokument o Zvjezdanim 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.

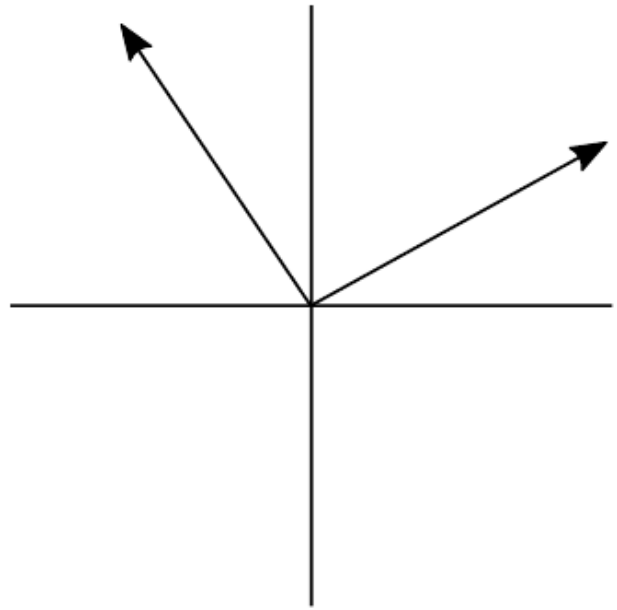
Kosinusna sličnost

- Sličnost dokumenata d_i i d_j:

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



(a) Primjer sličnih vektora.



(b) Primjer različitih vektora.

Demo 1

Demonstracija 1: uspoređivanje dokumenata

Kategorije zbirke dokumenata

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

Vizualizacija dokumenata

Pitanje:

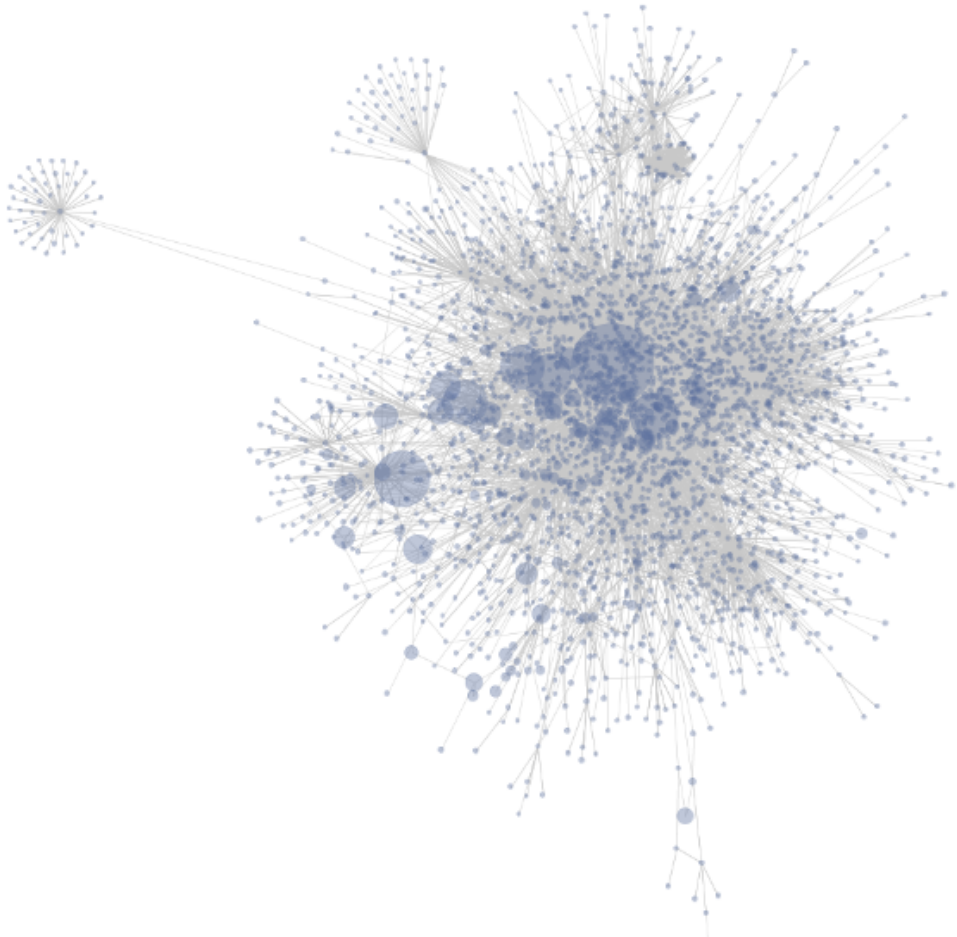
- Kako vizualizirati dobivene rezultate sličnosti dokumenata?

