TDT4117 Information Retrieval - Autumn 2022
Assignment 2
Deadline for delivery is 06.10.2022

September 22, 2022

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 - 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.

   d_1 = *failure is the opportunity to begin again more intelligently.*

   d_2 = *intelligence is the ability to adapt to change.*

   d_3 = *lack of will power leads to more failure than lack of intelligence or ability*

$q_1 = \textit{failure}$

$q_2 = \textit{intelligence opportunity}$

$q_3 = \textit{intelligence failure}$

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 2 - Evaluation of IR Systems

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

2. 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.

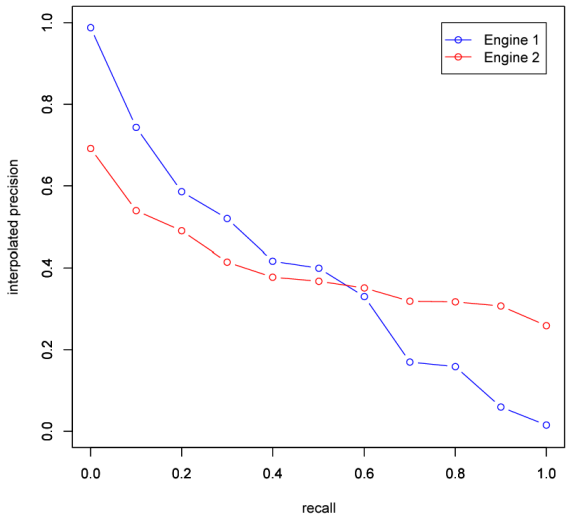
Task 3 - Interpolated Precision

1. What is interpolated precision?

2. Given the example in Task 2.2, find the interpolated precision and make a graph.

3. The figure below depicts interpolated precision-recall curves for two search engines that index research articles. There is no difference between the engines except in how they score documents. Imagine you're a scientist looking for all published work on some topic. You don't want to miss any citation. Which engine would you prefer and why?

Figure 1: interpolated precision-recall curves



Task 4 - True/False Questions

Please choose either True or False for each of the following statements. For the statement you believe it is False, please give your explanation of it (you do not need to explain when you believe it is True). **Note: the credit can only be granted if your explanation for the false statement is correct.**

1. A harmonic mean is used in the F1 score instead of arithmetic mean because of computational complexity.

2. If we could have a corpus with an infinite number of documents, smoothing is not needed when estimating those document language models.

3. The goal of retrieval models we have learnt is to improve some specific IR evaluation metrics, such as NDCG and MAP.

4. Given a very large IR evaluation collection, where System A achieves a MAP of 0.33 and System B achieves a MAP of 0.79, we can safely conclude that System B is significantly better than System A.

5. We usually make independence assumption for the purpose of reducing computational cost in retrieval models.