Najčastejšie chyby sú uvedené červenou farbou

Zadanie ku skúške z predmetu Vyhľadávanie informácií 23.1.2009

Dokumenty

Číslo	**Text Dokumentu** (linky sú podčiarknuté na konci je číslo dokumentu kde ukazujú)
1	Rusi zastavili plyn Ruský plynárenský koncern Gazprom(3) zastavil dodávky plynu do Európy. Gazprom obvinil Ukrajinu že kradne z plynu pre Európskych zákazníkov. Európa je znepokojená.
2	Zajtra by mal Gazprom(3) pustiť plyn cez Ukrajinu a tak obnoviť dodávky do Európy. Medzi najviac zasiahnuté krajiny patria Bulharsko, Srbsko a Slovensko. Súvisiace články: • Rusko zastavilo plyn(1)
3	Gazprom 16 Nametkina St., 117997, Moscow, V-420, GSP-7 Email: gazprom@gazprom.ru
4	SME Správy: ▪ 7. 1. 2009 Rusi zastavili plyn do Európy(1) ▪ 12.01.2009 12:50 Gazprom pustí plyn pravdepodobne zajtra (2) Adresa: Petit Press, a.s., Lazaretská 12, 811 08 Bratislava, Slovensko

1. Sťahovače (4b)

a) Aká je najlepšia stratégia sťahovania? (0,5b)
b) Ako sa definujú obmedzenia pre sťahovače a aké obmedzenia? (0,5b)
c) Nakreslite orientovaný graf liniek medzi dokumentmi 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. (3b)

2. Textové operácie (5b)

a) Čo je tokenizácia ? (1b)
b) Čo je Lematizácia a stemovanie? Aký je rozdiel ? (1b)
Lematizer slovnikovy, stemer algoritmicky – nie vždy pravda, nevýplýva z toho rozdiel pre IR systém
c) Lematizujte dokument 1. (1b)
Lematizovanie iba časti dokumentu 1
Chyby v nasledovných slovách – základný tvar prídavného mena musí byť prídavné meno a nie sloveso alebo podstatné meno: Ruský = ruský, Plynárenský = plynárenský, znepokojená = znepokojený
d) Tokenizujte dokument 3, tak ako by ste mali inteligentný tokenizátor. Vezmite do úvahy rozdiel medzi tokenmi a termami (2b)
Chýbali tokeny ako čiarky dvojbodka alebo email{word}
Tokenizovanie iba časti dokumentu 3

3. Indexovanie (7b)

a) Utvorte jednoduchý invertovaný index vyššie uvedených dokumentov, berte do úvahy iba podčiarknuté slová ostatné vynechajte. Slová v rôznych tvaroch berte akoby boli rovnaké. Indexujte podčiarknuté slová aj

v dokumentoch v ktorých podčiarknuté nie sú (napr. plyn aj v dokumente 1) (1b)
chýbalo zoradenie slov podľa abecedy

b) Utvorte invertovaný index kde vezmete do úvahy aj počet výskytu (frekvenciu) termov v dokumente, za rovnakých podmienok ako v úlohe a). Vezmite do úvahy anchor text (text liniek) ktorý patrí aj dokumentom na ktoré ukazuje pričom dajte dvojnásobnú váhu termom odkazujúcim z liniek v dokumentoch na ktoré odkazujú. Váhy nemusíte normalizovať. (4b)
chýbalo zoradenie slov podľa abecedy
zle vykonanie alebo vysvetlenie textových operácií

c) Prečo treba váhy normalizovať? (0,5b)

d) Čo je kosínusová miera a načo slúži? (0,5b)
chybná definícia prebratá z vlaňajšieho testu – nie je to vzdialenosť ale podobnosť. Vzdialenosť vektorov je iná metrika – bralo sa ako ok odpoveď

e) Vypočítajte kosínusovú mieru medzi dopytom „Gazprom plyn“ a dokumentmi 3,4. Vezmite do úvahy frekvenciu termov (2b)

