Exam in course
TDT4215 Web Intelligence
- Solutions and guidelines -
Wednesday June 4, 2009
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. A list of formulas
from the book and the papers/chapters are included at the end of the exam set.

Question 1. Vector Space model (20%)

Given the four documents

D1 (document length = 0.68): "The first car did not have a steering wheel. Drivers steered the car with
a lever."

D2 (document length = 1.38): “The automobile is the most recycled consumer product in the world
today.”

D3 : “Car pollution is becoming an increasing problem today.”

D4 (document length = 0.76): “Pollution is a global problem causing climate changes. The car is one
of the contributors to the pollution.”

a) Explain the motivation for the tf.idf formula and each of its components (tf and idf).
b) Calculate the length of document 3. Tokenize the document and remove the stop words (No stemming etc.).

Stop word list: {a, an, did, first, have, in, is, most, not, of, one, the, to, with}

<table>
<thead>
<tr>
<th>Term</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
<th>n</th>
<th>log(N/n)</th>
<th>di=tfidf</th>
<th>di^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>becoming</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.60206</td>
<td>0.60206</td>
<td>0.362476</td>
<td></td>
</tr>
<tr>
<td>car</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.124939</td>
<td>0.124939</td>
<td>0.01561</td>
<td></td>
</tr>
<tr>
<td>increasing</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.60206</td>
<td>0.60206</td>
<td>0.362476</td>
<td></td>
</tr>
<tr>
<td>pollution</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0.30103</td>
<td>0.30103</td>
<td>0.090619</td>
<td></td>
</tr>
<tr>
<td>problem</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.30103</td>
<td>0.30103</td>
<td>0.090619</td>
<td></td>
</tr>
<tr>
<td>today</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0.30103</td>
<td>0.30103</td>
<td>0.090619</td>
<td></td>
</tr>
</tbody>
</table>

\[ \sum di = 1.0124 \]
\[ \sqrt(\sum di) \approx 1.01 \]

Length of document 3 is 1.01

c) Calculate the tf.idf score for the terms “car”, “pollution”, and “wheel” for all four documents.

Car:
\[ \text{Idf} = \log(4/3) \]

D1:
\[ \text{tf}=2 \quad \text{maxf} = 2 \]
\[ \text{tfidf} = 2/2*\log(4/3) = 0.125 \]

D2:
\[ \text{Tf} = 0 \Rightarrow \text{tfidf} = 0 \]

D3:
\[ \text{Tf} = 1, \text{maxf} = 1 \]
\[ \text{tfidf} = 1/1*\log(4/3) = 0.125 \]
d) Calculate the cosine similarity between each of the documents and the query “car pollution wheel” (NOT a phrase query). Rank the documents with respect to the query.

Query:

\[
W_{i,q} = (0.5 + 0.5\times tf_{i,q}/maxf_{i,q})\times\log(N/n)
\]

\[
W_{\text{car},q} = (0.5 + 0.5\times 1)\times\log(4/3) = 0.125
\]

\[
W_{\text{pollution},q} = (0.5+0.5\times 1)\times\log(4/2) = 0.301
\]

\[
W_{\text{wheel},q} = (0.5 + 0.5\times 1)\times\log(4/1) = 0.602
\]

Length of query:

\[
L_q = \sqrt{(0.125^2 + 0.301^2 + 0.602^2)} = 0.68
\]
Document 1:
\[ \text{tfidf}_{\text{car}} = 0.125 \]
\[ \text{tfidf}_{\text{pollution}} = 0 \]
\[ \text{tfidf}_{\text{wheel}} = 0.301 \]

Document 2:
\[ \text{tfidf}_{\text{car}} = 0 \]
\[ \text{tfidf}_{\text{pollution}} = 0 \]
\[ \text{tfidf}_{\text{wheel}} = 0 \]

Document 3:
\[ \text{tfidf}_{\text{car}} = 0.125 \]
\[ \text{tfidf}_{\text{pollution}} = 0.301 \]
\[ \text{tfidf}_{\text{wheel}} = 0 \]

