Exam in course
TDT4215 Web Intelligence
- Solutions and guidelines -
Friday May 20, 2011
Time: 0900-1300

Allowed means of assistance: D No printed or handwritten material allowed. Simple calculator allowed.

The weighting of the questions is indicated by percentages. All answers are to be entered directly in the designated boxes on the question sheet and no additional sheets of paper are to be handed in. For the multiple choice questions, you should set a cross in the box for the correct answer. There is only one correct answer for the multiple choice questions. A list of formulas from the book and the papers/chapters are included at the end of the exam set.

Question 1. IR Models and Indexing (15%)
Assume a document collection with the following four documents:

Document #1:
Several technologies are important for realizing the semantic web. (Length = 0.619)

Document #2:
Technologies that are important for semantic web are amongst other RDF and OWL. (Length = 0.781)

Document #3:
OWL is a language for knowledge representation. OWL has several sub languages. (Length = 0.661)

Assume the vector space model for the questions unless stated otherwise.
a) (2%) Explain the main differences between the Boolean model and the Vector Space Model

- Binary weighting vs. frequency based weighting
- Binary relevance vs. partial relevance
- No ranking vs. continuous ranking
- Boolean queries vs. more flexible queries

b) (4%) Given the documents above, lemmatize the text and remove stop words. Ignore punctuations, uppercase and lowercase characters, they are treated the same. Create an inverted index for the three documents. Use character positions.

Stop words = \{a, be, for, has, other, that, the\}

Documents after lemmatization and stop word removal (not required as part of the answer).
D1 = Several technology important realize semantic web
D2 = Technology that important semantic web amongst rdf owl
D3 = owl language knowledge representation owl several sub language

Inverted index:

<table>
<thead>
<tr>
<th>Terms</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
</tr>
</thead>
<tbody>
<tr>
<td>amongst</td>
<td></td>
<td></td>
<td>54</td>
</tr>
<tr>
<td>important</td>
<td>26</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>knowledge</td>
<td></td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>language</td>
<td></td>
<td>10, 68</td>
<td></td>
</tr>
<tr>
<td>owl</td>
<td>76</td>
<td>1, 48</td>
<td></td>
</tr>
<tr>
<td>rdf</td>
<td>68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>realize</td>
<td>40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>representation</td>
<td></td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>semantic</td>
<td>54</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>several</td>
<td>1</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>sub</td>
<td></td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>technology</td>
<td>9</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>web</td>
<td>63</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

If the student has not ignored the punctuation in the middle of D3, the occurrence of owl, several, sub, and language may have a char position 1 larger.
c) (5%) Assume a vector space model and rank the documents according to the query “semantic owl” (assume no phrases). Use the cosine similarity and show your calculations.

\[ W_{i,q} = (0.5 + 0.5*tf_{i,q}/maxf_{i,q})*\log(N/n) \]
\[ W_{\text{semantic},q} = (0.5 + 0.5*1)*\log(3/2) = 0.176 \]
\[ W_{\text{owl},q} = (0.5 + 0.5*1)*\log(3/2) = 0.176 \]

Query length, \( l_q = \sqrt{0.176^2 + 0.176^2} = 0.249 \)

D1:
- \( tfidf_{\text{semantic}} = 0.176 \)
- \( tfidf_{\text{owl}} = 0 \)

D2:
- \( tfidf_{\text{semantic}} = 0.176 \)
- \( tfidf_{\text{owl}} = 0.176 \)

D3:
- \( tfidf_{\text{semantic}} = 0 \)
- \( tfidf_{\text{owl}} = 0.176 \)

\[
\text{cossim}(D1, q) = (0.176*0.176 + 0*0.176)/(0.619*0.249) = 0.201 \\
\text{cossim}(D2, q) = (0.176*0.176 + 0.176*0.176)/(0.781*0.249) = 0.319 \\
\text{cossim}(D3, q) = (0*0.176 + 0.176*0.176)/(0.661*0.249) = 0.188 
\]

Rank: \( D2 > D1 > D3 \)

Max \( f \) can be 2 for \( D2 \) before stop word removal, then we get: \( D1 > D3 > D2 \)
d) (2%) Assume a Boolean retrieval model and the query “semantic AND (rdf AND NOT(owl))”. What are the relevant documents for the query?

