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. Vector Space model (25%)**

Assume a document collection with the following four documents:

**Document #1:**
Trondheim is the third largest city in Norway and is situated on the Nidelva river. (Length = 0.89)

**Document #2:**
Trondheim is a university city with many students.

**Document #3:**
The Nidaros cathedral is a popular tourist attraction. (Length = 1.13) Correct length: 1.28

**Document #4:**
Nidaros was the original name of Trondheim. Trondheim is the modern name. (Length = 0.89)

Small letters and capital letters are treated as the same. Do not perform any stemming, lemmatization or stop word removal.

a) Explain the idea behind the cosine similarity in vector space information retrieval (2.5%).
b) Explain the motivation behind the tf-idf formula and each of its components (2.5%).

c) Calculate the length of document #2. Show the calculation (5%).

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Doc 1</th>
<th>Doc 2</th>
<th>Doc 3</th>
<th>Doc 4</th>
<th>n</th>
<th>log(N/n)</th>
<th>di=tfidf</th>
<th>di^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td>0,30103</td>
<td>0,30103</td>
<td>0,090619</td>
</tr>
<tr>
<td>city</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td>0,30103</td>
<td>0,30103</td>
<td>0,090619</td>
</tr>
<tr>
<td>is</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Term</td>
<td>Frequency</td>
<td>Term ID</td>
<td>tfidf value</td>
<td>Maxfreq</td>
<td>Square sum:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
<td>-------------</td>
<td>---------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>many</td>
<td>1</td>
<td>1</td>
<td>0,60206</td>
<td>0,60206</td>
<td>0,362476</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>students</td>
<td>1</td>
<td>1</td>
<td>0,60206</td>
<td>0,60206</td>
<td>0,362476</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trondheim</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0,124939</td>
<td>0,124939</td>
<td>0,01561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>university</td>
<td>1</td>
<td>1</td>
<td>0,60206</td>
<td>0,60206</td>
<td>0,362476</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with</td>
<td>1</td>
<td>1</td>
<td>0,60206</td>
<td>0,60206</td>
<td>0,362476</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\sum (d_i^2) = 1.646753 \\
\sqrt{\sum (d_i^2)} \approx 1.28
\]

Length of document #2 is 1.28.

d) Calculate the tfidf score for the terms “trondheim” and “nidaros” for each of the documents (5%).

**Nidaros:**
Idf(nidaros) = \(\log(4/2)\)

D1, D2 = 0 (does not contain nidaros)

D3:
tfidf(nidaros) = \(\frac{1}{1}\log(4/2) = 0.301\) (maxFreq = 1)

D4:
tfidf(nidaros) = \(\frac{1}{2}\log(4/2) = 0.151\) (maxFreq = 2)

**Trondheim**
Idf(Trondheim) = \(\log(4/3)\)

D1:
tfidf = \(\frac{1}{2}\log(4/3) = 0.062\) (maxFreq = 2)
D2:
Tfidf(trondheim) = 1/1*\log(4/3) = 0.125 \text{ (maxFreq = 1)}

D3:
Tfidf(trondheim) = 0 \text{ (does not contain trondheim)}

D4:
Tfidf(Trondheim) = 2/2*\log(4/3) = 0.125

e) Calculate the cosine similarity for each of the documents with respect to the query (10%).

Query: trondheim nidaros

Show the list of ranked documents.

Query:
\[ W_{i,q} = (0.5 + 0.5 \times \frac{tf_{i,q}}{maxf_{i,q}}) \times \log(\frac{N}{n}) \]
\[ W_{nidaros,q} = (0.5 + 0.5 \times 1/1) \times \log(\frac{4}{2}) = 0.301 \]
\[ W_{trondheim,q} = (0.5 + 0.5 \times 1/1) \times \log(\frac{4}{3}) = 0.125 \]

Length og query:
\[ L_q = \sqrt{(0.301^2 + 0.125^2)} = 0.326 \]

D1:
Tfidf_nidaros = 0
Tfidf_trondheim = 0.062
Length = 0.89

D2:
Tfidf_nidaros = 0
Tfidf\textsubscript{trondheim} = 0.125
Length = 1.28

D3:
Tfidf\textsubscript{nidaros} = 0.301
Tfidf\textsubscript{trondheim} = 0
Length = 1.13

D4:
Tfidf\textsubscript{nidaros} = 0.151
Tfidf\textsubscript{trondheim} = 0.125
Length = 0.89

\[ \text{Sim}(d1,q) = \frac{0 \times 0.301 + 0.062 \times 0.125}{(0.89 \times 0.326)} = 0.0267 \]
\[ \text{Sim}(d2,q) = \frac{0 \times 0.301 + 0.125 \times 0.125}{(1.28 \times 0.326)} = 0.0374 \]
\[ \text{Sim}(d3,q) = \frac{0.301 \times 0.301 + 0 \times 0.125}{(1.13 \times 0.326)} = 0.2459 \quad \text{Correct length is 1.28} \]
\[ \text{Correct sim: } 0.217 \]
\[ \text{Sim}(d4,q) = \frac{0.151 \times 0.301 + 0.125 \times 0.125}{(0.89 \times 0.326)} = 0.2105 \]

