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	Kiska vyhlásil referendum o ochrane rodiny, bude 7. februára - Aliancia za rodinu(2) - Ústavný súd, vyjadrenie(5) SITA | 27.11.2014 14:52, aktualizované: 16:49 Pravda.sk © P E R E X , a. s. 2015	2	Aliancia za rodinu - 420 000 podpisov - Rozhodnutie ÚS SR(5) - Referendum 7.2.2015 - Prieskum(3) Aliancia za rodinu, Gorkého 15, 811 01 Bratislava
3	Prieskum: Referendum o rodine má podporu, ale neistú účasť štvrtok 25. 9. 2014 11:35, zdroj: Focus(4) - Pravdepodobná účasť: 45,5% - proti adopcii homosexuálnymi pármi: 76% SME.sk © Copyright 1997-2015 Petit Press, a.s.	4	FOCUS Aktuality: 23.12.2014 Vianočné zvyky - december 2014 ... Grösslingová 37, 81000 Bratislava 1

1. Sťahovače (3b)

a) Aká je najlepšia stratégia sťahovania: do hĺbky, do šírky, čiastočný PageRank, OPIC? (1b)
b) Ako funguje Focused Crawler: HTTP head request, podľa prípony v URL, podľa URL, sťahuje všetky stránky. (1b)
c) V akom poradí sa stiahnu dokumenty pri vyhľadávaní do šírky _____ a v akom poradí do hĺbky _____, keď začneme od dokumentu 1 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)

a) Aký je počet tokenov reprezentujúcich číselné a časové informácie v dokumente 2, keď máte inteligentný tokenizátor? _____ Uveďte jeden token _____a jeden z neho odvodený term _____. (2b)
b) Lematizujte prvý riadok dokumentu 1: _____ (1b)
c) Tokenizujte posledný riadok dokumentu 1, tak ako by ste mali tokenizátor ktorý rozpoznáva slová, firmy a čísla: _____. (1b)
d) Tokenizujte ten istý riadok pomocou whitespace tokenizátora: _____ (1b)

3. Indexovanie, váhovanie a podobnosť (8b)

a) Tvoríte 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. Koľko slov je v indexe? _____. V ktorých dokumentoch je term súd? _____ (2b)
b) Tvorte 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) priraďte aj tým dokumentom, na ktoré ukazuje, pričom termom z anchor textu dajte pre odkazované dokumenty dvojnásobnú frekvenciu. Akú frekvenciu má term prieskum v dokumente 3? _____ V ktorých dokumentoch je term súd? _____ Ako sú zoradené slová v indexe? _____ Akou dátovou štruktúrou môžeme reprezentovať invertovaný index? _____ (4b)
c) Vypočítajte euklidovskú vzdialenosť medzi dopytom *rodina, focus* a dokumentmi 1, 3. Váha termov je frekvencia výskytu termov. Uvažujte s lematizáciou. Ignorujte všetky termy/slová z dokumentov 1, 3 okrem termov *rodina, focus, prieskum.* Vzorec: _____. Výsledok: _____. (2b)

4. Usporiadanie (5b)

a) Akým spôsobom je možné kombinovať usporiadania, napr. euklidovskú vzdialenosť a PageRank? Možnosti: sčítaním, normalizovaním, násobením, prepočet hodnoty. (2b)
b) Napíšte vzťah pre výpočet PageRank pomocou Google Matice. _____(2b)

c) Aké iné algoritmy usporiadania pomocou analýzy liniek poznáte? Napíšte aspoň dva _____. (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.

_____(2b)

c) Výsledky úloh extrakcie informácií na dokumentov 1 a 2. Udalosť(i):_____,

Entita a alias:_____, entita a vlastnosť: _____ (3b)

6. Regulárne výrazy (6b)

a) Úlohy z oblasti vyhľadávania informácií, kde je možné použiť regulárne výrazy (regex). Možnosti: tokenizácia, obmedzenia sťahovania, indexovanie, usporiadanie, extrakcia informácií. (2b)

b) Napíšte regexy na vyhľadanie: čísel:_____, PSČ:_____,

_____, rokov:_____ (4b)

7. Hodnotenie (6b)

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 *rodina* AND *focus* bez a s lematizáciou: _____.(2b)

d) Vypočítajte miery hodnotenia pre IR systém, ktorý pre dopyt *rodina* AND *focus* (s a bez lematizácie) vráti dokumenty 3, 4. _____(2b)

8. Sémantický web (2b)

a) Opíšte štandard RDF(S): trojice, OWL, ontológie, slovník, taxonómia, graf, objekt, literál? (2b)

9. Softvérové knižnice a systémy (4b)

a) Na čo je možné použiť knižnice: *Solr*, *Nutch* a *Tika*. Spojte čiarou s vlastnosťami:

Indexovanie, konverzia dokumentov, sťahovanie dokumentov, vyhľadávanie? (2b)

b) Ktoré z knižníc z úlohy a) používajú Hadoop? _____(1b)

c) Ktoré z knižníc z úlohy a) používajú Lucene? _____(1b)

10. Vyhľadávanie informácií na internete a MapReduce (5b)

a) Vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlíšil od dovtedy známych systémov pre vyhľadávanie na internete: Invertovaný index, PageRank, spracovanie anchor textov, MapReduce, PC architektúra, AdWords (2b)

b) V čom sa líši vyhľadávanie na internete od vyhľadávania na disku počítača: typ dokumentov, dôveryhodnosť informácie, škálovateľnosť na počet používateľov, na veľkosť dát, hardware failures, sieť prepojení (2b)

c) Napíšte aspoň dva softvérové systémy postavené nad (alebo využívajúce) Hadoop. (1b)

_____