Empty result.

e) (2%) What is the main advantage of block addressing over full inverted indexes? What is the main disadvantage of block addressing over full inverted indexes?

+ smaller index size, smaller pointers since address space is reduced, multiple occurrences collapsed to one inside block

- resolving proximity queries is harder, need to sift through the text, proximity might be found over block boundaries. Text must be available at time of search.

Accepts listing of several advantages too.
Question 2. Text Preparation (5%)

a) (2,5%) Manning, Raghavan and Schütze (Chpt. 2) describe normalization of tokens (equivalence classing of terms). What is normalization, and why do we perform normalization? What is the effect of normalization in a retrieval system?

{USA, U.S.A.} => both map to USA (token normalization)
Map several terms to the same token by e.g. deleting periods from terms, delete hyphens.
Make the system retrieve more documents (e.g. USA vs U.S.A.). Improves recall.

b) (2,5%) What is stemming and lemmatization? Describe the differences. How do the techniques affect retrieval?

Stemming and lemmatization collapse tags to stems and lemmas, respectively (base form of the word).
Stemming:
- generates stems that are not necessarily words.
- suffix stripping
- may collapse to many words

Lemmatization:
- returns the dictionary form of a word (are->be)
- different results depending on if the word is a noun or a verb

Improves recall, might hurt precision.
Question 3. Queries (5%)

a) (2%) Baeza-Yates and Ribeiro (Chpt. 4) mentions two types of context queries, name and describe these two query types.

- Phrase queries – find phrases in the documents
- Proximity queries – find query terms that are found close to each other in the document, relaxed phrase queries

b) (3%) Baeza-Yates and Ribeiro (Chpt. 5) mentions user relevance feedback. Describe relevance feedback and the basic approach.

First part of 5.2

Question 4. Retrieval Evaluation (10%)

a) (5%) Given a set of relevant documents Ra for a given query q, and the returned (ranked) result from an experimental information retrieval engine for query q, Aq, calculate the interpolated precision at 11 standard recall levels. You are not required to graph the result.
Rq = \{D3, D7, D30, D33, D39, D55, D56, D89, D103, D184\}


<table>
<thead>
<tr>
<th>R</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>10%</td>
<td>67%</td>
</tr>
<tr>
<td>20%</td>
<td>67%</td>
</tr>
<tr>
<td>30%</td>
<td>60%</td>
</tr>
<tr>
<td>40%</td>
<td>57%</td>
</tr>
<tr>
<td>50%</td>
<td>43%</td>
</tr>
<tr>
<td>60%</td>
<td>43%</td>
</tr>
<tr>
<td>70%</td>
<td>0%</td>
</tr>
<tr>
<td>80%</td>
<td>0%</td>
</tr>
<tr>
<td>90%</td>
<td>0%</td>
</tr>
<tr>
<td>100%</td>
<td>0%</td>
</tr>
</tbody>
</table>

b) (5%) Baeza-Yates and Ribeiro (Chpt. 3) describe the Harmonic Mean (aka F-Measure) and the E-Measure. Explain briefly the differences between these measures.

3.2.2 in Modern IR
**Question 5. Web Search (15%)**

*a) (3%)* Manning, Raghavan and Schütze described in chapter 20 the purpose of an URL frontier. Describe the purpose of an URL frontier.

Its goals are to ensure that (i) only one connection is open at a time to any host; (ii) a waiting time of a few seconds occurs between successive requests to a host and (iii) high-priority pages are crawled preferentially.

*b) (3%)* Manning, Raghavan and Schütze described in chapter 21 the PageRank algorithm and how it can be specialized to be topic specific. Describe shortly how the PageRank algorithm can be specialized to be topic specific.

Chapter 21.2.3 in Introduction to IR. By teleporting to a random web page chosen non-uniformly. Are then able to derive PageRank values tailored to particular interests. A non-uniformly set of pages can be created manually (e.g. the open directory project).