Ranking: D3 > D4 > D2 > D1

---

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

a) Given a set of relevant documents $R_q$ for a given query $q$, and the returned (ranked) result from an experimental information retrieval engine for query $q$, $A_q$, calculate the interpolated precision at 11 standard recall levels. You are not required to graph the result. (5%)

$R_q = \{D4, D12, D23, D26, D30, D33, D60, D73, D83, D99\}$

<table>
<thead>
<tr>
<th>recall</th>
<th>precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>100 %</td>
</tr>
<tr>
<td>10 %</td>
<td>100 %</td>
</tr>
<tr>
<td>20 %</td>
<td>100 %</td>
</tr>
<tr>
<td>30%</td>
<td>75%</td>
</tr>
<tr>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>40%</td>
<td>67%</td>
</tr>
<tr>
<td>50%</td>
<td>56%</td>
</tr>
<tr>
<td>60%</td>
<td>46%</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) What is one of the biggest problems of precision/recall in web information retrieval, and which measure out of the two is more important to a user in web search? (2.5%)

We do not have detailed knowledge of the document collection searched; and hence no exact definition of the set of related documents for a given query. This leads to only estimation of the correct set of documents.

Precision will most likely be the most important measure, since the user is interested in getting valuable hits high in the ranking.

c) Calculate the R-precision for the query and the ranked list given in a). (2.5%)

Number of relevant documents = 10.
In the first 10 documents retrieved, there are 5 relevant documents;
R-precision is 5/10 = 0.5 (50%).

Question 3. Text Categorization (20%)

a) What is the basic hypothesis in using the vector space model for classification and what is the content of the hypothesis? (5%)

The basic hypothesis in using the vector space model for classification is the contiguity hypothesis.

Contiguity hypothesis: Documents in the same class form a contiguous region and regions of
different classes do not overlap.

b) Explain briefly the rationale of k-Nearest Neighbor (kNN) text classification. (5%)

The rationale of kNN classification is that, based on the contiguity hypothesis, we expect a test document d to have the same label as the training documents located in the local region surrounding d.

c) The table below shows a test set consisting 10 documents, where each document is annotated with its corresponding class (A, B, C). We wish to use this test set to find which class document X belong to. The distance between document X and the documents in the test set is given in the table. Use majority voting scheme kNN with k=5 to calculate which class document X belongs to. (5%)

<table>
<thead>
<tr>
<th>Document</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>C</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Distance from X</td>
<td>5.31</td>
<td>2.89</td>
<td>0.63</td>
<td>1.12</td>
<td>3.52</td>
<td>4.31</td>
<td>2.33</td>
<td>1.35</td>
<td>0.89</td>
<td>2.15</td>
</tr>
</tbody>
</table>

The five closest documents are: 3(B), 4(A), 8(C), 9(B), 10(B). B has three out of 5 closest documents, while A and C only have 1 out of the 5 closest documents. Document X is annotated with class B.

d) With the table given in question c), use weighted sum voting scheme kNN with k=5 to calculate which class document X belongs to. (5%)

The five closest documents are: 3(B), 4(A), 8(C), 9(B), 10(B). The maximum distance 2.15 is used for normalization and calculation of weights.

A = 1-1.12/2.15 = 0.479
B = (1-0.63/2.15) + (1-0.89/2.15) + (1-2.15/2.15) = 1.293
C = 1.35/2.15 = 0.37
Question 4. Clustering (15%)

a) Explain the main steps of the k-means clustering algorithm (5%)

See chapter 4 in Chakrabarti: Mining the web – discovering knowledge from hypertext data.

b) Assume the four documents below. Use the suffix tree clustering (STC) method explained in “Contextualized Clustering in Exploratory Web Search” and build a suffix tree for the documents (5%).

Doc 1: {gustave eiffel}
Doc 2: {eiffel, paris sightseeing}
Doc 3: {eiffel tower, paris attractions}
Doc 4: {alexandre gustave eiffel, paris}
c) Identify the three base clusters and calculate their scores (5%).

