

## Zadanie ku skúške z predmetu Vyhľadavanie 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 <u>Gazprom(3)</u> zastavil dodávky plynu do Európy. Gazprom obvinil Ukrajinu že kradne z plynu pre Európskych zákazníkov. Európa je znepokojená.
2	Zajtra by mal <u>Gazprom(3)</u> pustiť plyn cez Ukrajinu a tak obnoviť dodávky do Európy. Medzi najviac zasiahnuté krajiny patria Bulharsko, Srbsko a Slovensko.  Súvisiace články: <ul style="list-style-type: none"><li>• <u>Rusko zastavilo plyn(1)</u></li></ul>
3	Gazprom 16 Nametkina St., 117997, Moscow, V-420, GSP-7  Email: gazprom@gazprom.ru
4	SME  Správy: <ul style="list-style-type: none"><li>▪ 7. 1. 2009 <u>Rusi zastavili plyn do Európy(1)</u></li><li>▪ 12.01.2009 12:50 <u>Gazprom pustí plyn pravdepodobne zajtra (2)</u></li></ul> Adresa: Petit Press, a.s., Lazaretská 12, 811 08 Bratislava, Slovensko

### 1. Sťahovače (4b)

- Aká je najlepšia stratégia sťahovania? (0,5b)
- Ako sa definujú obmedzenia pre sťahovače a aké obmedzenia? (0,5b)
- 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)

- Čo je tokenizácia ? (1b)
- Čo je Lematizácia a stemovanie? Aký je rozdiel ? (1b)
- Lematizujte dokument 1. (1b)
- Tokenizujte dokument 3, tak ako by ste mali inteligentný tokenizátor. Vezmite do úvahy rozdiel medzi tokenmi a termami (2b)

### 3. Indexovanie (7b)

- 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)
- 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)
- Prečo treba váhy normalizovať? (0,5b)
- Čo je kosínusová miera a načo slúži? (0,5b)
- 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  $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ . (2b)
- d) Napíšte vzťah pre výpočet PageRank pomocou Google Matice. (1b)

#### **5. Extrakcia informácií (5b)**

- a) Akých je 5 základných úloh extrakcie informácií? (definované 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)

#### **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)

#### **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 „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)
- b) Z extrakcie informácií dostaneme jednoduché objekty typov ako Location (geografické miesto), Settlement (sídlu, dedina, mesto). Vytvorte graf inštancií týchto objektov získaných v úlohe 5b). Tieto inštancie zároveň majú vlastnosť „title“ kde je uvedený textový reťazec získaný 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)
- 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)

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

- a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlišil od dovtedy známych systémov pre vyhľadávanie na internete (1b)
- b) Opíšte princíp MapReduce a nakreslite architektúru (2b)
- c) Opíšte algoritmus MapReduce pre nejaký konkrétny príklad – word count alebo iné (3b)