*c) (4%)* Manning, Raghavan and Schütze described in chapter 21 hubs and authorities. What are hubs and authorities? Describe them shortly and how they can be used as part of the PageRank algorithm.

Chapter 21.3 in Introduction to IR.
d) (5%) Manning, Raghavan and Schütze described in chapter 19 a technique to detect near duplicates of documents. Describe and exemplify how the technique called shingling can be used to detect near duplicates.

Chapter 19.6 in Introduction to IR.
Question 6. Text Mining (5%)

a) (5%) What is text mining? Please give a definition and list at least three examples of text mining technology.

Text mining is to discover new and previously unknown information by computer by automatically extracting information from different written resources. (2p)

Examples of text mining technology include information extraction, topic tracking, text summarization, text classification/categorization, text clustering, etc. (3p)

Question 7. Text Categorization (20%)

a) (4%) What are the two different ways to set up a Naïve Bayes (NB) classifier?

Multinomial Naïve Bayes (multinomial NB model) (2p)
Multivariate Bernoulli Model (Bernoulli model) (2p)
b) (16%) The table below describes a text data set. There are 4 documents in the training set and 1 document in the testing set. In the training set, each document is labeled “Yes” if it is in “Sport” class, otherwise “No”. Use the two different ways to set up a Naïve Bayes (NB) classifier to test whether document 5 can be classified into the class “Sport”. Write down your test results with the detailed calculation process.

<table>
<thead>
<tr>
<th>Sets</th>
<th>Id</th>
<th>Words in documents</th>
<th>c=Sport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>1</td>
<td>football football basketball</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>football football tennis</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>football swim</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>water ocean football</td>
<td>No</td>
</tr>
<tr>
<td>Testing</td>
<td>5</td>
<td>Football football football water ocean</td>
<td>?</td>
</tr>
</tbody>
</table>

For multinomial NB model: (8p)

\[
P(c) = \frac{3}{4}, \ P(\text{not-}c) = \frac{1}{4}
\]

\[
P(\text{football} \mid c) = \frac{5+1}{8+6} = \frac{3}{7}
\]

\[
P(\text{water} \mid c) = P(\text{ocean} \mid c) = \frac{0+1}{8+6} = \frac{1}{14}
\]

\[
P(\text{football} \mid \text{not-}c) = \frac{1+1}{3+6} = \frac{2}{9}
\]

\[
P(\text{water} \mid \text{not-}c) = P(\text{ocean} \mid \text{not-}c) = \frac{1+1}{3+6} = \frac{2}{9}
\]

\[
P(c \mid \text{doc5}) = \frac{3}{4} \times (\frac{3}{7})^3 \times (\frac{1}{14})^2 = 0.0003
\]

\[
P(\text{not-}c \mid \text{doc5}) = \frac{1}{4} \times (\frac{2}{9})^3 \times (\frac{2}{9})^2 = 0.0001
\]

So the test doc5 is classified as in the class “Sport”.

For Bernoulli model: (8p)

\[
P(c) = \frac{3}{4}, \ P(\text{not-}c) = \frac{1}{4}
\]

\[
P(\text{football} \mid c) = \frac{3+1}{3+2} = \frac{4}{5}
\]

\[
P(\text{basketball} \mid c) = P(\text{swim} \mid c) = P(\text{tennis} \mid c) = \frac{1+1}{3+2} = \frac{2}{5}
\]

\[
P(\text{water} \mid c) = P(\text{ocean} \mid c) = \frac{0+1}{3+2} = \frac{1}{5}
\]
\[ P(\text{football} \mid \text{not-c}) = \frac{1+1}{1+2} = \frac{2}{3} \]
\[ P(\text{basketball} \mid \text{not-c}) = P(\text{swim} \mid \text{not-c}) = P(\text{tennis} \mid \text{not-c}) = \frac{0+1}{1+2} = \frac{1}{3} \]
\[ P(\text{water} \mid \text{not-c}) = P(\text{ocean} \mid \text{not-c}) = \frac{(1+1)}{(1+2)} = \frac{2}{3} \]

\[ P(c \mid \text{doc5}) = 0.005 \]
\[ P(\text{not-c} \mid \text{doc5}) = 0.022 \]

So the test doc5 is classified as not in the class “Sport”.

