Najčastejšie chyby/výsledky sú uvedené červenou farbou

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)
chýbalo vysvetlenie prečo
nesprávne vysvetlenie že sa zacyklí pri prehľadávaní do hĺbky

b) Opíšte ako môže fungovať FocusedCrawler, ktorý sa rozhoduje o stiahnutí stránky pred jej stiahnutím. Čo je jeho hlavnou nevýhodou? (1b)
Neuvedené nevýhody, napríklad problém objavovania nových liniek na sťahovanie.
Opisovaný Focused Crawler vo všeobecnosti a nie len požadovaný typ Focused Crawlera.

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)
Tokeny majú typ a obsahujú textový reťazec z analyzovaného textu. Termy nemajú typ ale môžu byť reprezentované iným reťazcom ako je v texte (lematizácia a podobne). Teda napríklad termy reprezentujúce dátumy mali byť prevedené na jednotný tvar, najlepšie v tvare YYYY-MM-DD aby sa dalo aj utriediť a robiť range query

b) Lematizujte text všetkých liniek (podčiarknutý text). (1b)
Zaujímavé: správna lematizácia *Sme* na *byť*. Nevyžadovalo sa.

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)
neutriedenie termov podľa abecedy

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)
Nezahrnutie dokumentu 5 do indexu
zlý výpočet váh, nezarátanie dvojnásobnej váhy od anchor textov

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)
Nemal skoro nik správne. Toto je jednoduchá Pytagorova veta v n-rozmernom priestore
kapitola 5.2.1 v učebnici

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)
HITS, OPIC, SALSA ... kapitola 5.2 v učebnici

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)
Často neuvedený typ NE

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)
uvádzame príklady správnych odpovedí nie úplne riešenie ....
Často neuvedený typ NE a zasa niekedy uvádzaný typ NE ako úloha ktorá sa rieši v CO/aliasoch.
Nebol to príliš vhodný text lebo je to dosť nejednoznačné prezident SR môže byť alias (CO) k Ivan Gašparovič ale môže byť aj vlastnosť. Podobne predseda môže byť CO k Paška ale aj vlastnosť.
CO – môže byť aj žiadne alebo Gašparovič je CO k Ivan Gašparovič ak berieme celý text dokumentu 2
Vlastnosti NE – Jozef Čenteš: kandidát, zvolený kandidát, ...
Vzťahy medzi NE – ak je Národná rada rozpoznaná ako NE tak potom NR a Paška je vzťah, NR je správne aj ako vlastnosť

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)
Extrakcia firiem len cez a.s., bolo dobre dať napríklad aj extrakciu pomocou slov ako spoločnosť, firma a za tým meno spoločnosti začínajúce veľkým. Dala sa tým extrahovať firma ELET.

V čase chýbala extrakcia času cez bodku, potom sa neextrahoval čas 15.00.
V dátume chýbala extrakcia mesiacov pomocou názvov mesiacov – alebo aspoň ľubovoľného textu reprezentujúci názov mesiaca, čo stačilo v tomto prípade
Uvedené boli často iba 3 regexy. V tomto teste bolo potrebné spraviť všetky štyri nie ako po minulé roky keď ich bolo vypísaných viac.

7. *Hodnotenie (5b)*

a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (IR - Information Retrieval)? (1b)
Neuvedenie F1
slovenské názvy (nestrhávali sa body len na poučenie): Recall = Úplnosť (preklad pokrytie nevystihuje túto mieru pri všetkých úlohách kde je ju možné použiť), Precision = Presnosť

b) Napíšte ich vzorce. (1b)
Neuvedenie F1

c) Definujte, aké dokumenty vráti dopyt *Gašparovič, Čentéš* bez a s lematizáciou.(1b)
Bez lematizácie: žiadne
S lematizáciou: 1, 2

d) Vypočítajte miery hodnotenia pre IR systém, ktorý pre dopyt *Gašparovič, Čentéš* (s a bez lematizácie) vráti dokumenty 1, 2, 3. (2b)
Bez lematizácie: Precision=0, Recall=0 alebo sa nedá vypočítať – delenie nulou. Niekto by mohol napísať aj že je 100% aspoň pre Recall lebo vráti všetky relevantné. Všetky 3 možnosti sa brali ako správne. Pričom správna je že sa nedá určiť.
S lematizáciou: Precision=2/3=67%, Recall=2/2=1=100%

8. *Sémantický web (5b)*

a) Opíšte štandard RDF(S) slovne, aj na konkrétnom príklade. Čo je jeho podstatou, na čom je založený? (1b)
Opísaný iba RDF a nie RDF(S). RDFS umožňuje navyše napríklad definovať typy objektov a ich inštancie. Často príklad uvedený cez XML, to je fajn, ale lepšie je uvažovať o RDF, RDFS ako o trojiciach vyjadrujúcich hranu v grafe a teda rozmýšľať o dátach ako o grafe/sieti

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)
Nutch neumožňuje indexovanie, toto bolo možné v starších verziách ale teraz využíva Solr. Nemá vyhľadávacie rozhranie lebo využíva Solr. Často nebola spomínaná vlastnosť crawlowania a škálovania pomocou Hadoop.

b) Ktoré z knižníc z úlohy a) používajú Lucene? (1b)
Nutch používa lucene na analýzu dát a predspracovanie, indexovanie je cez Solr
Solr používa lucene na indexovanie aj vyhľadávanie ako aj na analýzu dát.

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)
Tu boli 3 body. Bolo treba napísať viac a porozmýšľať - väčšina tých čo niečo napísali získala 0,5 alebo 1 bod.
Regulárne výrazy nie sú veľmi dobrý nápad keďže chceme aj iné entity ako vymenované ....

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)
PageRank, spracovanie anchor textov liniek, PC architektúra na základe Mooroveho pravidla
MapReduce a kontextová reklama až neskôr

b) V čom sa líši vyhľadávanie na internete od vyhľadávania napríklad v knižniciach (1b)
Dôveryhodnosť informácie, na internete si každý publikuje čo chce. Dostupnosť informácií, objem
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)
Princíp: každá úloha ako Map a Reduce metódy kde vstupom a výstupom sú asociatívne polia. Map alebo sekvencia Map úloh rieši problém nad konkrétnou časťou dát. Reduce kombinuje, spracúva a triedi výsledky z viacerých úloh (musí kopírovať dáta, výstup z Map spracovaný pomocou Reduce by mal byť čo najmenší). Zložitejšia úloha je riešená pomocou sequencie Map a Reduce úloh. Úloha (program) ide k dátam nie opačne ako pri niektorých úlohách.
Výhody: škálovateľnosť; riešený failover; Nie je nutné aby programátor rozumel distribuovanému systému, stačí keď problém naprogramuje ako Map a Reduce metódy; Možnosť ladiť program lokálne. Vhodná architektúra pre spracovanie rozsiahlych dát.