Language models and evaluation in IR

Task 1 - Relevance Feedback

1. Explain the difference between automatic local analysis and automatic global analysis.

With automatic local analysis, we dynamically determine similar terms based on analysis of top-ranked retrieved documents at query time. The base correlation analysis is only performed on a local set of retrieved documents for a specific query. It avoids ambiguity by determining similar (correlated) terms only within relevant documents. This can be an example:

$"Apple\ computer" \rightarrow "Apple\ computer\ Powerbook\ laptop"$

Then to the difference between local and global: In automatic local analysis, documents retrieved are examined to automatically determine query expansion, i.e. no relevance feedback is needed. In automatic global analysis, on the other hand, you can perform a global analysis once and produce a thesaurus (synonyms dictionary), and use it for expanding queries with similar words later on.

2. What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?

The main idea behind relevance feedback is to take the result set initially returned from a given query, gather user feedback, and to use information about whether or not the result set are relevant to use in a new query. Or rephrased as *involving the user in the retrieval process as to improve the final result set.*

Query Expansion consists of selecting and adding terms to the query with the goal of minimizing query-document mismatch and thereby improving retrieval performance. In other words, it is done by evaluating the users input and expanding the query for better retrieval performance.

Term Re-weighting is about making weights of unrelevant terms less significant, and increasing the weight of relevant terms. The difference between Query Expansion and Term Re-weighting is that the process of reweighting does not include expanding the query, only making some terms less/more significant based on relevance.

Task 2 - Language Model

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

A language model is a probability distribution over sequences of terms. Given such a sequence, it assigns a probability to the whole sequence. In the query likelihood model in IR, a language model is constructed for each document in the collection. It is then possible to rank each document by the probability of specific documents given a query. This is interpreted as being the likelihood of a document being relevant given a query.

The strengths is that the model is intuitive, mathematical precise and conceptually simple. The weaknesses on the other hand, is that it is hard to take into account the users' wishes, and improve relevancy.

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

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

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

First of all, I want to explain the different elements of smoothing function, and count the terms per document

$\hat{P}(t|M_d)$ means the probability of the term to occur in the model for each document.

$\hat{p}_{mle}(t|M_d)$ means the probability of the term to occur in the model based on the current document. Practically speaking, this is done by dividing the number of occurrences of the selected term, by the number of therms in the document.

$\hat{p}_{mle}(t|C)$ means the probability of the term to occur in the whole collection. Practically speaking, this means you divide the number of occurrences in the whole collection based on all terms in the collection.

So, to the count of terms per document, which will be used in the later typing:

$d_1 = 8\ terms$

$d_2 = 12\ terms$

$d_3 = 7\ terms$

$Total = 27\ terms$

q1 = doctor

$\hat{P}(q_1|d_1) = (1-0.5) * \frac{1}{8} + 0.5 * \frac{2}{27} = 0.099537$

$\hat{P}(q_1|d_2) = (1-0.5) * \frac{1}{12} + 0.5 * \frac{2}{27} = 0.078703$

$\hat{P}(q_1|d_2) = (1-0.5) * \frac{0}{7} + 0.5 * \frac{2}{27} = 0.037037$

Based on the query and the generated number of relevance the ranking would be $d1 > d2 > d3$

q2 = apple orange

$\hat{P}(q_2|d_1) = ((1-0.5) * \frac{1}{8} + 0.5 * \frac{2}{27}) * ((1-0.5) * \frac{0}{8} + 0.5 * \frac{0}{27}) = 0$

$\hat{P}(q_2|d_2) = ((1-0.5) * \frac{0}{12} + 0.5 * \frac{2}{27}) * ((1-0.5) * \frac{0}{12} + 0.5 * \frac{0}{27}) = 0$

$\hat{P}(q_2|d_3) = ((1-0.5) * \frac{1}{7} + 0.5 * \frac{2}{27}) * ((1-0.5) * \frac{0}{7} + 0.5 * \frac{0}{27}) = 0$