**Question 8. Text Clustering (10%)**

(a) (10%) Manning, Raghavan and Schütze (Chpt. 17) described four different notions of cluster similarity used by the HAC algorithms. Please describe each of them shortly.

See 17, summarized in Fig 17.3. 2,5% for each explained with a similar figure.
Question 9. Ontologies (10%)

a) (4%) Order the following languages, from 1 to 4, according to their expressiveness from weak (weakest indicated by 1) to strong (strongest indicated by 4) semantics.

- OWL = 4
- RDF/S = 2
- XML = 1
- UML = 3

b) (3%) Describe shortly the differences between the current Web and the Semantic Web.

Based around “The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.”

c) (3%) OWL is a language for knowledge representation. OWL has several sub-languages. Describe these different sub-languages and their differences.
OWL Lite supports those users primarily needing a classification hierarchy and simple constraints. It has lower formal complexity than OWL DL.

• OWL DL supports the users who want the maximum expressiveness while retaining computational completeness and decidability.
• OWL Full is meant for users who want maximum expressiveness and the syntactic freedom of RDF; however, with no computational guarantees.

Question 10. Ontology Applications (5%)

a) (3%) In the paper "A Hybrid Approach to Constructing Tag Hierarchies" by Solskinnsbakk & Gulla the concept of tag vector is described. Please describe shortly how such a tag vector is constructed.

- remove markup, stop word removal, stemming
- basic vector construction, add all terms of tagged resource to the tag vector according to the term frequency and adjusting frequencies according to the tagging frequency of the document
- calculate final weight based on tf idf (vectors are used in place of documents in the idf calculation)
- (noramlize vectors)

b) (2%) In the paper "Measuring intrinsic quality of semantic search based on features vectors" by Tomassen and Strasunskas the concept of feature vector is described. Please describe shortly the purpose of these feature vectors.

Connect the domain terminology provided in an ontology and the terminology provided in textual
Appendix A. Formulas

\[
\text{sim}(q,d) = \frac{\sum_{i=1}^{n}(q_i \cdot d_i)}{\sqrt{\sum_{i=1}^{n} q_i^2 \cdot \sum_{i=1}^{n} d_i^2}} = \frac{q \cdot d}{\sqrt{|q| \cdot |d|}}
\]

\[
w_{k,q} = (0.5 + \frac{0.5 \cdot \text{freq}_{k,q}}{\max_i \text{freq}_{i,q}}) \cdot \log \left( \frac{N}{n_i} \right)
\]

\[
w_{i,j} = f_{i,j} / \max_i (f_{i,j}) \cdot \log (N/n_i)
\]

\[
score(c \mid x) = \sum_{d \in \Delta(x)} \text{sim}(x,d) \cdot I(d,c)
\]

\[
sim(d,q) = \frac{P(R \mid d)}{P(R)} = \frac{P(R) \cdot P(d \mid R) \cdot P(R)}{P(d) \cdot P(R)} \approx \frac{P(d \mid R)}{P(d)}
\]

\[
\text{sim}_q(d) = \sum_{i=1}^{n} w_{i,q} \cdot \log \left( \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} \right)
\]

\[
\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{12})(O_{12} + O_{22})(O_{21} + O_{22})}
\]

\[
P(d|c) = P((t_{1}, \ldots, t_{n})|c) = \prod_{1 \leq k \leq n} P(T_k = t_k|c)
\]

\[
P(d|c) = P((e_{1}, \ldots, e_{M})|c) = \prod_{1 \leq l \leq M} P(U_l = e_l|c)
\]

\[
X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}
\]

\[
DR_{i,k} = \frac{p_i(d_k)}{\sum_{j=1}^{n} p_i(d_j)} \quad E\left( P_i(d) \right) = \frac{f_{i,k}}{\sum_{r \in D_k} f_{r,k}}
\]

\[
DC_{i,k} = \sum_{d \in D_k} \left( P_i(d) \log \frac{1}{P_i(d)} \right)
\]