

Zadanie ku skúške z predmetu Vyhľadávanie informácií, rok 2014

Dokumenty

Číslo	Text Dokumentu (linky sú podčiarknuté, na konci je číslo dokumentu, kde ukazujú)	Číslo	Text Dokumentu (linky sú podčiarknuté, na konci je číslo dokumentu, kde ukazujú)
1	Prezident <u>SR</u> (5) 2014 Top kandidáti abecedne: Ján Čarnogurský, <u>Robert Fico</u> (3), Pavol Hrušovský, Andrej Kiska, Radoslav Procházka. Prvé kolo bude 15., druhé 29. marca 2014	2	Čo po Smere? O tom, že prezidentská kandidatúra Roberta Fica bude znamenať oslabenie <u>Smeru</u> (4), niet sporu... .josef Majchrák 8. január 2014 Vydavateľstvo W PRESS a.s., Partizánska 2, 811 03 Bratislava, Slovensko / ISSN: 1336-653X
3	Robert Fico Doc. JUDr. Robert Fico, CSc. (* 15. september 1964, Topoľčany) je predseda politickej strany <u>SMER</u> (4) a predseda vlády <u>Slovenska</u> (5).	4	SMER SMER-SD, je slovenská ľavicová politická strana. Správy: <u>Čo po SMER-e?</u> (2), <u>Prezident 2014</u> (1)

1. Sťahovače (3b)

- Aká je najlepšia stratégia sťahovania a prečo (kratko)? (1b)
- Opíšte ako môže fungovať FocusedCrawler, ktorý sa rozhoduje o stiahnutí stránky pred jej stiahnutím. (1b)
- Nakreslite orientovaný graf liniek medzi dokumentami a definujte v akom poradí sa stiahnu dokumenty pri vyhľadávaní do šírky a v akom poradí do hĺbky, keď začneme od dokumentu 4 a v rámci stránky sú linky objavené v poradí, akom sa nachádzajú v texte dokumentu. (predpokladajte, že dok. 5 neobsahuje linky) (1b)

2. Textové operácie (5b)

- Vypíšte tokeny a z nich odvodené termy reprezentujúce číselné a časové informácie z dokumentu 2 pomocou inteligentného tokenizátora a analyzéra. (2b)
- Lematizujte text všetkých liniek (podčiarknutý text). (1b)
- Tokenizujte posledný odsek dokumentu 2, tak ako by ste mali tokenizátor ktorý rozpoznáva slová, firmy a adresy. (1b)
- Tokenizujte posledný riadok dokumentu 1 pomocou whitespace tokenizátora. (1b)

3. Indexovanie, váhovanie a podobnosť (8b)

- Vytvorte jednoduchý invertovaný index vyššie uvedených dokumentov. Berte do úvahy iba podčiarknuté slová (slová v linkách) a slová v nadpisoch dokumentov, ostatné vynechajte. Vynechajte stop slová. Slová v rôznych tvaroch berte, akoby boli rovnaké. Indexujte tieto slová aj na miestach, kde nie sú podčiarknuté alebo nie sú v nadpise (1b)
- Vytvorte invertovaný index, kde vezmete do úvahy aj počet výskytu (frekvenciu) termov v dokumente ktorý bude váhou termu, za rovnakých podmienok ako v úlohe a). Anchor text (text liniek) priradte aj tým dokumentom, na ktoré ukazuje, pričom termom z anchor textu dajte pre odkazované dokumenty dvojnásobnú váhu. Váhy nemusíte normalizovať. (3b)
- Vypočítajte váhu termu *SMER* pomocou miery tf-idf (term frequency–inverse document frequency) v dokumentoch 2 a 4. Do úvahy berte celú kolekciu dokumentov (dokumenty 1 až 4) a lematizáciu. (2b)
- Vypočítajte euklidovskú vzdialenosť medzi dopytom *Fico* *SMER* a dokumentmi 3, 4. Váha termov je frekvencia výskytu termov. Uvažujte s lematizáciou. Ignorujte všetky termy/slová z dokumentov 3, 4 okrem termov *Robert*, *Fico*, *SMER*. (2b)

