Zadanie ku skúške z predmetu Vyhľadávanie informácií 24.1.2008

Dokumenty

Číslo	**Text Dokumentu** (linky sú podčiarknuté na konci je číslo dokumentu kde ukazujú)
1	Automobilka KIA(3) sa rozhodla investovať pri Žiline(4), kde vybudovala svoju prvú továreň(2) v Európe.
2	Kia Motors Slovakia, s.r.o. P.O.Box 2, 01301 Teplička nad Váhom Slovakia
3	Kia Motors Corp. www.kiamotors.com First Plant in Europe(2)
4	Mesto Žilina Správy: ▪ 2. 3. 2004 Kia investuje(1) ▪ 30.03.2004 12:50 Kia vytvorila dcérsku spoločnosť(2) Adresa: Mestský úrad v Žiline Nám. obetí komunizmu 1 011 31 Žilina

1. Sťahovače

a) Aká je najlepšia stratégia sťahovania? (1b)
b) Ako sa definujú obmedzenia pre sťahovače a aké obmedzenia? (1b)
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 1 a v rámci stránky sú linky objavené v poradí akom sa nachádzajú v texte dokumentu. (3b)

2. Textové operácie

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

3. Indexovanie

a) Utvorte jednoduchý invertovaný index vyššie uvedených dokumentov, berte do úvahy iba podčiarknuté slová ostatné vynechajte. (1b)
b) Utvorte invertovaný index kde vezmete do úvahy aj počet výskytu (frekvenciu) slov v dokumente, vezmite do úvahy iba podčiarknuté termy.
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) Prečo treba váhy normalizovať? (1b)
d) Čo je kosínusová miera a načo slúži? (1b)

4. Usporiadanie

a) Akým spôsobom je možné kombinovať usporiadania na základe napr. váh termov a PageRank? (1b)
b) Opíšte vlastnými slovami princíp PageRank (1b)
c) Vypočítajte Google Maticu. Dumping factor je 0,5. Pravdepodobnosť že používateľ odskočí na ľubovoľnú stránku z dangling nódu a personalizačný vektor je (¼, ¼, ¼, ¼). (2b)
d) Napíšte vzťah pre výpočet PageRank pomocou Google Matice.(1b)

5. Extrakcia informácií

a) Akých je 5 základných úloh extrakcie informácií? (definovane konferenciami MUC) (2b)
b) Identifikujte názvoslovné entity v dokumente 1 a definujte ich typ. (1b)
c) Preveďte všetky úlohy extrakcie informácií na dokumente 1. (2b)

6. Regulárne výrazy

a) na aké úlohy sa dajú v oblasti vyhľadávania informácií (information retrieval) použiť regulárne výrazy (regex)? (1b)
b) Napíšte regex na vyhľadanie sídel (miest a dedín) v uvedených dokumentoch.
Napíšte regex na extrakciu s.r.o. firiem z textu. Napíšte regex na extrakciu lokalít z textu.
Napíšte regex na extrakciu PSČ z textu. Napíšte regex na extrakciu dátumov z textu.
Stačí definovať 3 regexy. (4b)

7. Hodnotenie

a) Aké sú základné miery hodnotenia systémov pre vyhľadávanie informácií (information retrieval) (1b)
b) napíšte ich vzorce (1b)
c) definujte aké dokumenty vráti dopyt „Žilina“ bez a s použitím lematizácie. Vypočítajte základné miery hodnotenia pre IR systém, ktorý pre dopyt „Žilina“ vráti dokumenty 2,3,4 (3b)

8. Sémantický web

a) Opíšte vlastnými slovami čo je sémantický web, aké sú jeho ciele, uveďte základné štandardy pre sémantický web (1b)
b) Z extrakcie informácií dostaneme jednoduché objekty typov ako Location (geografické miesto), Settlement (sídlo, dedina, mesto). Vytvorte graf inštancií týchto objektov získaných v úlohe 5. Tieto inštancie zároveň majú vlastnosť „title“ kde je uvedený textový reťazec zistený z dokumentu (2b)
c) Napíšte SPARQL dopyt na získanie všetkých inštancií typu Settlement. Výstupom je vlastnosť „title“ týchto inštancií. (2b)

9. Softvérové knižnice a systémy

a) uveďte aspoň 3 softvérové knižnice alebo systémy ktoré je možné použiť pri vytváraní systémov pre vyhľadávanie informácií (1b)
b) Opíšte aké vlastnosti, časti musí obsahovať systém na vyhľadávanie v slovenských mailoch. (2b)
c) Opíšte pomocou akých softwarových knižníc je možné takýto systém vytvoriť a čo je nutné doprogramovať. (2b)

10. Vyhľadávanie informácií na internete

a) Uveďte vlastnosti vyhľadávača Google, ktorými sa už v roku 1998 odlíšil od dovtedy známych systémov pre vyhľadávanie na internete (3b)
b) Čím sa líši prostredie internetu od vyhľadávania v iných prostrediach (ako napr. vyhľadávanie v knižniciach, súboroch na disku, ...) (1b)
c) Opíšte princíp MapReduce (1b)