Allowed means of assistance: D No printed or handwritten materials 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 (18%)**

a) **Describe the idea behind the tf.idf score. (4%)**

   2.5.3 i Baeza-Yates & Ribeiro-Neto: Modern Information retrieval

   The score contains two elements:
   1) *tf*: The score increases with high frequency of the term relative to the frequency of the most frequent term in the document.
   2) *idf*: The score decreases if the term is found in many of the other documents in the collection. The log function is used to dampen the score

   *tfidf* score of a term that does not appear in the document collection is zero.
   *tfidf* score of a term that appears in all documents is zero.

b) **Given the following documents:**

   **Document #1:** Salmon is the common name for several species of Fish of the family Salmonidae.

   **Document #2:** Several other fish in the family are called trout.
Document #3: Salmon live in both the Atlantic Ocean and the Pacific Ocean.

Document #4: Typically, salmon are anadromous; they are born in fresh water, migrate to the ocean, then return to fresh water to reproduce.

Assume the query “fish” (not a phrase). Calculate the cosine similarity between Document #1 and the query. Show the calculations. Do not consider stop word removal, stemming and other preprocessing techniques. (10%)

<table>
<thead>
<tr>
<th>Term:</th>
<th>Doc1</th>
<th>doc2</th>
<th>doc3</th>
<th>doc4</th>
<th>n</th>
<th>log(N/n)</th>
<th>d_i = tf.idfi</th>
<th>d_i^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>salmon</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0.125</td>
<td>0.06247</td>
<td>0.003902</td>
</tr>
<tr>
<td>is</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
<tr>
<td>the</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>common</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
<tr>
<td>name</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
<tr>
<td>for</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
<tr>
<td>several</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.301</td>
<td>0.15051</td>
<td>0.022655</td>
</tr>
<tr>
<td>species</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
<tr>
<td>of</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.60206</td>
<td>0.362476</td>
</tr>
<tr>
<td>fish</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.301</td>
<td>0.15051</td>
<td>0.022655</td>
</tr>
<tr>
<td>family</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.301</td>
<td>0.15051</td>
<td>0.022655</td>
</tr>
<tr>
<td>salmonidae</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.602</td>
<td>0.30103</td>
<td>0.090619</td>
</tr>
</tbody>
</table>

\( \sum d_i^2 = 0.978 \)

Length = \( \sqrt{\sum d_i^2} = 0.989 \)

\( w_{fish,1} = 0.1505 \)

\( W_{fish,q} = (0.5 + 0.5*1/1)*\log(4/2) = 0.301 \)

Length of query: \( \sqrt{0.301^2} = 0.301 \)

Cosine similarity: \( (0.301*0.1505)/(0.301*0.989) = 0.15 \)

c) What is polysemy and what is its main problem in IR? (4%)

Polysemy: A word may have several interpretations depending on context.

Problem: Precision is affected negatively by non-relevant documents that are returned to the user.

Question 2 : The Index (22 %)
The *Index* is a central component in most retrieval systems. One of the main indexing techniques is the so-called inverted files. An inverted file has two main elements, the *Vocabulary* and the list of *Occurrences*.

**a) Define briefly the two main elements Vocabulary and Occurrences. (5%)**

- *Vocabulary* is simply the set of words in the text.
- *The set of occurrences* is the list of all the occurrences of each word in the vocabulary.

**b) What are the required operations on an index? (4%)**

- *Look-up word*
- *Insert word*
- *Retrieve occurrences*
- *Manipulate occurrences*

In the curriculum\(^1\) several data-structures are discussed for the representation of indices, such as: Tries / Prefix-Trees, Suffix-Trees, Suffix Arrays and Supra Indices.

**c) Create a Prefix tree (trie) for the words beginning with the letters ‘a’, ‘i’, ‘s’ and ‘t’ for the documents given in question 1. Discuss briefly the performance issues of this data structure, for example with regards to the required operations you have listed in b) and with respect to the other possible data structures listed above. (8%)**

---

\(^1\) Ricardo Baeza-Yates and Berthier Riberio-Neto, “Modern Information Retrieval”, Chapter 8.2 and 8.3
During the indexing process a variety of Natural Language Processing (NLP) techniques may be applied to the document text during a phase we may call “index term selection”.
d) Describe how the two techniques Stemming and Stop-word-removal would alter the index you made in c). (You do not have to reconstruct the index.) (5%)

- Stopword removal would remove a substantial part of the index (the, they, then, is, in, are, the, …)
- Stemming would not have the desired effect in this corpus, Salmon -> Salmonidae. However, proper lemmatisation would detect ‘is’ and ‘are’ as inflections of ‘be’.

