Najčastejšie chyby sú uvedené červenou farbou – zatiaľ iba k otázkam 2, 3a), 5, 6, 8, 9.

Zadanie ku skúške z predmetu Vyhľadávanie informácií 22.1.2010

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	Slovak Man Takes Hidden Explosive on Dublin Flight Published: January 5, 2010 Explosive traveled undetected through security at Poprad-Tatry Airport in central Slovakia onto a Danube Wings aircraft. The Slovak carrier launched services to Dublin last month. Copyright 2010 The Associated Press(3)	2	RTE News: Explosive reached Ireland after failed test Tuesday, 5 January 2010 23:04 A quantity of explosive, found in a flat on Dorset Street in Dublin this morning, was brought into Ireland following a failed security operation in Slovakia
3	THE ASSOCIATED PRESS Headquarters 450 W. 33rd St. New York, NY 10001 Main Number: +1-212-621-1500	4	BBC News: Slovaks plant explosives on air traveller Page last updated at 02:01 GMT, Wednesday, 6 January 2010
5	.týždeň: Analýza Irish Times Nebezpečenstvu bolo vystavené celé lietadlo. Výbušnina semtex je stabilná iba pri teplotách -5 až 30 °C, tvrdí Tom Clonan, analytik novín a bývalý armádny dôstojník na dôchodku, špecializujúci sa na výbušniny. číslo 01-02/2010 Copyright © 2009 - 2010 Vydavateľstvo W Press a.s., Partizánska 2, 811 03 Bratislava, Slovensko / ISSN: 1336-653X	6	Policajti na letisku schovali výbušniny do batožín. Jedna odletela do Dublinu Aktualizované streda 6. 1. 2010 12:15 O nepodarenej akcii našej polície informovali RTE(2), New York Times(1), BBC(4) a ďalší. SME © Petit Press, a.s.

1. Sťahovače (3b)

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 6 a v rámci stránky sú linky objavené v poradí akom sa nachádzajú v texte dokumentu. (2b)

2. Textové operácie (5b)

a) Čo je tokenizácia ? (1b)
Nie je to rozdeľovanie na termy ale na tokeny. Token može byť aj čiarka, dvojbodka, ... Takéto tokeny sa potom neindexujú ale možu slúžiť pri extrakcii alebo pri tvorbe lepších termov, napríklad na prevedenie dátumov na jednotný tvar alebo na zistenie slov ktoré su v jednej vete a podobne ...
b) Čo je Lematizácia a stemovanie? Aký je rozdiel ? (1b)
Neuvedený rozdiel
c) Lematizujte dokument 6. (1b)
Chyby podobné ako vlani – základný tvar prídavného mena musí byť prídavné meno a nie sloveso; tiež boli zle lemaizované číslovky: Jedna = jeden, aktualizované = aktualizovaný (nie aktualizovať)
Tiež Times = time. – za toto sa nestrhávalo keďže je anglické, ale kto to mal dobre prihliadlo sa.

d) Tokenizujte dokument 3, tak ako by ste mali inteligentný tokenizátor. Vezmite do úvahy rozdiel medzi tokenmi a termami (2b)
   - Chýbali tokeny ako čiarka dvojbodka. Telefónne číslo bol jeden token dokopy aj z textom a dvojbodkou čo je zle. Tel číslo by sa malo upraviť na jednotný tvar.
   - Adresa: „450 W. 33rd St.“ môže byť jeden token ale mal by sa upraviť – nestrhávali sa body.
   - Tokenizovanie iba časti dokumentu.

3. *Indexovanie, váhovanie a podobnosť (8b)*

a) Utvorte jednoduchý invertovaný index vyššie uvedených dokumentov, berte do úvahy iba podčiarknuté slová a slová *Dublin* a *explosive* ostatné vynechajte. Slová v rôznych tvaroch berte akoby boli rovnaké. Indexujte slová aj na miestach kde nie sú podčiarknuté (napr. RTE v dokumente 2) (1b)
   - Termy majú byť s malými písmenami.
   - „New York Times“ alebo „The Associated Press“ nie sú ako jeden term ale ako 3. Mohlo sa uznať za správne pri uvedení vysvetelenia. V takom systéme by ste to podľa „times“ potom nikdy nenašli.

