Assignment 2 - TDT4117

Hermann Owren Elton, Olaf Rosendahl

September 30, 2022

Task 1 : Language Model

1. Explain the language model, what are the weaknesses and strengths of this model?

   The language model is a model which is used to determine the likelihood of generating a specific sentence based on the text in a document. This is done with the use of various probabilistic and statistical techniques. The strengths of this is that it's mathematically precise and conceptually simple and intuitive. A weakness is that it's difficult to predict the user's needs. It can also be difficult to improve the relevance.

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 = *failure*
   q_2 = *intelligence opportunity*
   q_3 = *intelligence failure*

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

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

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

   $M_{d1} = 9$
   $M_{d2} = 8$
   $M_{d3} = 14$
   $C = 9 + 8 + 14 = 31$

   $\hat{P}(q_1|d_1) = 0.5 * (\frac{1}{9} + \frac{2}{31}) = 0.0878$
   $\hat{P}(q_1|d_2) = 0.5 * (\frac{0}{8} + \frac{2}{31}) = 0.0323$
   $\hat{P}(q_1|d_3) = 0.5 * (\frac{1}{14} + \frac{2}{31}) = 0.0680$

$q_1 : d_1 > d_3 > d_2$

$$\hat{P}(q_2|d_1) = 0.5 * (\tfrac{0}{9} + \tfrac{2}{31}) * 0.5 * (\tfrac{1}{9} + \tfrac{1}{31}) = 0.0023$$
$$\hat{P}(q_2|d_2) = 0.5 * (\tfrac{1}{8} + \tfrac{2}{31}) * 0.5 * (\tfrac{0}{8} + \tfrac{1}{31}) = 0.0015$$
$$\hat{P}(q_2|d_3) = 0.5 * (\tfrac{1}{14} + \tfrac{2}{31}) * 0.5 * (\tfrac{0}{14} + \tfrac{1}{31}) = 0.0011$$

$q_2 : d_1 > d_2 > d_3$

$$\hat{P}(q_3|d_1) = 0.5 * (\tfrac{0}{9} + \tfrac{2}{31}) * 0.5 * (\tfrac{1}{9} + \tfrac{2}{31}) = 0.0028$$
$$\hat{P}(q_3|d_2) = 0.5 * (\tfrac{1}{8} + \tfrac{2}{31}) * 0.5 * (\tfrac{0}{8} + \tfrac{2}{31}) = 0.0031$$
$$\hat{P}(q_3|d_3) = 0.5 * (\tfrac{1}{14} + \tfrac{2}{31}) * 0.5 * (\tfrac{1}{14} + \tfrac{2}{31}) = 0.0046$$

$q_3 : d_3 > d_2 > d_1$

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

   Smoothing is a technique which is used to avoid zeros when evaluating the occurrence of a specific query-word in a document. When you don't use smoothing, a document which contains all but one of the query word won't match because the missing word creates a zero in the search. This zero eliminates that document.

   Document *d2* in the previous subtask seems like a quite good result for the query *q2*. But since *d2* doesn't contain *opportunity* from *q2*, *d2* will be excluded from the results without smoothing.

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.

   - **MAP** (Mean Average Precision) is calculated by the mean average precision for each query.
     - Pros:
       - Calculates a single metric that represents the area below the precision-recall curve.
       - Weights errors which happens high in the list significantly more then errors deeper in the list.
     - Cons:
       - It's best at lists where each document can be seen as either relevant or not relevant. It's not fit where there's for example numerical ratings as an error measure can't be found in these cases.
       - In order to deal with fine-grained ratings such as numerical ratings, one have to threshold the ratings to transform them into binary relevance's which then can be evaluated.

- **MRR** (Mean Reciprocal Rank) attempts to find the position of the first relevant item. The MRR is computed with the average score between multiple scores.
  - Pros:
    - Easy to compute and interpret
    - Well suited for queries looking for "the best item" since it looks for the first relevant item
  - Cons:
    - Does only evaluate the first relevant item and not the rest of the list.
    - A list of only 1 relevant item is given as much weight as a list which contains many relevant items.

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.