**Question 3. Retrieval Evaluation (16%)**

a) Given the set of relevant documents for a query q,

Rq = {d1, d4, d9, d17, d22, d39, d41, d48, d63, d70}

Show the calculation of the interpolated precision at 11 standard levels for the following ranked results for the query q: (8%)

1) d1, 2) d74, 3) d41, 4) d13, 5) d50, 6) d9, 7) d7, 8) d61, 9) d22, 10) d44, 11) d19, 12) d28, 13) d17, 14) d51, 15) d2

<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>67 %</td>
</tr>
<tr>
<td>30 %</td>
<td>50 %</td>
</tr>
<tr>
<td>40 %</td>
<td>44 %</td>
</tr>
<tr>
<td>50 %</td>
<td>38 %</td>
</tr>
<tr>
<td>60 %</td>
<td>0 %</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) For a given query there are 16 relevant documents in the collection. The precision for the query is 0.40, and the recall for the query is 0.25. How many documents are in the result set? (4%)

\[
|A| = \text{unknown} \\
|Ra| = r * |R| \text{ (from the definition of recall)} \\
|A| = |Ra| / p \text{ (from definition of precision)} \\
|A| = (r * |R|)/p = (0.25 * 16)/0.4 = 10 \\
10 documents in the result set.
\]

c) What is a precision histogram, and what is it used for, how is it calculated (and what does negative, positive, and 0 values mean)? (4%)

A precision histogram is a graphical representation (histogram) that is used to compare the performance of two retrieval algorithms.

The histogram values are calculated based on R-precision values for each query and algorithm, and are based on \( \text{RP}_{A,i} = \text{RP}_{A,i} - \text{RP}_{B,i} \), where \( \text{RP}_{A,i} \) and \( \text{RP}_{B,i} \) are the R precision measures for algorithm A and B respectively for query i. A negative value indicates better performance for B, while a positive value indicates better performance for A, while 0 indicates equal performance.
Question 4. Clustering (8%)

How many iterations does the k-means algorithm need to identify the four clusters in Figure 1? The crosses show the starting points for the clusters. (8%)  

**Fasit:** 4 iterations. See solution below. An iteration consists of 1) drawing new borders and 2) moving the cluster centroid.

Iteration 1:

Iteration 2:
Iteration 3:

Iteration 4:
### Question 5. Classification (10%)  

<table>
<thead>
<tr>
<th>Dokument</th>
<th>Klasse</th>
<th>Avstand fra dokument X</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>5,21</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>2,19</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>5,10</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>1,20</td>
</tr>
<tr>
<td>5</td>
<td>B</td>
<td>0,17</td>
</tr>
<tr>
<td>6</td>
<td>C</td>
<td>3,45</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>1,23</td>
</tr>
<tr>
<td>8</td>
<td>B</td>
<td>6,54</td>
</tr>
<tr>
<td>9</td>
<td>C</td>
<td>4,31</td>
</tr>
<tr>
<td>10</td>
<td>B</td>
<td>3,12</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>4,32</td>
</tr>
<tr>
<td>12</td>
<td>C</td>
<td>5,66</td>
</tr>
<tr>
<td>13</td>
<td>C</td>
<td>6,74</td>
</tr>
</tbody>
</table>

The table shows a test set consisting of 13 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 belongs to. The distance between document X and the documents in the test set is given in the rightmost column. Use kNN (k Nearest Neighbor) with $k = 7$ to calculate which class document X belongs to using:

a) Weighted sum voting scheme (5%)

b) Majority voting scheme (5%)

Remember to give the arguments of your answer!

The seven closest documents are: 2 (A), 4 (A), 5 (B), 6 (C), 7 (A), 9 (C), 10 (B). The maximum distance 4.31 is used for normalization and calculation of weights.

**a)** \[ A = \frac{1-2,19}{4,31} + \frac{1-1,20}{4,31} + \frac{1-1,23}{4,31} = 3 - \frac{2,19 + 1,20 + 1,23}{4,31} = 1,92 \]
\[ B = 2 - \frac{3,12 + 0,17}{4,31} = 1,23 \]
\[ C = 2 - \frac{3,45 + 4,31}{4,31} = 0,19 \]

Document X is annotated with class A.

