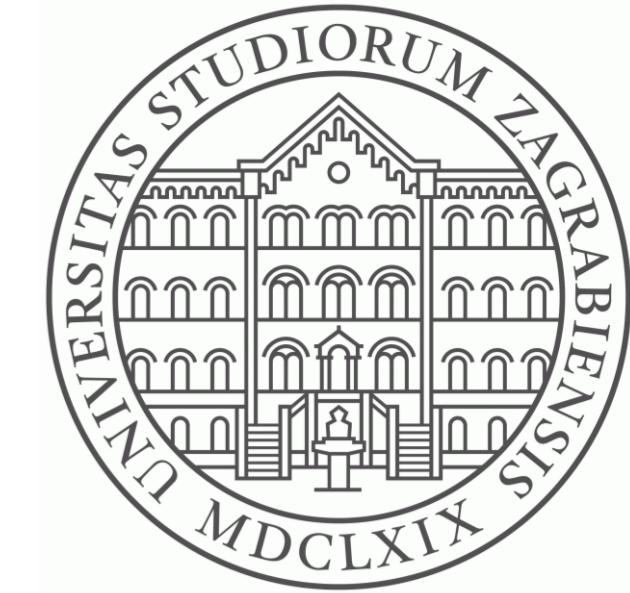


FER



TakeLab

Izgradnja i pretraživanje zbirk često postavljenih pitanja



Mladen Karan

mentor: doc. dr. sc. Jan Šnajder

Sveučilište u Zagrebu Fakultet elektrotehnike i računarstva

1. Uvod

Informacije o nekoj specifičnoj temi često se organiziraju u zbirke često postavljenih pitanja (engl. *frequently asked question – FAQ*). Posebno praktičan slučaj primjene FAQ-zbirki jest strukturiran prikaz znanja o proizvodima i uslugama. Takve su zbirke redovito dio sadržaja na internetu za, između ostalog, banke, telekom-operatere ili javne ustanove. Osim za krajnje korisnike, FAQ-zbirka može biti vrlo korisna i kao pomagalo korisničkoj službi. Tema ovog istraživanja jest razvoj metodologije za učinkovitu izgradnju i pretraživanje domenski specifičnih FAQ-zbirki.

2. Opis problema

Potrebno je napraviti metodologiju za:

- 1) **Izgradnju FAQ-zbirke** – prepostavlja se da imamo skup dokumenata D koji sadrži informacije o nekoj temi i skup korisničkih upita Q . Potrebno je poluautomatski izgraditi FAQ-zbirku koja bi u potpunosti pokrivala informacijske potrebe iz Q koristeći isječke teksta iz D ;
- 2) **Pretraživanje FAQ-zbirke** – potrebno je za zadani korisnikov upit q poredati elemente f_i FAQ-zbirke F tako da se pri vrhu liste nađu elementi koji sadrže odgovor na upit q .

Izazovi tog zadatka jesu:

- 1) Korisnički upiti su vrlo **kratki tekstovi**. Isto vrijedi i za elemente FAQ-zbirke. U kontekstu tehnologije za obradu prirodnoga jezika ovo značajno **otežava** zadatak;

Upit: Ne radi mi mreža.

FAQ: Česti problemi sa spajanjem
na Internet ...

FAQ je relevantan, iako nema niti jednu zajedničku riječ s upitom.

- 2) Cilj je dohvatiti elemente f_i koji **odgovaraju na pitanje**, premda možda nisu leksički **slični** korisničkom upitu;

Upit: Procedura ugradnje ADSL-a?

FAQ1: Procedura otkazivanja ADSL priključka ...

FAQ2: Postupak instalacije ADSL-a ...

**FAQ2 je znatno bolji odgovor na upit, iako FAQ1
ima više zajedničkih riječi s upitom.**

- 3) Analiza teksta i prirodnog jezika općenito je **AI-potpun problem** zbog problema **jezične varijabilnosti** (isti koncept može se izreći različitim riječima) i **višeznačnosti** (ista riječ može označavati više različitih koncepata).

3. Metodologija

Metodologija istraživanja oslanjaće se prvenstveno na modele temeljene na **statističkom strojnom učenju**.

- 1) Za izgradnju zbirke bit će potrebni nadzirani modeli za **grupiranje**, kao što su npr. algoritam K-Means ili algoritam hijerarhijskoga aglomerativnog grupiranja;
- 2) Kod pretraživanja će se, radi veće učinkovitosti, koristiti nadzirani modeli prilagođeni **učenju rangiranja**. Tu ubrajamo varijante stroja potpornih vektora (SVM) i neuronskih mreža prilagođenih analizi teksta (npr. konvolucijske mreže).

4. Rezultati

Napravljena je ispitna FAQ-zbirka i skup od 1200 korisničkih upita s označenim relevantnim odgovorima. Preliminarno su isprobani neki nadzirani modeli rangiranja:

- **BM25** – često korišten i dokazano dobar probabilistički model;
- **tf-idf** – model u kojem su dokumenti točke u vektorskem prostoru;
- **word2vec** – model temeljen na neuronskim mrežama;
- **kombinacija** – zbraja rangove prethodna tri modela.

Preliminarni rezultati			
	MAP	MRR	P@5
BM25	0,166	0,638	0,375
tf-idf	0,140	0,579	0,357
word2vec	0,130	0,588	0,352
Kombinacija	0,150	0,642	0,410

Pokazuje se da nadzirana kombinacija modela radi najbolje. Idući korak je nadziranim strojnim učenjem naučiti model koji je posebno prilagođen da dobro radi na danoj FAQ-zbirki.

5. Zaključak

Tema istraživanja jest izgradnja i pretraživanje FAQ-zbirki. U razvoju rješenja koristit će se prvenstveno statističko strojno učenje. Očekivani doprinos istraživanja jest postupak za izgradnju i pretraživanje domenski specifične FAQ-zbirke. Rezultati istraživanja mogu izravno doprinijeti povećanju kvalitete pretraživanja FAQ-zbirki, kao i povećanju učinkovitosti rada agenata u službama za korisnike.