Question 1. Vector Space Information Retrieval Model (27%)

Assume a document collection that consists of the following 4 documents:

Document 1 (document length: 0.51):
The Eiffel Tower was built for the International Exhibition of Paris of 1889 commemorating the centenary of the French Revolution. Of the 700 proposals submitted in a design competition, Gustave Eiffel’s was unanimously chosen.

Document 2 (document length: 0.87):
A signature of the Las Vegas skyline, the replica Eiffel Tower at Paris Las Vegas is an exact reproduction of one of Europe’s highest towers.

Document 3 (document length: 0.58):
Alexandre Gustave Eiffel was born in Dijon, France, in 1832

Document 4 (document length: 0.70):
Another undisputed great engineering feat is the Eiffel Tower, a huge wrought-iron skeleton in Paris. Alexandre Gustave Eiffel designed the 984-foot tower for the exposition of 1889. The tower contains about 7,000 short tons of iron and steel.

Assume that hyphens and apostrophes are parts of words and not word boundaries (e.g. Eiffel’s and 984-foot are each one word). Periods and commas are ignored, and small letters and capital letters are treated as the same. Also assume the vector space model using the formulas in Appendix A.
a) Explain the idea behind the cosine similarity measure in vector space information retrieval! (8%)

See Chapter 2.5.3 in Baeza-Yates & Ribeiro-Neto: Modern Information Retrieval

b) What is the tf.idf score of the term “tower” for each of the four documents? (6%)

Show the calculations in the box below. Use the log table below if needed.

\[ idf_{tower} = \log(4/3) = 0.125 \]

Doc 1: \( tf = 1/5 = 0.20 \), (“the” has frequency 5)
\[ tf.idf = 0.20 \times 0.125 = 0.025 \]

Doc 2: \( tf = 1/3 = 0.333 \), (“of” has frequency 3)
\[ tf.idf = 0.333 \times 0.125 = 0.0417 \]

Doc 3: \( tf.idf = 0 \) since \( tf = 0 \)

Doc 4: \( tf = \frac{1}{4} = 0.75 \), (“the” has frequency 4)
\[ tf.idf = 0.75 \times 0.125 = 0.09375 \]

<table>
<thead>
<tr>
<th>( x )</th>
<th>1.00</th>
<th>1.33</th>
<th>1.67</th>
<th>2.00</th>
<th>2.33</th>
<th>2.66</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(x) )</td>
<td>0</td>
<td>0.125</td>
<td>0.222</td>
<td>0.301</td>
<td>0.368</td>
<td>0.426</td>
<td>0.477</td>
<td>0.602</td>
</tr>
</tbody>
</table>
c) Calculate the cosine similarity scores for each document with respect to the query below. (10%)

Query $q = \text{highest tower}$

$$w_{\text{tower},q} = (0.5 + 0.5) \times \log(4/3) = 0.125$$
$$w_{\text{highest},q} = (0.5 + 0.5) \times \log(4/1) = 0.602$$

Length of query: $\sqrt{0.125 \times 0.125 + 0.602 \times 0.602} = 0.615$

$$\text{sim}(d_1, q) = \frac{(0.025 \times 0.125 + 0 \times 0.602)}{(0.51 \times 0.615)} = 0.0099$$
$$\text{sim}(d_2, q) = \frac{(0.0417 \times 0.125 + 0.0417 \times 0.602)}{(0.87 \times 0.615)} = 0.057 \quad \{w_{\text{highest},d_2} = 0.0417\}$$
$$\text{sim}(d_3, q) = \frac{(0 \times 0.125 + 0 \times 0.602)}{(0.58 \times 0.615)} = 0$$
$$\text{sim}(d_4, q) = \frac{(0.09375 \times 0.125 + 0 \times 0.602)}{(0.70 \times 0.615)} = 0.027$$

Ranking: $\text{Doc2} > \text{Doc4} > \text{Doc1} > \text{Doc3}$

\[\]

d) What role does the log function serve in tf.idf scores? (3%)

\[Dampening the scores of the idf part.\]
**Question 2. Boolean Retrieval Model (5%)**

Assume a boolean retrieval model and the same document collection as in question 1.

What is the set of relevant documents for the query $q = \text{NOT Alexandre AND (Gustave OR Eiffel)}$?

$\{\text{document1, document2}\}$

**Question 3. Linguistic Techniques (20%)**

Make the same assumptions as for question 1 above.