We use formula \( s(B) = |B| \times f(|P|) \)

<table>
<thead>
<tr>
<th>Base cluster</th>
<th>Documents</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>eiffel</td>
<td>1,2,4</td>
<td>3</td>
</tr>
<tr>
<td>paris</td>
<td>2,3,4</td>
<td>3</td>
</tr>
<tr>
<td>gustave eiffel</td>
<td>1,4</td>
<td>4</td>
</tr>
</tbody>
</table>

**Question 5. Latent Semantic Analysis (5%)**

Explain briefly the purpose of Singular Value Decomposition in latent semantic analysis (5%)

See paper on latent semantic analysis

**Question 6. Semantic Web (15%)**

a) What role does XML play in the Semantic Web? (3%)

- [ ] Definition of OWL
- [ ] Serialization of OWL ontology models = TRUE
- [ ] Unified format for ontology mapping
- [ ] Representation of ontology layout
- [ ] Semantic markup language
b) Which is the most expressive language that retains computational completeness and decidability? (3%)

- XML
- RDF
- RDFS
- OWL DL = TRUE
- OWL Full


c) Below is an OWL definition described using the abstract syntax from Antoniou et al.

\[
\text{SubClassOf} (\text{student complete intersectionOf} (\text{human smart})) \\
\text{ObjectProperty} (\text{hasStudNo InverseFunctional}) \\
\text{Class} (\text{student partial}) \\
\quad \text{restriction} (\text{read someValueFrom} (\text{book})) \\
\quad \text{restriction} (\text{read allValuesFrom} (\text{scientificBook})) \\
\text{Individual} (\text{john type(student)}) \\
\text{Individual} (\text{john value(hasStudNo 12345)}) \\
\text{Individual} (\text{Mary value(hasStudNo 12345)})
\]

Phrase these definitions in natural language. (3%)

- All students are smart humans.
- Things have unique student numbers.
- Students must read at least one book and they only read scientific books.
- John is a student. Both John and Mary have student number 12345.


d) What cannot be inferred from the statements in c)? (3%)

- Mary is a student
- John and Mary are the same individuals
- Mary reads only scientific books
- John is smart
- Each student has exactly one student number = TRUE

100%: Correctly answered otherwise 0%
e) Translate the following definition of professors into the OWL abstract syntax (3%)

*Scientific staff are either professors or researchers.*

*Scientific staff teach courses*

*Professors teach at the most one course each.*

```
SubClassOf(scientificStaff unionOf(professor researcher))
ObjectProperty(teach domain(scientificStaff) range(course))
Class(professor partial restriction(teach maxCardinality(1)))
```

**Question 7. Semantic Search (10%)**

a) Recall the article "Indexing a Web Site with a Terminology Oriented Ontology" by Demontils et al., they mention "cumulative similarity measure" by Demontils & Jacquin. Why did they stem instead of lemmatize the words before they were annotated with part of speech tags? (2.5%)

- Stemming is the process to find the common root form of words and hence an ideal basis for a part of speech tagger;
- Stemming is the process to find the canonical form of words and hence an ideal basis for a part of speech tagger;
- Lemmatization is the process to find the canonical form of words and hence not suitable as a good basis for a part of speech tagger;
- Lemmatization is the process to find the common root form of words and hence not suitable as a good basis for a part of speech tagger;
- They did not use stemming, but lemmatization. = TRUE

100%: Correctly answered otherwise 0%

b) Recall the article "Construction of Ontology based Semantic-Linguistic Feature Vectors for Searching: the Process and Effect" by Tomassen & Strasunskas. The development of the approach is inspired by a linguistic method for describing the meaning of objects, where a FV "connects" something. What does a feature vector "connect"? (2.5%)

- An ontology property with associated textual documents;
- An ontology property with associated domain terminology;
□ An ontology entity with associated textual documents;
□ An ontology entity with associated domain terminology; = TRUE
□ An ontology with associated textual documents;
□ An ontology with associated domain terminology.

100%: Correctly answered otherwise 0%

c) Recall the article "Measuring intrinsic quality of semantic search based on Feature Vectors" by Tomassen & Strasunskas. A set of evaluation measures were proposed. What is the purpose of the Average Fv Similarity measure? (2.5%)

□ Provides an indication of the uniqueness of the FVs; = TRUE
□ Provides an indication of the degree of semantic relatedness between neighbouring entities;
□ Provides an indication of the overall quality of the FVs;
□ Provides an indication of the FV quality relative to the ontology quality;
□ Provides an indication of the semantic distance between the FV terms.

100%: Correctly answered otherwise 0%

d) What are ontologies normally not used for in semantic search applications? (2.5%)

□ Disambiguation of queries
□ Query expansion with synonyms
□ Semantic indexing with ontology concepts
□ Latent semantic indexing = TRUE
□ Semantic annotation of documents

100%: Correctly answered otherwise 0%

Appendix A. Formulas

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

\[ w_{kq} = (0.5 + \frac{0.5 \cdot \text{freq}_{kq}}{\max_i \text{freq}_{iq}}) \cdot \log \left( \frac{N}{n_k} \right) \]

\[ w_{ij} = \frac{f_{ij}}{\max_i (f_{ij})} \cdot \log \left( \frac{N}{n_j} \right) \]