Problem vizualizacije dokumenata

- Problem prevelike dimenzionalnosti
- Problem broja dokumenata

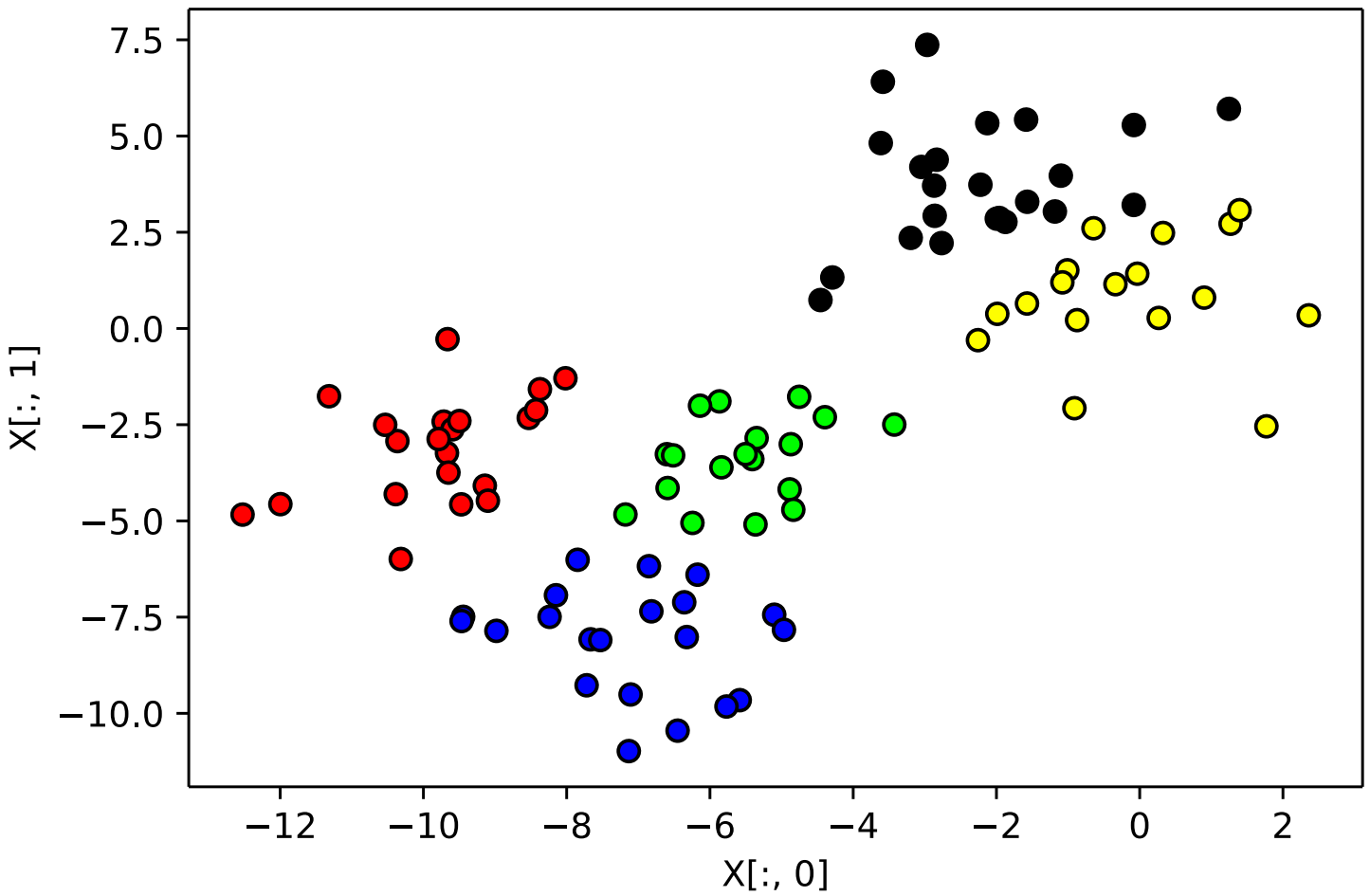
Silom usmjereno crtanje grafova