Document 5:

The Eiffel Tower was opened in Paris in 1889 and was the world’s tallest building. Opening hours today are restricted

a) Create an inverted index for document 5 above (with positions). (5%)
b) Lemmatize the text in Document 5 and create a new inverted index for the lemmatized text. (5%)

1889: 40
be: 18, 49, 103
and: 45
building: 73
eiffel: 5
hour: 91
in: 28, 37
open: 22, 83
paris: 31
restrict: 107
tall: 65
today: 97
the: 1, 53
tower: 12
world: 57

c) What is the difference between stemming and lemmatization? List advantages and disadvantages for these two methods! (7%)

Stemming is an algorithmic approach, where suffixes are analyzed and removed from the word. With lemmatization we look up the word in a dictionary. Whereas the result in stemming is a stem, the result in lemmatization is the base form (aka canonical form, lemma) of the word.

Stemming:
+ no need for lexicon
+ can handle unknown words
- more error-prone
- stem is not a real word

Lemmatization
+ higher accuracy
+ faster (questionable)
- memory-demanding
- acquisition and maintenance of lexicon
d) What is the typical effect of introducing lemmatization in search systems? (3 %)

Exactly one of the options below should be marked with a cross.

- [ ] Improved precision
- [ ] Improved recall = TRUE
- [ ] Reduced index size
- [ ] Language independent index
- [ ] Improved speed of retrieval

**Question 4. Collocations (12%)**

Assume a document collections of 893,484 tokens. There are 7 occurrences of Gustave Eiffel in the collection and a total of 642 occurrences of Eiffel and 1287 occurrences of Gustave.

a) Calculate the chi-square (X²) value for Gustave Eiffel. Show the calculation below. (10%)

\[
N=893484
\]
\[
\begin{array}{c|c|c}
& = \text{gustave} & \text{NOT gustave} \\
= \text{eiffel} & 7 & 635 \\
\text{NOT eiffel} & 1280 & 891562 \\
\end{array}
\]

\[X^2 = 40.0\]

(There is a mistake in the book/presentation, where N does not include the number in the upper left corner. The answer should be about the same also with that mistake)

b) Is Gustave Eiffel a good candidate for a collocation? (2%)

Use the X² distribution table below. Justify your answer.

\[
\begin{array}{cccccccc}
p & 0.99 & 0.95 & 0.10 & 0.05 & 0.01 & 0.005 & 0.001 \\
d.f. & & & & & & & \\
1 & 0.00016 & 0.0039 & 2.71 & 3.84 & 6.63 & 7.88 & 10.83 \\
2 & 0.020 & 0.10 & 4.60 & 5.99 & 9.21 & 10.60 & 13.82 \\
3 & 0.115 & 0.35 & 6.25 & 7.81 & 11.34 & 12.84 & 16.27 \\
4 & 0.297 & 0.71 & 7.78 & 9.49 & 13.28 & 14.86 & 18.47 \\
100 & 70.06 & 77.93 & 118.5 & 124.3 & 135.8 & 140.2 & 149.4 \\
\end{array}
\]

For a 2x2 table we have one degree of freedom. A probability level of 0.05 is appropriate for this analysis.

Since 40.0 > 3.84, gustave eiffel is a good candidate for a collocation
**Question 5. Text Categorization (10%)**

a) Classifiers can be divided into two groups. Answer shortly: what are these? Explain the difference and give at least one example for each of them. (4%)

*Manually and automatically*

b) Calculate probability below. (2%)

\[
P(A) = 0.10 \\
P(B) = 0.30 \\
P(A \lor B) = 0.25
\]

A and B are not independent.

Which is the correct answer of \( P(A \mid B) \)?

- 0.25
- 0.50 = TRUE
- 0.80
- 0.15

c) Answer shortly: what is the difference of Bayesian and Naïve Bayesian categorization? (4%)

*Bayesian assumes dependency while Naïve Bayesian assumes independency.*
**Question 6. Ontology Learning (5%)**

Which three properties are considered important for a good domain ontology in "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites" by R. Navaglì and P. Velardi?

- **Convergence**
- **Consensus** = TRUE
- **Accessibility** = TRUE
- **Broadness**
- **Number of relationship types**
- **Coverage** = TRUE
- **Number of community members**
- **Number of service levels**
- **Number of concepts**

**Question 7. Ontology Applications (10%)**

a) In the article “Indexing a Web Site with a Terminology Oriented Ontology” by Demontils and Jacquin, four coefficients are calculated and presented:

- Direct Indexing Degree (DID)
- Indirect Indexing Degree (IID)
- Ontology Cover Degree (OCD)
- Mean of the Representativeness of the candidate Concepts (MRC)

Describe shortly each of them. (6%)