Document 4:
\[ \text{tfidf}_{\text{car}} = 0.062 \]
\[ \text{tfidf}_{\text{pollution}} = 0.301 \]
\[ \text{tfidf}_{\text{wheel}} = 0 \]

\[ \cosSim(d1,q) = \frac{(0.125*0.125 + 0*0.301 + 0.301*0.602)/(0.68*0.68)}{\approx 0.43} \]

\[ \cosSim(d2,q) = \frac{(0*0.125 + 0*0.301 + 0*0.602)/(1.38*0.68)}{= 0} \]

\[ \cosSim(d3,q) = \frac{(0.125*0.125 + 0.301*0.301 + 0*0.602)/(1.01*0.68)}{\approx 0.15} \]

\[ \cosSim(d4,q) = \frac{(0.062*0.125 + 0.301*0.301 + 0*0.602)/(0.76*0.68)}{\approx 0.19} \]

Rank:
D1 > D4 > D3 (D2 is not returned by the query)

**Question 2. Retrieval Evaluation (10%)**

a) A query Q1 has a precision of 0.4, a recall of 0.2, and an answer set of 30 documents. How many relevant documents were returned? How many relevant documents exist for the query?

\[ |R| = \text{Unknown} \]
\[ |Ra| = \text{Unknown} \]
\[ |A| = 30 \text{ docs} \]
\[ p = 0.4 \]
\( r = 0.2 \)

\[ |Ra| = |A| * p \text{ (from the definition of precision)} \]

\[ |Ra| = 30 * 0.4 = 12 \text{ relevant documents in the result set.} \]

\[ |R| = |Ra|/r \text{ (from the definition of recall)} \]

\[ |R| = 12/0.2 = 60 \text{ relevant documents for the query.} \]

b) For a query Q2, the novelty ratio for the query is 0.25, the coverage is 0.5, and the user has the previous knowledge that 30 relevant documents exist for the query. How many new documents were revealed to the user?

Coverage = \( |Rk|/|U| \) (\( Rk \) = retrieved relevant documents known to the user, \( U \) = relevant documents known to the user)

\[ |Rk| = \text{unknown} \]

\[ |Rk| = c * |U| \]

\[ |Rk| = 0.5 * 30 = 15 \]

Novelty = \( |Ru|/(|Ru|+|Rk|) \) (\( Ru \) = retrieved relevant documents previously unknown to the user)

\[ |Ru| = \text{unknown} = x \]

\[ 0.25 = x/(x+15) \]

\[ x = 0.25x + 0.25*15 \]

\[ 0.75x = 0.25*15 \]

\[ x = 0.25*15/0.75 = 5 \]

The search revealed 5 new relevant documents to the user.

**Question 3. Indexing (10%)**

a) Inverted files

Assume a document collection that consists of the three first documents (D1, D2 and D3) from question 1 and the stopword list from question 1b). Lemmatize the documents and remove the stopwords. Construct an inverted file (vocabulary and occurrences) for the document collection. Use block addressing with blocks equal to documents.
Lemmatized documents with stopwords removed:

D1: car, steer, wheel, driver, steer, car, lever
D2: automobile, recycle, consumer, product, world, today
D3: car, pollution, become, increase, problem, today

Inverted file for documents D1, D2 and D3:

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automobile</td>
<td>2</td>
</tr>
<tr>
<td>Become</td>
<td>3</td>
</tr>
<tr>
<td>car</td>
<td>1, 3</td>
</tr>
<tr>
<td>Consumer</td>
<td>2</td>
</tr>
<tr>
<td>driver</td>
<td>1</td>
</tr>
<tr>
<td>increase</td>
<td>3</td>
</tr>
<tr>
<td>Lever</td>
<td>1</td>
</tr>
<tr>
<td>Pollution</td>
<td>3</td>
</tr>
<tr>
<td>Problem</td>
<td>3</td>
</tr>
<tr>
<td>product</td>
<td>2</td>
</tr>
<tr>
<td>Recycle</td>
<td>2</td>
</tr>
<tr>
<td>steer</td>
<td>1</td>
</tr>
<tr>
<td>today</td>
<td>2, 3</td>
</tr>
<tr>
<td>Wheel</td>
<td>1</td>
</tr>
<tr>
<td>world</td>
<td>2</td>
</tr>
</tbody>
</table>

b) Signature files

Explain briefly how signature files work for text indexing.