4. Usporiadanie (4b)

- Akým spôsobom je možné kombinovať usporiadania, napr. euklidovskú vzdialenosť a PageRank? Prečiarknite nevyhovujúce a zakrúžkujte vyhovujúce možnosti a zdôvodnite: sčítaním, normalizovaním, násobením, prepočet hodnoty – ak áno aký? (2b)
- Napíšte vzťah pre výpočet PageRank pomocou Google Matice a opíšte jeho princíp. (1b)

- c) Aké iné algoritmy usporiadania pomocou analýzy liniek poznáte? Napíšte aspoň dva a opíšte v krátkosti ich výhody a nevýhody. (1b)

5. **Extrakcia informácií (6b)**

- a) Akých je 5 základných úloh extrakcie informácií? (Definované konferenciami MUC) (1b)
b) Identifikujte názvoslovné entity v dokumente 2 definujte ich typ. (1b)
c) Opíšte, akým spôsobom sa dajú extrahovať mená osôb z textu a extrahujte ich zo všetkých dokumentov. (1b)
d) Preveďte všetky úlohy extrakcie informácií na dokumente 3. (3b)

6. **Regulárne výrazy (5b)**

- a) Opíšte aspoň jeden typ úlohy z oblasti vyhľadávania informácií (iný ako extrakcia informácií), kde je možné použiť regulárne výrazy (regex) a napíšte k nej príslušný regex (1b)
b) Napíšte regexy na vyhľadanie dátumov a ľudí tak, aby boli aj všeobecnejšie použiteľné. Regexy musia správne fungovať minimálne na uvedených textoch. Regexy vysvetlite, uveďte ktoré časti čo extrahujú. Uveďte ktoré entity z uvedených textov extrahujú. (2b)
c) Napíšte regex na extrakciu liniek z HTML (2b)

7. **Hodnotenie (5b)**

- a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (IR - Information Retrieval)? Vymenujte: (1b)
b) Zakrúžkujte a prečiarknite ich vzorce, k zakrúžkovanému napíšte názov miery hodnotenia:
počet získaných/počet všetkých; počet relevantných získaných/počet všetkých;
počet relevantných získaných /počet získaných; počet relevantných získaných /všetkých relevantných;
počet získaných / počet relevantných; (1b)
c) Definujte, aké dokumenty vráti dopyt „Fico AND SMER“ bez a s lematizáciou. (1b)
d) Vypočítajte miery hodnotenia pre IR systém, ktorý pre dopyt „Fico AND SMER“ (s a bez lematizácie) vráti dokumenty 1, 2, 4. (2b)

8. **Sémantický web (5b)**

- a) Opíšte štandard RDF(S) slovne, aj na konkrétnom príklade. Čo je jeho podstatou, na čom je založený? (1b)
b) Z extrakcie informácií dostaneme jednoduché objekty typov ako People (ľudia). Vytvorte graf inštancií týchto objektov získaných v úlohe 5c). Tieto inštanície zároveň majú vlastnosť *fn*, kde je uvedený textový reťazec zistený z dokumentu po lematizácii. Inštanície typu People obsahujú vlastnosti *firstname*, *lastname*. V grafe zakomponujte aj vlastnosť že je niekto kandidát (2b)
c) Napíšte SPARQL dopyt na získanie všetkých inštancií typu People, ktorí sú kandidáti. Výstupom je vlastnosť *firstname* týchto inštancií. Napíšte aj výsledok, ktorý vráti. (2b)

9. **Softvérové knižnice a systémy (5b)**

- a) Opíšte jednotlivo na čo je možné použiť knižnice: Solr, Nutch a Tika. Aké sú ich základné vlastnosti? (1b)
b) Ktoré z knižníc z úlohy a) používajú Lucene? (1b)
c) Opíšte svoj nápad ako by sa dal naprogramovať systém na hľadanie entít reálneho sveta (ľudia, firmy, produkty, lokality, a iné) pomocou spracovania Wikipedie alebo Freebase (3b)

10. **Vyhľadávanie informácií na internete a MapReduce (4b)**

- a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlišil od dovtedy známych systémov pre vyhľadávanie na internete. (1b)
b) V čom sa líši vyhľadávanie na internete od vyhľadávania na disku počítača (1b)
c) Napíšte aspoň dva softvérové systémy postavené nad (alebo využívajúce) Hadoop a opíšte na čo slúžia. (1b)
d) Opíšte princíp MapReduce a uveďte aspoň dve jeho výhody. (1b)