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



Grupiranje *k*-srednjih vrijednosti

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



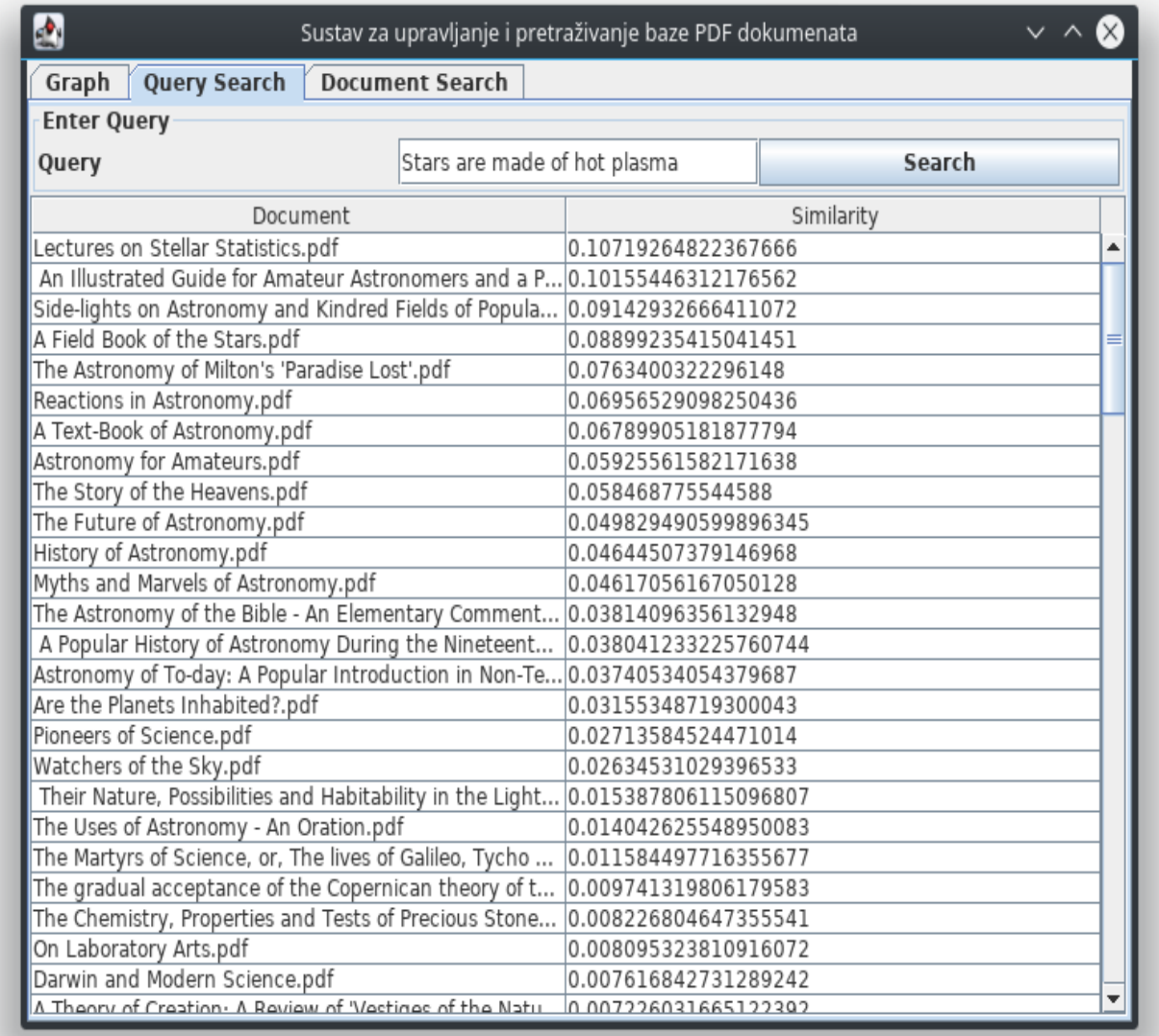
Demo 2

Demonstracija 2: vizualizacija dokumenata

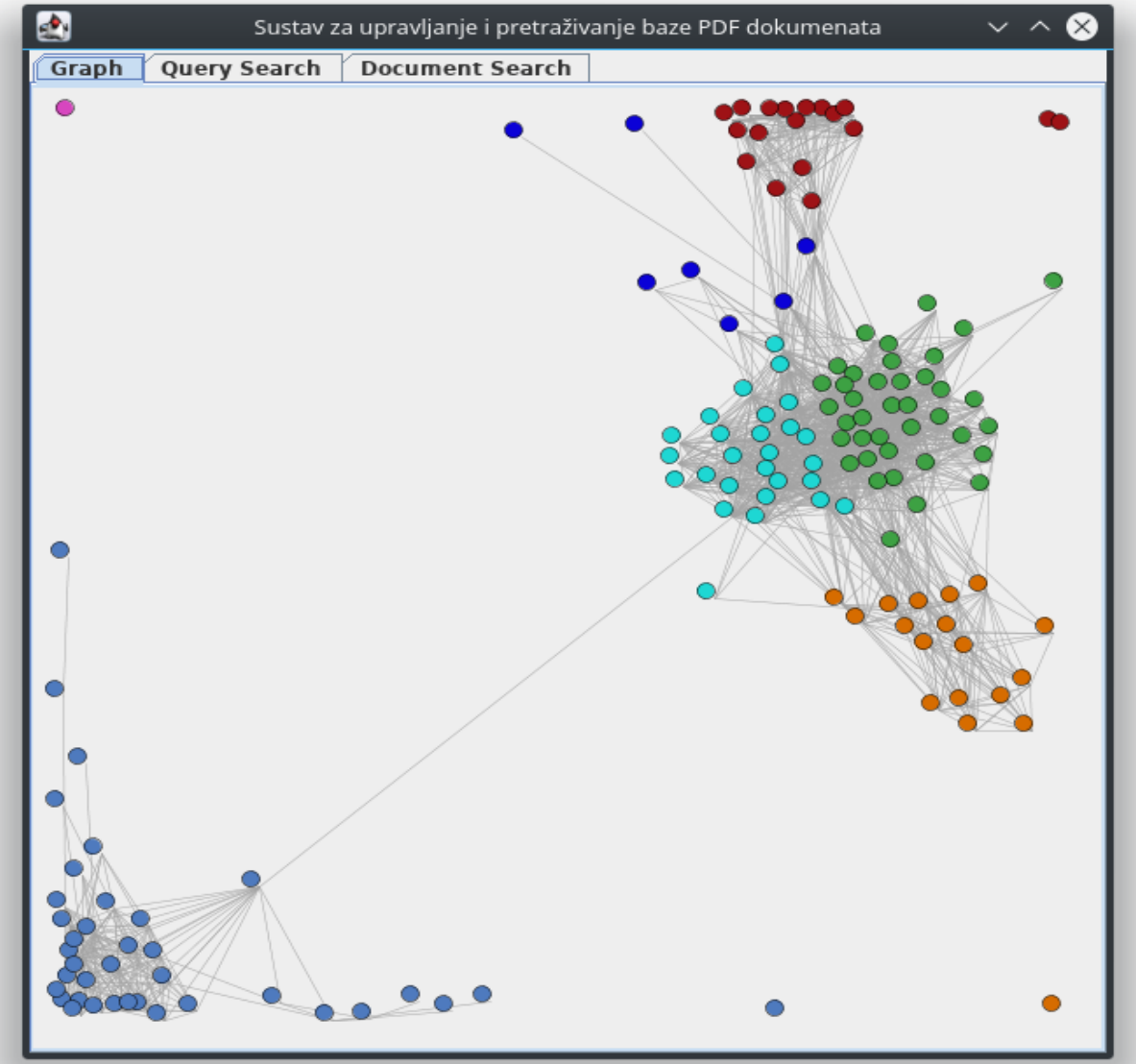
Dodaci

Slijede dodaci...