Based on the query and the generated number of relevance the ranking would be $d1 = d2 = d3$

q3 = doctor apple

$\hat{P}(q_1|d_1) = (1-0.5) * \frac{1}{8} + 0.5 * \frac{2}{27}) * (1-0.5) * \frac{1}{8} + 0.5 * \frac{2}{27}) = 0.009907$

$\hat{P}(q_1|d_2) = (1-0.5) * \frac{1}{12} + 0.5 * \frac{2}{27}) * (1-0.5) * \frac{0}{12} + 0.5 * \frac{2}{27}) = 0.002914$

$\hat{P}(q_1|d_2) = (1-0.5) * \frac{0}{7} + 0.5 * \frac{2}{27}) * (1-0.5) * \frac{1}{7} + 0.5 * \frac{2}{27}) = 0.004017$

Based on the query and the generated number of relevance the ranking would be $d1 > d3 > d2$

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

subtask.

Smoothing is used for removing the null-values from the search, i.e. terms not appearing. This is done for creating a smoother model with less noice. Jelinek-Mercer smoothing uses a λ for determining a query terms' effect on the search. An example of this is q_2, where all the ranks are 0 as orange is not occurring in the documents.

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.

Precision is the amount of retrieved documents which is relevant, in other words, how accurate the system is in the documents it returns (1.0 means it only returned the relevant documents)

$Precision = (relevant\ items\ retrieved)/(retrieved\ items) = P(relevant|retrieved)$

Recall is the amount of relevant documents which is retrieved, in other words, what percentage of the relevant documents the system found (1.0 means it found them all)

$Recall = (relevant\ items\ retrieved)/(relevant\ items) = P(retrieved|relevant)$

To evaluate the quality of these metrics, we can use the following table:

	Relevant	Nonrelevant
Retrieved	True positives (TP)	False positives (FP)
Not retrieved	False negatives (FN)	True negatives (TN)

One may then see that the formulas for quality in $Precision$ and $Recall$ is:

$Precsion = \frac{TP}{(TP+FP)}$

$Recall = \frac{TP}{(TP+FN)}$

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

For answering on the question about pros and cons, I wanted to read a little bit more about the ranking methods, and read this article by Dr. Moussa Taifi.

The Mean Reciprocal Rank is the simplest metric of the two. It tries to measure "Where is the first relevant item". It is closely linked to the binary relevance family of metrics.

Mean Reciprocal Rank (MRR) pros:

- Simple to compute and easy to interpret
- Good focus on first relevant element, therefore best suited for targeted searches such as users asking for the "best item for me"

Mean Reciprocal Rank (MRR) cons:

- It focuses on a single item, and does not evaluate the rest of the list of recommended items.
- Might not be a good evaluation metric for users that want a list of related items. The goal of the users might be to compare multiple related items.

The Mean Average Precision metric, on the other hand, tries to approximate the weighting sliding scale for cutting the errors in the first few elements rather than much later in the list. The goal is to weight heavily the errors on top of the list, then gradually decrease the significance of the errors as we go down the lower items in a list.

Mean Average Precision (MAP) pros:

- Gives a single metric that represents the complex area under the Precision-Recall curve, which provides the $(average\ precision)/(list)$.
- This metric is able to give more weight to errors that happen high up in the recommended lists. Conversely, it gives less weight to errors that happens deeper in the recommended lists. This matches the need to show as many relevant items as possible high up the recommended list.

Mean Average Precision (MAP) cons:

- This metric shines for binary, relevant or non-relevant, ratings. However, it is not fit for fine-grained numerical ratings. This metric is unable to extract an error measure from this information.
- With fine-grained ratings, for example on a scale from 1 to 5 stars, the evaluation would need first to threshold the ratings to make binary relevancies. One option is to consider only ratings bigger than 4 as relevant. This introduces bias in the evaluation metric because of the manual threshold. Besides, we are throwing away the fine-grained information. This information is in the difference between a 4 and 5 stars ratings, as well as the information in the non-relevant items. This makes a 1 star rating count the same as a 3 star rating, which is unfortunate.