Table 1: Precision/Recall

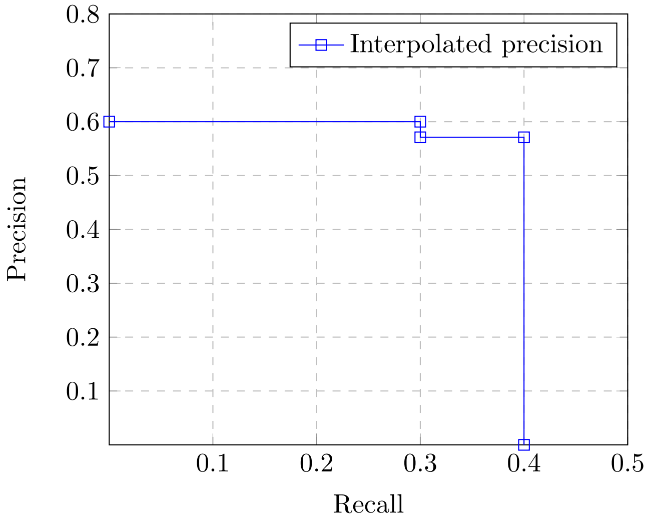
Documents retrieved	Doucment retrieved	Relevant	Recall	Precision
0	-	-	0	-
1	55	0	0	0
2	500	1	0,1	0,5
3	2	0	0,1	0,333
4	23	1	0,2	0,5
5	72	1	0,3	0,6
6	79	0	0,3	0,5
7	82	1	0,4	0,571
8	215	0	0,4	0,5

Task 3 : Interpolated Precision

1. What is interpolated precision?

   Interpolated precision smoothes the precision/recall graph. This is achieved by letting the precision be the maximum of all future points. This allows computing the precision at recall-level 0.

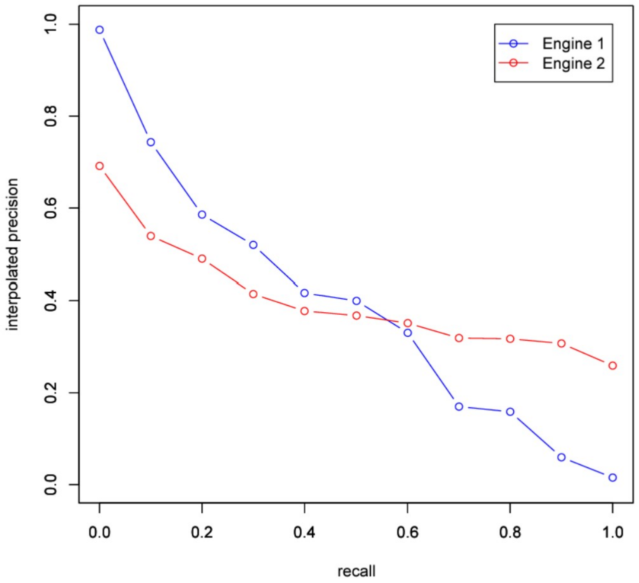
2. Given the example in Task 2.2, find the interpolated precision and make a graph.



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?

   I would prefer engine 2. If I look for **all** published work, it isn't super-important that the very first results aren't relative, as long as all relevant results can be found. Engine 1 finds good relevant results at the start, but the quality of the search decreases quickly. Engine 2 maintains the relevance of the retrieved documents much better as the recall increases and it would therefore be easier to find all the relevant documents I'm looking for.

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.

   > **False** - Harmonic mean (HM) is used because it's closer to the minimum when two values differ greatly compared to Arithmetic Mean (AM). If for example the recall is 1.0 and the precision is 0.0, AM would be 0.5 meaning that the result is 50% correct, even though it isn't correct at all. The HM would be 0.0 which represents the result better.

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

   > **False** - In single documents in a corpus with an infinite number of documents, we'll still not observe all words in the vocabulary. This means that smoothing is needed to estimate the document language model.

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

   > **False** - The goal of the retrieval models we've learned about is to satisfy the information need of the user using the model, the evaluation metrics are used to tell how good a certain model is.

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.

   False - You can't safely say that B is better than A because B might be better optimized for the evaluation metrics than A and the result might change if you use something like NDCG or MRR.

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

   True