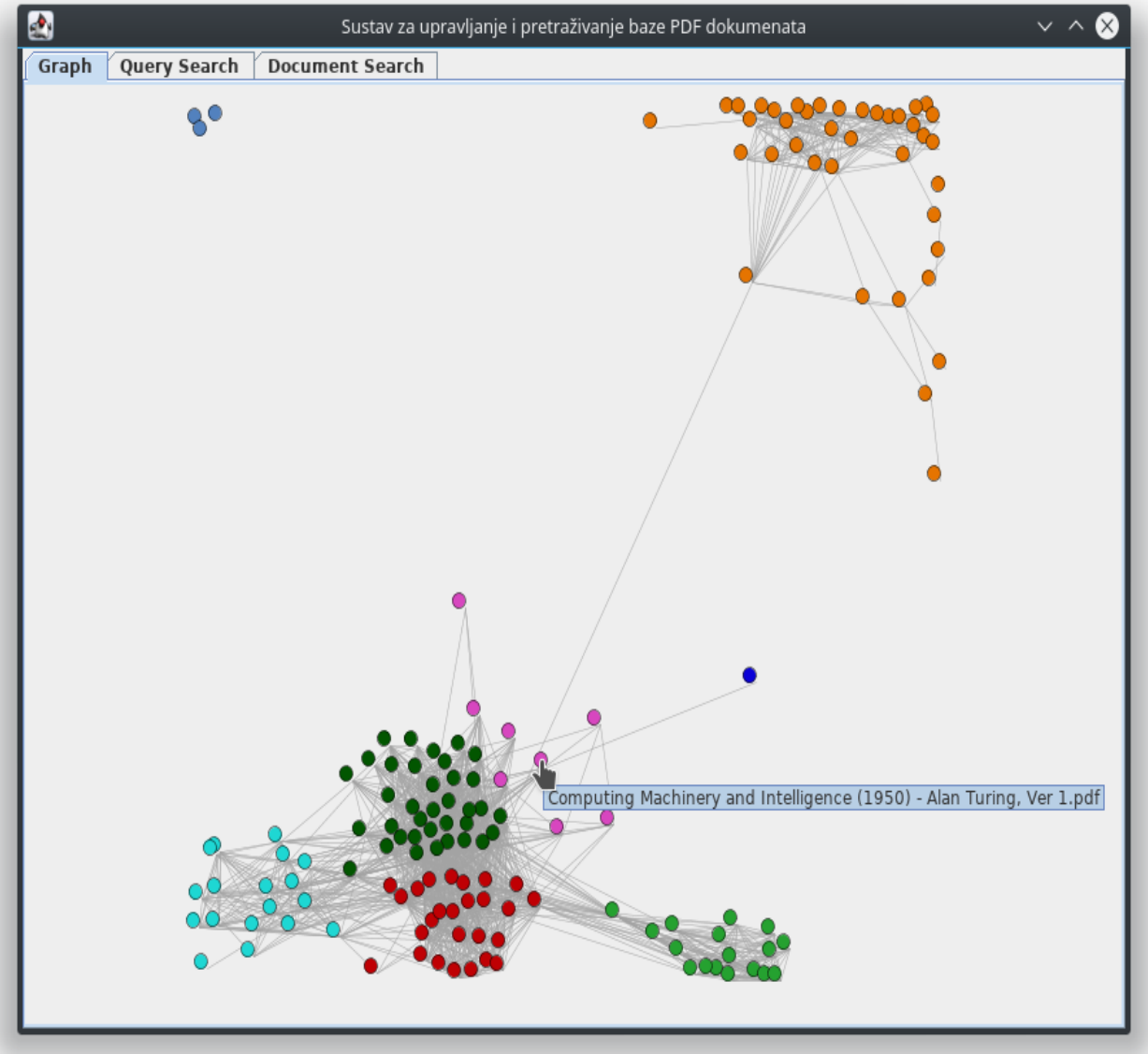
Prikaz rezultata



Vizualizacija dokumenata



Zanimljivost s Turingom



¯_(ツ)_/¯

Hvala na pažnji!
Pitanja?