4. *Usporiadanie (5b)*

a) Akým spôsobom je možné kombinovať usporiadania na základe napr. váh termov a PageRank? (1b)

b) Opíšte vlastnými slovami princíp PageRank (1b)

c) Určte Google Maticu a spravte prvú iteráciu. Dumping factor je 0,5. Pravdepodobnosť že používateľ odskočí na ľubovoľnú stránku z dangling nódu a personalizačný vektor je (¼, ¼, ¼, ¼). (2b)
zle vynásobené matice

d) Napíšte vzťah pre výpočet PageRank pomocou Google Matice.(1b)
nevysvetlené premenné

5. *Extrakcia informácií (5b)*

a) Akých je 5 základných úloh extrakcie informácií? (definovane konferenciami MUC) (1b)

b) Identifikujte názvoslovné entity v dokumente 1 a 2 a definujte ich typ. (1b)

c) Preveďte všetky úlohy extrakcie informácií na dokumente 1. (3b)
vzťahy medzi NE ktoré nevyplývali z textu

6. *Regulárne výrazy (5b)*

a) na aké úlohy sa dajú v oblasti vyhľadávania informácií (information retrieval) použiť regulárne výrazy (regex)? (1b)

b) Napíšte regex na vyhľadanie objektov v uvedených dokumentoch tak aby boli aj všeobecnejšie použiteľné. Sídel (miest a dedín), a.s. firiem, lokalít, PSČ, dátumov.
Stačí definovať 3 regexy. (4b)
regexy neextrahovali objekty na dokumentoch v zadaní napr:
firmy – chýbajúce alebo zlé pokrytie dvojslovných názvov. Extrakcia podľa s.r.o. a nie a.s.
lokality – „v”, “pri” namiesto “do” ktoré bolo v texte
PSČ – nepokrytie ruského PSČ kde je 6cifier

7. *Hodnotenie (5b)*

Chýbajúca F1 štatistika

a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (information retrieval) (1b)

b) napíšte ich vzorce (1b)
nevysvetlené premenné

c) definujte aké dokumenty vráti dopyt „Európa“ bez a s použitím lematizácie. Vypočítajte základné miery hodnotenia pre IR systém, ktorý pre dopyt „Európa“ vráti dokumenty 2,3 (3b)

8. *Sémantický web (5b)*

a) Opíšte vlastnými slovami čo je sémantický web, aké sú jeho ciele, uveďte základné štandardy pre sémantický web, opíšte jeden zo štandardov (1b)
neopísaný aspoň jeden štandard

b) Z extrakcie informácií dostaneme jednoduché objekty typov ako Location (geografické miesto), Settlement (sídlo, dedina, mesto). 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 (2b)
c) Napíšte SPARQL dopyt na získanie všetkých inštancií typu Country. Výstupom je vlastnosť „title“ týchto inštancií. (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)
   často uvedené len 2 knižnice alebo systémy alebo neopísane vlastnosti.
b) Opíšte aké vlastnosti, časti musí obsahovať systém na vyhľadávanie v slovenských emailoch vrátane príloh. Opíšte pomocou akých softwarových knižníc je možné takýto systém vytvoriť a čo je nutné doprogramovať. (2b)
   Neuvedené vlastnosti: podpora email standardov, MIME, ...; tiež konverzia príloh na text
   Neuvedené JavaMail alebo iné na prácu s emailom
   Neuvedené konvertory na text pre prílohy napr. PDFBox alebo POI

10. Vyhľadávanie informácií na internete a MapReduce (6b)

a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlíšil od dovtedy známych systémov pre vyhľadávanie na internete (1b)
b) Opíšte princíp MapReduce a nakreslite architektúru (2b)
   neuvedenie DFS v architektúre
c) Opíšte algoritmus MapReduce pre nejaký konkrétny príklad – word count alebo iné (3b)
   väčsina písala tzv. “rozprávky starej matere” namiesto napr. pseudokódu