A signature file uses a hash function (or ‘signature’) that maps words to bit masks of B bits. It divides the text in blocks of b words each. To each text block of size b, a bit mask of size B will be assigned by ORing the bit masks of all the words of the block.

The idea is that if a word is present in a text block, then all the bits set in its signature are also set in the bit mask of the text block.

Searching a single word is carried out by:

– hashing it to a bit mask \( W \),
– comparing the bit mask \( B_i \) of all the text blocks.
– if \( (W \& B_i = W) \), the text block may contain the word.
c) What is a false drop in signature files? Show an example of a false drop.

A false drop occurs if all the bits of a bit mask of a word \( w \) is set for a block, even though the block does not contain \( w \).

Example: Block B: a b c

Bit masks:
- \( h(a) = 101010 \)
- \( h(b) = 000011 \)
- \( h(c) = 001001 \)
- \( h(d) = 001011 \)
- \( h(B) = 101011 \)

\( h(d) \& h(B) = h(d) \), which means that B may contain \( d \). As seen above, however, \( d \) is not included in B.

**Question 4. Text Categorization (15%)**

a) Explain briefly the principles of Rocchio text classification

Page 269 in Introduction to Information Retrieval, chapter 14.

b) Discuss the differences between Rocchio and k nearest neighbor (kNN). Show how the two methods may produce different classification results!

Page 269-275 in Introduction to Information Retrieval, chapter 14.
**Question 5. Clustering (15%)**

a) Assume the following document vectors: \( d_1 = (1,1), d_2 = (2,1), d_3 = (4,3), d_4 = (5,4) \). We want to cluster these 4 documents into \( k=2 \) clusters using the k-means clustering method. Use \( d_1 \) and \( d_2 \) as the initial cluster centroids \( c_1 \) and \( c_2 \), run two iterations of the k-means method, and use cosine similarity for allocating documents to clusters. Show the calculations.

**1st iteration:**

No calculations needed. \( d_1 \) is closest to centroid \( c_1=(1,1) \), all the other documents are closest to \( c_2=(2,1) \). This conclusion is wrong. Euclidean distance has been assumed. In first iteration \( d_1,d_3, \text{and } d_4 \) are closest using cosine. This stabilizes after 1st iteration.

**Modifications of centroids:**

- \( c_1 \) stays the same
- \( c_2 = \frac{(2+4+5)/3,(1+3+4)/3}{(11/3,8/3)} = (3.7,2.7) \)

**2nd iteration:**

\( d_1 \) is closest to \( c_1 \). \( d_3 \) and \( d_4 \) are obviously closer to \( c_2 \).

\[
\text{Sim}(d_2,c_1) = \frac{2+1}{\sqrt{5}\sqrt{2}} = \frac{3}{\sqrt{10}} \\
\text{Sim}(d_2,c_2) = \frac{2*3.7+1*2.7}{\sqrt{5}\sqrt{13}} = \frac{11}{\sqrt{65}}
\]

\( d_2 \) is closest to \( c_1 \).

We then have two clusters: \( C_1 = \{d_1, d_2\}, C_2 = \{d_3, d_4\} \)

**Resulting centroids:**

\[
c_1 = \frac{(2+1)/2, (1+1)/2}{(1.5, 1)} \\
c_2 = \frac{(4+5)/2,(3+4)/2}{(9/2,7/2)} = (4.5,3.5)
\]

b) What is the time complexity of the k-means clustering algorithm?

- [ ] \( O(n^2) = \text{TRUE} \)
- [ ] \( O(n^2 \log n) \)
- [ ] \( O(kn^2) \)
- [ ] \( O(kn) \)
- [ ] \( O(k^2n) \)

**Question 6. Semantic Web (15%)**
a) Jasper & Uschold presents a framework for ontology classification. What are the four ontology scenarios in the 'common access to information' category?

b) Given that the vocabularies used are strictly partitioned (also known as type separation), what is the correct language inclusion relationships among the ontology languages listed below (one is correct)?

