Zadanie ku skúške z predmetu Vyhľadávanie informácií, 17.1.2011

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	Hedviga Malinová má knihu Kniha s názvom Hedviga(4) od Márie Vrabcovej vyšla v slovenskom aj maďarskom jazyku po tritisíc kusov. štvrtok 9. 12. 2010, SME © Petit Press, a.s.	2	Hedviga Malinová • 29.5. 2009 - 5.11. 2010, Kauza Hedviga Malinová v časopise .týždeň(3) • 9. 12. 2010, Hedviga Malinová má knihu(1) (zdroj: SME)
3	.týždeň .kauza: Hedviga Malinová Napadnutie študentky, ktorej sa vlastný štát nezastal, je príbehom o našich právach … 29. máj 2009 Copyright © 2009 - 2011 Vydavateľstvo W Press a.s., Partizánska 2, 811 03 Bratislava, Slovensko / ISSN: 1336-653X	4	Hedviga Marie Vrabcová · LOAR (2010) NA SKLADE Kniha sa usiluje zhrnúť okolnosti útoku na Hedvigu Malinovú a udalosti štyroch rokov, ktoré odvtedy uplynuli, s osobitým dôrazom na spoločenské a politické súvislosti prípadu... Naša cena: 7,12 € (teda ušetríte 11%) © 2000-2010 Martinus.sk

1. Sťahovače (3b)

a) Aká je najlepšia stratégia sťahovania? (0,5b)
b) Ako sa definujú obmedzenia pre sťahovače a aké poznáme obmedzenia? (0,5b)
c) Čo je FocusedCrawler a ako funguje? (1b)
d) Nakreslite orientovaný graf liniek medzi dokumenatmi a definujte v akom poradí sa stiahnu dokumenty pri vyhľadávaní do šírky a v akom poradí do hĺbky, keď začneme od dokumentu 2 a v rámci stránky sú linky objavené v poradí, akom sa nachádzajú v texte dokumentu. (1b)

2. Textové operácie (5b)

a) Čo je tokenizácia? (0,5b)
b) Čo je lematizácia a stemovanie? (0,5b)
c) Tokenizujte všetky čísla v dokumentoch 1 a 4 pomocou inteligentného tokenizátora. (1b)
d) Lematizujte 2. odstavec dokumentu 3 (*Napadnutie* ...). (1b)
e) Tokenizujte dokument 1, tak ako by ste mali inteligentný tokenizátor. Vezmite do úvahy rozdiel medzi tokenmi a termami. (2b)

3. Indexovanie, váhovanie a podobnosť (8b)

a) Vytvorte jednoduchý invertovaný index vyššie uvedených dokumentov. Berte do úvahy iba podčiarknuté slová (slová v linkách) a slová v prvých riadkoch dokumentov, ostatné vynechajte. Slová v rôznych tvaroch berte, akoby boli rovnaké. Indexujte tieto slová aj na miestach, kde nie sú podčiarknuté (napr. *Kniha* v dokumente 4) (1b)
b) Vytvorte invertovaný index, kde vezmete do úvahy aj počet výskytu (frekvenciu) termov v dokumente, 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 váhu. Váhy nemusíte normalizovať. (3b)
c) Vypočítajte váhu termu *Kniha* pomocou miery tf-idf (term frequency–inverse document frequency) v dokumentoch 1 a 2. Do úvahy berte celú kolekciu dokumentov (dokumenty 1 až 4) a lematizáciu. (2b)
d) Vypočítajte kosínusovú mieru medzi dopytom *kauza Malinová* a dokumentmi 2, 3. Použite euklidovskú normalizáciu. Ignorujte číselné termy v dokumentoch 2 a 3. Kolekciu tvoria iba dokumenty 2 a 3. (2b)

4. Usporiadanie (5b)

a) Akým spôsobom je možné kombinovať usporiadania, napr. kosínusovú mieru a PageRank? (1b)
b) Opíšte vlastnými slovami princíp PageRank (1b)
c) Napíšte vzťah pre výpočet PageRank pomocou Google Matice.(1b)
d) Určte Google Maticu a spravte prvú iteráciu. Damping factor je 0,6. Pravdepodobnosť že používateľ odskočí na ľubovoľnú stránku z dangling nódu a personalizačný vektor je (1/4, 1/4, 1/4, 1/4). (2b)

5. Extrakcia informácií (6b)

a) Akých je 5 základných úloh extrakcie informácií? (Definované konferenciami MUC) (1b)
b) Identifikujte názvoslovné entity v dokumente 1 a 3 a 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 1. (3b)

6. Regulárne výrazy (5b)

a) Na aké úlohy sa dajú v oblasti vyhľadávania informácií (information retrieval) použiť regulárne výrazy (regex)? (1b)
b) Napíšte regexy na vyhľadanie sídel (miest a dedín), firiem, PSČ, dátumov, rokov, ľudí a cien v uvedených dokumentoch tak, aby boli aj všeobecnejšie použiteľné. Stačí definovať 3 regexy, musia správne fungovať minimálne na uvedených textoch. Je potrebné napísať regex na dátumy. (4b)

7. Hodnotenie (5b)

a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (IR - Information Retrieval)? (1b)
b) Napíšte ich vzorce. (1b)
c) Definujte, aké dokumenty vráti dopyt *Vrabcová* bez a s použitím lematizácie. Vypočítajte základné miery hodnotenia pre IR systém, ktorý pre dopyt *Vrabcová* (s a bez lematizácie) vráti dokumenty 1, 2. (3b)

8. Sémantický web (5b)

a) Opíšte vlastnými slovami, čo je sémantický web a aké sú jeho ciele. Uveďte základné štandardy pre sémantický web a jeden zo štandardov opíšte. (1b)
b) Z extrakcie informácií dostaneme jednoduché objekty typov ako People (ľudia), Organization (organizácie, média, ...) a Book (knihy). Vytvorte graf inštancií týchto objektov získaných v úlohe 5b). Tieto inštancie zároveň majú vlastnosť *title*, kde je uvedený textový reťazec zistený z dokumentu po lematizácii. Inštancie typu People obsahujú vlastnosti *firstname, lastname*. (2b)
c) Napíšte SPARQL dopyt na získanie všetkých inštancií typu People. Výstupom je vlastnosť *firstname* týchto inštancií. Napíšte aj výsledok, ktorý vráti. (2b)

9. Softvérové knižnice a systémy (3b)

a) Uveďte aspoň 3 softvérové knižnice alebo systémy, ktoré je možné použiť pri vytváraní systémov pre vyhľadávanie informácií. Opíšte ich základné vlastnosti. (1b)
b) Opíšte, aké vlastnosti a časti musí obsahovať systém na vyhľadávanie informácií o ľuďoch v slovenskom webe. Opíšte, pomocou akých softvérových knižníc je možné takýto systém vytvoriť a čo je nutné doprogramovať. (2b)

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

a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlíšil od dovtedy známych systémov pre vyhľadávanie na internete. (1b)
b) Napíšte aspoň dva systémy postavené nad MapReduce a opíšte na čo slúžia. (1b)
c) Opíšte princíp MapReduce a uveďte, aké sú jeho výhody. (1b)
d) Opíšte algoritmus MapReduce pre nejaký konkrétny príklad – word count alebo iné. (2b)