Zadanie ku skúške z predmetu Vyhľadávanie informácií, rok 2013

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	SME • Gašparovič sa rozhodol, že Čentéša nevymenuje(2) Dôvody prezident zverejnil po 15.00 na stránke prezident(3).sk • Gašparovič vyznamenal Trnku z Esetu aj Kočtúcha(5) Vyznamenia dostali aj odporcovia komunistického režimu Krčméry a Jukl(4). © Copyright 1997-2013 Petit Press, a.s.	2	Gašparovič nevymenuje Čentéša 2. január 2013 Prezident SR Ivan Gašparovič dňa 28.12.2012 listom predsedovi Národnej rady SR Pavlovi Paškovi oznámil, že zvoleného kandidáta Jozefa Čentéša do funkcie generálneho prokurátora nevymenuje. 1. augusta 2011 na blogu Sme(1) pod názvom ... Redaktor denníka RSS 5. augusta 2011 konštatoval ...
3	Prezident SR SPRÁVY TLAČOVÉHO ODDELENIA • Prezident nevymenuje generálneho prokurátora(2) 2. január 2013 • Prezident udelil štátne vyznamenania(5) 2. január 2013 • .... © 2005 Kancelária prezidenta SR. WebDesign, analýza prístupnosti a redakčný systém SwiftSite od spoločnosti ELET.	4	Vyznamenania aj Juklovi a Krčmérymu P:3, 01. 01. 2013 19:30, DOM „Vy máte moc, my máme pravdu.“ (Krčméry, jún 1954, Trenčiansky súd) Dve popredné osobnosti tajnej Cirkvi na Slovensku v minulosti Vladimír Jukl a Silvester Krčméry dnes získali najvyššie štátne vyznamenanie Rad Ľudovíta Štúra I. triedy. Ocenenie im udelil prezident Ivan Gašparovič pri príležitosti 20. výročia vzniku SR. ZOZNAM všetkých ocenených(5)

1. Sťahovače (3b)

a) Aká je najlepšia stratégia sťahovania a prečo? (1b)
b) Opíšte ako môže fungovať FocusedCrawler, ktorý sa rozhoduje o stiahnutí stránky pred jej stiahnutím. Čo je jeho hlavnou nevýhodou? (1b)
c) 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 3 a v rámci stránky sú linky objavené v poradí, akom sa nachádzajú v texte dokumentu. (predpokladajte, že dokument 5 neobsahuje linky) (1b)

2. Textové operácie (5b)

a) Vypíšte tokeny a z nich odvodené termy reprezentujúce číselné a časové informácie z dokumentoch 2 a 4 pomocou inteligentného tokenizátora a analýzéra. (2b)
b) Lematizujte text všetkých liniek (podčiarknutý text). (1b)
c) Tokenizujte posledný riadok dokumentu 3, tak ako by ste mali tokenizátor ktorý rozpoznáva slová, produkty, firmy a dátumy. (1b)
d) Tokenizujte posledný riadok dokumentu 1 pomocou whitespace tokenizátora. (1b)

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 nadpisoch dokumentov dlhšie ako dva znaky, ostatné vynechajte. Slová v rôznych tvaroch berte, akoby boli rovnaké. Indexujte tieto slová aj na miestach, kde nie sú podčiarknuté alebo nie sú v nadpise (napr. *prezident* v dokumente 4) (1b)
b) 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) 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 *Čenteš* 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 euklidovskú vzdialenosť medzi dopytom *Ivan Gašparovič* a dokumentmi 1, 2. Váha termov je frekvencia výskytu termov. Uvažujte s lematizáciou. Ignorujte všetky termy/slová z dokumentov 1, 2 okrem termov *Ivan, Gašparovič, Čentéš a Trnka.* (2b)

4. Usporiadanie (4b)

a) Akým spôsobom je možné kombinovať usporiadania, napr. kosínusovú mieru a PageRank? Prečiarknite nevyhovujúce a zakrúžkujte vyhovujúce možnosti a zdôvodnite: sčítaním, normalizovaním, násobením. (2b)
b) 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 tele správy dokumentu 2. (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 firiem, dátumov, údaje o čase a regexy pre vyhľadanie rokov v uvedených dokumentoch 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. (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 *Gašparovič, Čentéš* bez a s lematizáciou.(1b)
d) Vypočítajte miery hodnotenia pre IR systém, ktorý pre dopyt *Gašparovič, Čentéš* (s a bez lematizácie) vráti dokumenty 1, 2, 3. (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) a ich vzťahy. Vytvorte graf inštancií týchto objektov získaných v úlohe 5c). Tieto inštancie zároveň majú vlastnosť *title*, kde je uvedený textový reťazec zistený z dokumentu po lematizácii. Inštancie typu People obsahujú vlastnosti *firstname, lastname*. V grafe zakomponujte aj vlastnosť že bol niekto vyznamenaný(2b)
c) Napíšte SPARQL dopyt na získanie všetkých inštancií typu People, ktorí boli vyznamenaní. Výstupom je vlastnosť *title* 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 spomínaných na webe (ľudia, firmy, produkty, lokality, a iné). (3b)

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

a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlíšil od dovtedy známych systémov pre vyhľadávanie na internete. (1b)
b) V čom sa líši vyhľadávanie na internete od vyhľadávania napríklad v knižniciach (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)