*(DID) the rate of concepts directly involved in HTML pages*  
*(IID) the rate of concepts indirectly involved in HTML pages*  
*(OCD) the rate of pages concerned with the ontology concepts*  
*(MRC) the mean of the representativeness of the candidate concepts*
b) The query language RDQL is discussed in the paper “KnOWLer - Ontological Support for Information Retrieval Systems”. What is the natural language version of following RDQL query? (4%)

\[
\forall (D,W,A) \; ; \\
\quad \text{wordnet}\#\text{antonymOf}(A,W) \land \\
\quad \exists SW : \text{stem}\#\text{stemFromWord}(SW,W) \land \\
\quad \exists SW.DE : \text{stem}\#\text{stemToDoc}(SW, SW.DE) \land \\
\quad \text{corpus}\#\text{doclink}(SW.DE, D) \land \\
\quad \exists SA : \text{stem}\#\text{stemFromWord}(SA,A) \land \\
\quad \exists SA.DE : \text{stem}\#\text{stemToDoc}(SA, SA.DE) \land \\
\quad \text{corpus}\#\text{doclink}(SW.DE, D). \\
\]

“find all documents that contain a word as well as at least one of its antonyms”

**Question 8. Ontology Web Language (11%)**

a) There exists a number of languages with various ability to represent semantics. Rank (strongest semantics to weakest semantics) the following lists of languages based on their ability to represent semantics: UML; OWL, RDF(S); ER; Description Logic. (2%)

3. UML; 2. OWL, 4. RDF(S); 5. ER; 1. Description Logic.

b) Recall Guarino’s paper on “Formal Ontology and Information Systems”. Briefly describe the difference between domain ontology and application ontology. (2%)

*Domain ontology describes the vocabulary related to a generic domain (like medicine, automobiles, etc.) and specializes the terms introduced in the top-level ontology.*

*Application ontology describe concepts depending both on a particular domain and task, which are often specializations of both task and domain ontology.*
c) Select the language which provides maximum expressiveness: (1%)  
- OWL Lite
- OWL Full = TRUE
- RDF(S)
- OWL DL

d) Interpret OWL statement SubClassOf(c_i unionOf(c_j c_k)) and choose the most correct interpretation: (1%)  
- c_i is superclass for both c_j and c_k;
- c_i is subsumed by union of c_j and c_k; = TRUE
- c_i is superclass of union of c_j and c_k;
- c_i is subsumed by either c_j or c_k

e) Recall the main characteristics of OWL Full. Which of the following statement(s) are/is not true: (1%)  
- unrestricted expressiveness;
- flexibility of RDF;
- full reasoning; = TRUE
- class can be treated as individual.

f) Given the following statements:  
hasChild(JOCASTA, OEDIPUS), i.e. Jocasta has a child, called Oedipus  
hasChild(JOCASTA, POLYNEIKES), i.e. Jocasta has a child, called Polyneikes  
hasChild(OEDIPUS, POLYNEIKES), i.e. Oedipus has a child, Polyneikes  
hasChild(POLYNEIKES, THERSANDROS), i.e. Polyneikes has a child, called Thersandros  
Patricide(OEDIPUS), i.e. Oedipus is a patricide  
¬Patricide(THESANDROS), i.e. Thersandros is not a patricide

And query:  
∃hasChild.(Patricide ∧ ¬hasChild.¬Patricide))(JOCASTA), i.e. whether Jocasta has a child that is a patricide and that itself has a child that is not a patricide?

What will be the result of the query (one or more answers)? (2%)  
- no answer (incomplete information);
- Oedipus and his child Polyneikes = TRUE
- Polyneikes and his child Thersandros = TRUE
- Oedipus and Thersandros
g) Select all wrong statements. (2%)

OWL supports hierarchical decomposition of:

- classes; [ ] TRUE
- individuals; [ ] TRUE
- properties; [ ] TRUE
- cardinality. [ ] TRUE

Appendix A. Formulas

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

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

\[
w_{i,j} = \frac{f_{i,j}}{\text{maxl} \ast \text{freq}_{i,q}} \ast \log \left(\frac{N}{n_i}\right)
\]

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

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

\[
sim(d_j,q) = \sum_{i=1}^{m} w_{i,q} \ast w_{i,j} \ast \left(\log \left(\frac{P(k_i \mid R)}{1 - P(k_i \mid R)}\right) + \log \left(1 - P(k_i \mid R)\right)\right)
\]

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

\[
t = \frac{\bar{x} - \mu}{s / \sqrt{n}}
\]

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

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

\[
P(E \mid c_j) = P(e_1 \land e_2 \land \cdots \land e_m \mid c_j) = \prod_{j=1}^{m} P(e_j \mid c_j)
\]