Exam
TDT4215 Web Intelligence
Monday June 4, 2012 * 0900-1300
Course taught in English, so no Norwegian text

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

Tasks are given a weight in percent of total score. All answers should be entered directly in the designated boxes on the question sheet and no additional sheets of paper are to be handed in.

Question 1. Vector space models (20%)

Assume a document collection with the following four documents:

**Document #1:**
"The first car did not have a steering wheel. Drivers steered the car with a lever." (length = 0.68)

**Document #2:**
"The automobile is the most recycled consumer product in the world today." (length = 1.38)

**Document #3:**
"Car pollution is becoming an increasing problem today." (length = ?)

**Document #4**
"Pollution is a global problem causing climate changes. The car is one of the contributors to the pollution." (length = 0.76)

a) (5%) Explain the components, motivation and assumptions behind this variant of the tf.idf formula:

\[ w_{ij} = \begin{cases} 
(1 + \log_2 f_{ij}) \left( \log_2 \frac{N}{n_i} \right) & \text{for } f_{i,j} > 0 \\
0 & \text{otherwise} 
\end{cases} \]

*for both documents and queries (both represented by j)*
b) (5%) Tokenize the document and remove the following stop words: {a, an, did, first, have, in, is, most, not, of, one, the, to, with}. Calculate the length of document 3. (Don’t do stemming or other preprocessing.)

c) (5%) Calculate the tf.idf-score for the terms “car”, “wheel” and “pollution” for all documents.
d) (5%) Explain and compare the properties of different distance measures in the vector model.

Question 2. Text preprocessing (10%)

a) (5%) Medical records usually contain many misspellings, mixed languages (english, scandinavian, latin), abbreviations, synonyms (eg. different brand names for tests and drugs, but with exactly the same meaning). What preprocessing methods would you suggest for such text? Why?
b) (5%) For a selection of the preprocessing techniques you know, compare their properties, strengths and weaknesses.

Question 3. Queries and semantic web (10%)

a) 5% Which types of semantic knowledge can be used for query processing. Referring to the project, how would this be relevant for treating patient record text as queries to a corpus of clinical guidelines?
b) (5%) Explain the concept of relevance feedback. How could it be used for the project?

**Question 4. Retrieval evaluation (20%)**

a) (5%) Using a Venn-diagram, explain the terms Answer set for a query a (A), Relevant documents (R), Relevant documents in the answer set for a (Ra), and Document collection (D). Give formulas for precision and recall.

b) (10%) Given a set of relevant documents Rq for a given query q, and the returned (ranked) result Aq from a retrieval engine for query q, calculate the interpolated precision at 11 standard recall levels and draw the corresponding graph.

Rq = \{D_3, D_5, D_9, D_{25}, D_{39}, D_{56}, D_{71}, D_{89}, D_{123}\}

c) (5%) What are the measures Harmonic Mean (aka F-Measure) and the E-Measure. Explain briefly the utility and difference between these, and precision and recall. When are precision and recall inappropriate measures?
Question 5. Text clustering (15%)

a) (5%) Describe properties of different text clustering methods.

b) (5%) How can clustering be used to improve the search experience?
c) (5%) Explain in detail the two-step, batchwise, K-means clustering algorithm:
Question 6. Text categorization (10%)

a) (5%) Discuss the differences between Rocchio and kNN categorization.

b) (5%) Give an example where the Rocchio and kNN categorization will yield different results.
Question 7. Ontologies (15%)

a) (5%) Translate the following OWL-statements into prose:

Class: HogwartsHouse
  EquivalentTo: { Gryffindor, Slytherin, Ravenclaw, Hufflepuff }

Class: HogwartsStudent
  SubClassOf: hasEmail some string
  SubClassOf: hasPet max 1

ObjectProperty: hasPet
  Inverses: isPetOf

Class: Phoenix
  SubClassOf: isPetOf only Wizard

b) (5%) What are the main properties of the OWL language(s).
c) (5%) What are the differences between reasoning with a theory in OWL, a logic program (Prolog) or a database system? What properties of a knowledge representation can be controlled?