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. (3b)

c) Vypočítajte váhu termu *Dublin* pomocou miery tf-idf v dokumentoch 1 a 2. (berte do úvahy celú kolekciu dokumentov a lematizáciu) (2b)

d) Vypočítajte kosínusovú mieru medzi dopytom *Slovak explosive* a dokumentmi 1,2. Použite euklidovskú normalizáciu. Počet termov v dokumente 1 je 43 a dokumente 2 je 40 termov. (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 (ignorujte vrchol č. 5) a spravte prvú iteráciu. Damping factor je 0,8. Pravdepodobnosť že používateľ odskočí na ľubovoľnú stránku z dangling nódu a personalizačný vektor je (1/5,1/5, 1/5, 1/5, 1/5). (2b)

5. *Extrakcia informácií (6b)*

a) Akých je 5 základných úloh extrakcie informácií? (definovane konferenciami MUC) (1b)
Vymenovanie skratiek nestačí, treba aj slovenský opis.

b) Identifikujte názvoslovné entity v dokumente 5 a 6 a definujte ich typ. (1b)
Stupne, dátumy, ulica, čas sú vlastnosti iných názvoslovných entít (NE) aj keď napr. ulica sa môže brať ako NE. Letisko, policajti nie NE kým nie je pomenované. Napr. „naša polícia“ by sa už mohlo brať ako NE. Meno a Priezvisko je dokopy jedna NE nie dve.

c) Opíšte akým spôsobom sa dajú extrahovať mená osôb z textu (1b)
Odpoveď “pomocou regulárnych výrazov” sa brala za 0b. Bolo treba uviesť aspoň identifikáciu dvoch slov začínajúcich veľkými písmenami a tiež použitie slovníka krstných mien.

d) Preveďte všetky úlohy extrakcie informácií na dokumente 6. (3b)

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), firiem, lokalít, PSČ, dátumov, rokov.
Stačí definovať 3 regexy, musia správne fungovať minimálne na uvedených textoch. (4b)
   - Rok: [20] znamená že vyhľadá jeden znak 2 alebo 0. [19|20] Znamená že hľadá jeden z 5-tich znakov v zátvorke. Medzera na začiatku roku nenájde rok v dokumente 5. Neošetrenie štvorčíslia – nájde aj rok v ISSN.
   - Lokalita: *(v|pri) +(\\p{Lu}\\p{L}+)* je naučené zo starého testu. Nemôže nič nájsť keď je v texte slovo „do“ „do Dublinu“.
   - PSČ: chýbajúce ohraničenie slova napr. pomocou \b. Potom nájde PSČ aj v dátume z dokumentu 2 (podčiarknuté): Tuesday, 5 January 2010 23:04

7. *Hodnotenie (5b)*

a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (information retrieval) (1b)
b) napíšte ich vzorce (1b)
c) Definujte aké dokumenty vráti dopyt „Dublin“ bez a s použitím lematizácie. Vypočítajte základné miery hodnotenia pre IR systém, ktorý pre dopyt „Dublin“ (s a bez lematizácie) vráti dokumenty 2,3,4,6 (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) a Organization (organizácie, média, ...). 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 City. Výstupom je vlastnosť „title“ týchto inštancií. Napíšte aj aký výsledok vráti. (2b)
Naučené SPARQL query z minulých rokov, bez toho aby bol nakreslený graf z úlohy b). Neuvedenie vráteného výsledku.

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)
neopísane vlastnosti knižníc.
b) Opíšte aké vlastnosti a časti musí obsahovať systém na vyhľadávanie v slovenskom webe s následným zobrazením výsledkov na mape podľa geografickej lokality extrahovanej z textu. Opíšte pomocou akých softwarových knižníc je možné takýto systém vytvoriť a čo je nutné doprogramovať. (2b)
Neuvedenie geokódovania ako dôležitej vlastnosti ktorú treba riešiť.
Uvádzanie nezmyselných knižníc a nástrojov ako JavaMail (slúži na prácu s emailami) alebo Beagle (čo je lokálny vyhľadávač založený na lucene s podobnou funkcionalitou ako Google Desktop Search)
Tieto nezmysly boli naučené z minuloročných testov kde mali zmysel.

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) Opíšte princíp MapReduce, uveďte výhody (2b)
c) Opíšte algoritmus MapReduce pre nejaký konkrétny príklad – word count alebo iné (2b)