- OWL Lite $\subset$ OWL DL $\subset$ OWL Full
- RDF(S) $\subset$ OWL Lite $\subset$ OWL DL - yes
- OWL Lite $\subset$ OWL DL $\subset$ RDF(S)
- RDF(S) $\subset$ OWL DL $\subset$ OWL Full
- XML $\subset$ OWL DL $\subset$ OWL Full

c) What is the range of a property $P$ in RDF?

- The class of those resources that may appear as subjects in a triple with predicate $P$
- The class of those resources that may appear as values in a triple with predicate $P$ - yes
- The set of instances of $P$
- The set of subproperties of $P$
- The number of instances of $P$

d) What cannot be expressed in RDF Schema (tick one box)?
Question 7. OWL Modeling (15%)

a) What is the relationship between individuals and classes in OWL?

☐ An individual is related to a class by means of an object property

☐ An individual is related to a class by means of a data property

☐ An individual is an instance of one or more classes – **yes**

☐ An individual is an instance of exactly one class

☐ An individual is a specialization of a class

b) What does it mean that a property is functional?

☐ The property has exactly one value for each object

☐ The property has at most one value for each object - **yes**

☐ Two different objects for the property cannot have the same value

☐ All objects for the property must have the same value

c) Below is an OWL definition of administrative staff, serialized in XML. Write the definition in natural language.

```
<owl:Class rdf:ID="adminStaff">
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#staffMember"/>
    <owl:Class>
      <owl:complementOf>
        <owl:Class>
          <owl:unionOf rdf:parseType="Collection">
            <owl:Class rdf:about="#faculty"/>
```
Administrative staff is a member of the staff that is neither faculty nor support staff.

d) Below is an OWL definition of cat owners and old ladies, described using the abstract syntax from Antoniou et al.

Class(OldLady complete
   Person Elderly Female

Class(OldLady partial
   restriction(hasPet someValuesFrom(owl:Thing))
   restriction(hasPet allValuesFrom(Cat)))

Class(CatOwner complete
   Person
   restriction(hasPet someValuesFrom(Cat)))

Phrase these definitions in natural language

Old ladies are elderly female persons. All old ladies must have a pet. All old ladies can have only cats as pets. A cat owner is any person that has a cat as a pet.

e) What can we infer about the relationship between old ladies and cat owners from the definitions in d) (multiple answers may be correct)?

- Some old ladies are cat owners
- All old ladies are cat owners = true
- Old ladies and cat owners are equivalent classes
- All cat owners are old ladies
- Some cat owners are old ladies = true

f) Translate the following definition of international students into the OWL abstract syntax:
“An international student is any student who is following only the Erasmus program and/or the International Master program.”

Class(InternationalStudent complete
Student
restriction(follows
  allValuesFrom(union of(ErasmusProgram InternationalMasterProgram))))

g) Translate the description of web intelligence courses into the OWL abstract syntax:

“Web intelligence courses are taught by Professor Gulla and have at least some good students”

Class(WebIntelligenceCourse partial
  restriction(isTaughtBy hasValue(ProfGulla))
  restriction(isAttendedBy someValuesFrom(GoodStudent)))

### Appendix A. Formulas

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

\[
score(c | x) = \frac{\sum_{d \in \text{NN of } x} \text{sim}(x,d) \cdot I(d,c)}{\sum_{d \in \text{NN of } x} \text{sim}(x,d)}
\]

\[
sim(d,q) = \frac{P(R|d)}{P(R|d)} = \frac{P(R) \times P(d|R)}{P(R) \times P(d|R)} = \frac{P(d|R)}{P(R)}
\]

\[
sim(dq) \approx \sum_{q \in \text{NN}} W_q \times W_{dq} \times \left( \log \left( \frac{P(k_i | R)}{1 - P(k_i | R)} \right) + \log \left( 1 - \frac{P(k_i | R)}{P(k_i | R)} \right) \right)
\]

\[
s = \frac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}
\]

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

\[
DR_{t,k} = \frac{P(i|D_k)}{\sum_{j=1}^{m} P(i|D_j)}
\]

\[
E(P(i|D_k)) = \frac{f_{i,k}}{\sum_{t \in D_k} f_{t,k}}
\]

\[
DC_{t,k} = \sum_{d \in D_t} \left( P_t(d) \log \frac{1}{P_t(d)} \right)
\]