**b)** A has three out of 7 docs, while B and C only have 2 out of seven each. Document X is annotated with class A.

### Question 6. Ontology Web Language (12%)
a) Recall the article "A Framework for Understanding and Classifying Ontology Applications" by Jasper et al. Briefly describe the difference between two ontology usage scenarios, namely, Authoring Operational Data and Authoring Ontology. (2%)

In Authoring Operational Data, an ontology is used to author operational data and operational data are translated to be used by an application.

In Authoring Ontology, an ontology author creates an ontology, which they convert into an operational target system (e.g., a knowledge base). Application users then interact with an operational system to perform their desired tasks.

b) OWL is layered into three different layers called; OWL Lite, OWL DL, and OWL Full. Describes shortly the main differences between these layers? (2%)

- **OWL Lite** adds the possibility to express definitions and axioms, together with a limited use of properties to define classes.
- **OWL DL** supports those users who want the maximum expressiveness while retaining good computational properties.
- **OWL Full** is meant for users who want maximum expressiveness with no computational guarantees.

c) Select correct statement/statements about importance of reasoning support in OWL: (2%)

- [ ] Check for unintended relationships between classes; **= TRUE**

- [ ] Automatically instantiate classes;

- [ ] Derive explicitly all the statements that are false in the ontology, to better understand its properties;
- Reduce the redundancy of an ontology; = TRUE
- Discover unequal descriptions, reuse concept descriptions, and refine the definitions;

**d) Select correct statement/statements about OWL DL class construct: (2%)**

- owl:Class is equivalent to rdfs:Class;
- rdfs:Class is a proper superclass of owl:Class; = TRUE
- owl:Class is a proper subclass of rdfs:Class; = TRUE
- classes can be treated as individuals;
- rdfs:Class is a proper subclass of owl:Class.

**e) Interpret the following statement in abstract OWL syntax and write a corresponding statement in a natural language sentence. (2%)**

Class(author super(person) super(restriction(written (someValuesFrom(book) or someValuesFrom(article)))))
An author is a subclass of person that has written at least one book or article.

f) What is the essential difference between open world and closed world assumption? (2%)  

- **Open world**: the system's knowledge is incomplete and hence cannot conclude that a statement is false if not inferred. E.g. OWL  
- **Closed world**: The system's knowledge is complete and hence always possible to draw the conclusion that a statement is either true or false. Everything that is not inferred to be true is false. E.g. SQL

**Question 7. Ontology Engineering (5%)**

a) In the article "Ontology Learning and Its Application to Automated Terminology Translation" by Navigli et al. they describe three levels of generality for domain ontologies.

Describe shortly these three levels of generality. (5%)  

1. **Foundational ontology (top or upper ontology also accepted)**: Consist of a few basic principles that support a model’s generality to ensure reusability across different domains.  
2. **Core ontology**: Key domain conceptualizations (of Foundational ontology) and describe them according to the organizational structure established, usually application-domain concepts.  
3. **Specific domain ontology**: Highly specific concepts of the domain.
Question 8. Ontology Applications (9%)

a) In the article "KnOWLer - Ontological Support for Information Retrieval Systems" by Ciorascu et al. they present knOWLer, which is an ontology-based information management system. In this paper they use an extended version of OWL Lite by W3C called Extended OWL Lite. What is the difference between OWL Lite and Extended OWL Lite? (3%)

They have extended OWL Lite to allow the usage of RDF Statements

b) What is the special characteristic of the knOWLer system, emphasized by the authors, compared to other similar systems of the time the paper was published? (3%)

Scalability

c) In the article “Indexing a Web Site with a Terminology Oriented Ontology” by Demontils et al., they mention "cumulative similarity measure". Describe shortly the purpose of the "cumulative similarity measure". (3%)
For evaluating the relative importance of a concept in a page. It is the sum of all the similarity measures calculated between this concept and all the other concepts included in the studied page.

Appendix A. Formulas

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

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

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

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

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

\[
\text{sim}(d, q) \approx \sum_{k=1}^{n} w_{k,q} * w_{k,j} * \left( \log \frac{P(k_i | R)}{1 - P(k_i | R)} \right) + \log \frac{1 - P(k_j | R)}{P(k_j | R)}
\]

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

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

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

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

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