SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 5672

Sustav za upravljanje i pretraživanje baze PDF dokumenata

Luka Čupić

Zagreb, travanj 2018.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

SADRŽAJ

1. Uvod

Područje analize i pretraživanja teksta neizbježno je u današnjem svijetu tehnologije. Od internetskih tražilica koje pretražuju enormne količine podataka baziranih na zadanom upitu, osobnih pomoćnika na pametnim mobitelima koji procesiraju izgovorene riječi pa sve do analize i prepoznavanja *spam* elektroničkih poruka. Kratki osvrt na ove te mnoge druge primjene ukazuju na nepobitnu činjenicu da je pretraživanje teksta...

2. Pregled područja

Ovaj rad baziran je na proučavanju postojećih metoda analize i pretraživanja teksta. U sklopu ovog rada, biti će proučeno nekoliko metoda za analizu, pretraživanje i usporedbu tekstualnih dokumenata.

3. Let's dive right into it...

Prije *poniranja u dubine*, objasnimo prvo što u kontekstu analize i pretraživanja predstavlja dokument. Neformalno dokument možemo definirati kao kolekciju riječi. Ovakva kolekcija riječi ne mora nužno biti skup, pošto dokument može imati više ponavljanja istih riječi, što će doći do izražaja u **poglavlju X**. Umjesto toga, pretpostavljamo da se dokument sastoji od tzv. *vreće riječi* (engl. *bag of words*) kod koje nam nije bitna semantika samog dokumenta, pa čak niti poredak riječi, već je bitna samo činjenica da se riječi pojavljuju u dokumentu te učestalost njihovog pojavljivanja. Ovakav model često je korišten u području procesiranja prirodnog jezika te povrata informacija kako iz dokumenata tako iz drugih izvora tekstualnih informacija.

Da bi se dokumenti mogli predstaviti u obliku vreća riječi, potrebno je odrediti vokabular—skup svih riječi koje se nalaze u svim dokumentima promatranog skupa dokumenata (u daljnjem tekstu: zbirka). Iz ovako zadanog vokabulara ćemo, međutim, ukloniti sve zaustavne riječi (engl. *stop words*)—riječi koje su učestale u nekom jeziku te su stoga nebitne za sam postupak analize teksta. Primjeri zaustavnih riječi u hrvatskom jeziku su: *aha*, *nešto*, *okolo*, *zaboga*. Nakon stvaranja vokabulara te izbacivanja zaustavnih riječi, možemo krenuti s predstavljanjem dokumenata. Radi praktičnosti, najčešća metoda predstavljanja dokumenata jest vektorska reprezentacija dokumenata. Najjednostavnija metoda vektorskog predstavljanja dokumenata jest binarna: za svaku riječ iz vokabulara naprosto provjerimo nalazi li se u danom dokumentu te ukoliko se nalazi, odgovarajuća komponenta vektora biti će 1, a u suprotnom 0. Nadograđujući se na metodu prethodnu postoji i frekvencijski prikaz vektora. Umjesto obične binarne reprezentacije u kojoj jedinica znači da se riječ nalazi u dokumentu, u frekvencijskom prikazu su komponente vektora zapravo frekvencija (tj. broj) pojavljivanja određene riječi u dokumentu. Naposlijetku dolazimo i do najčešće metode vektorskog prikaza dokumenata—TF-IDF (engl. *term frequency–inverse document frequency*). Ova metoda zasniva se na dvije intuitivne pretpostavke:

– Riječ je važnija za semantiku dokumenta što se češće u njemu pojavljuje (TF komponenta)

– Riječ je manje važna za semantiku dokumenta što se češće pojavljuje po drugim dokumentima (IDF komponenta)

TF i IDF dakle predstavljaju dvije komponente vektora kojima ćemo predstavljati dokumente. Prva komponenta je već spomenuta, frekvencija pojavljivanja riječi *w* u dokumentu *d*—$f_{t,d}$ dok je druga komponenta obrnuta frekvencija pojavljivanja riječi u cijeloj zbirci. Za riječ *w* i dokument *d*, TF komponenta računa se prema sljedećoj formuli: IDF komponenta računa se na sljedeći način:

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

4. Programsko rješenje

Cilj ovog rada jest istražiti već ranije spomenute metode analize i pretraživanja teksta. Kako je implementacija programskog rješenja specifična za dokumente tipa PDF, takve dokumente prvo treba preprocesirati kako bi bili spremni za obradu. U svrhu preprocesiranja PDF dokumenata, koristi se Apache PDFBox...

5. Zaključak

Zaključak.

Sustav za upravljanje i pretraživanje baze PDF dokumenata

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.