TDT4117 Information Retrieval - Autumn 2021
Assignment 2
Deadline for delivery is 06.10.2021

September 22, 2021

Important notes

Please carefully read the following notes and consider them for the assignment delivery. Submissions that do not fulfill these requirements will not be assessed and should be submitted again.

1. The assignment must be delivered in **pdf format**. Other formats such as .docx and .txt are not allowed.
2. The assignment must be **typed**. Handwritten assignments are not accepted.
3. Final scores are **required**, but not sufficient. You need to explicitly write the details of your computations (with no redundancy).
4. You may work in groups of maximum 2 students.

Task 1 - Relevance Feedback

1. Explain the difference between automatic local analysis and automatic global analysis.
2. What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

Task 2 - Language Model

1. Explain the language model, what are the weaknesses and strengths of this model?
2. Given the following documents and queries, build the language model according to the document collection.

```
d1 = An apple a day keeps the doctor away.
d2 = The best doctor is the one you run to and can't find.
d3 = One rotten apple spoils the whole barrel.

q1 = doctor
q2 = apple orange
q3 = doctor apple
```

Use MLE for estimating the unigram model and estimate the query generation probability using the Jelinek-Mercer smoothing

$$\hat{P}(t|M_d) = (1-\lambda)\hat{p}_{mle}(t|M_d) + \lambda\hat{p}_{mle}(t|_C), \lambda = 0.5. \quad (1)$$

For each query, rank the documents using the generated scores.

3. Explain what smoothing means and how it affects retrieval scores. Describe your answer using a query from the previous subtask.

Task 3 - Evaluation of IR Systems

1. Explain the terms Precision and Recall, including their formulas. Describe how **differently** these metrics can evaluate the retrieval quality of an IR system.

2. Explain the terms MAP and MRR ranking methods. List two pros and cons of each of methods in information retrieval querying.

3. Given the following set of relevant documents $rel = \{$*23, 10, 33, 500, 70, 59, 82, 47, 72, 9*$\}$, and the set of retrieved documents $ret = \{$*55, 500, 2, 23, 72, 79, 82, 215*$\}$, provide a table with the calculated precision and recall at each level.

Task 4 - Interpolated Precision

1. What is interpolated precision?

2. Given the example in Task 3.2, find the interpolated precision and make a graph.