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.

Copy of Table

d	Relevant	Precision	Recall
55			
500	REL	1/2	1/10
2			
23	REL	2/4	2/10
72	REL	3/5	3/10
79			
82	REL	4/7	4/10
215			

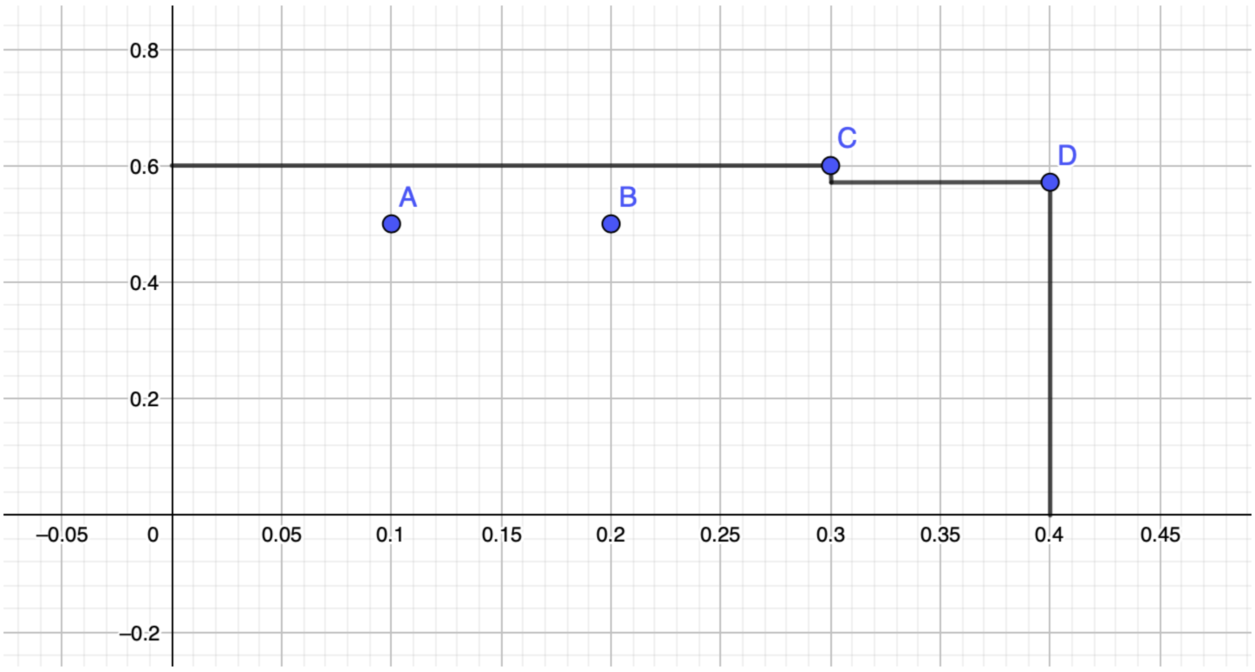
In the table above, each term occurring in both sets are given the value REL. The precision is calculated iteratively, being the number of relevant items based on the retrieved terms up to and including this term. The recall is calculated by looking at the amount of relevant items gathered so far based on the total number of relevant terms.

Task 4 - Interpolated Precision

1. What is interpolated precision?

Interpolated precision is where you pick a recall level r and for all recall levels $P(r') >= P(r)$, where $P(r)$ is the precision rank at r. It is the best precision you can achieve. Then in 11-pt interpolated average precision, you are looking at 11 recall levels $(0.0, 0.1, 0.2, ..., 1.0)$ and finding the interpolated precision at each point.

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



References

mAP (mean Average Precision) might confuse you!

One can be forgiven for taking mAP (mean average precision) to literally mean the average of precisions. Nevertheless, you couldn't be further from the truth! Let me explain. In computer vision, mAP is a popular evaluation metric used for object

https://towardsdatascience.com/map-mean-average-precision-might-confuse-you-5956f1bfa9e2

MRR vs MAP vs NDCG: Rank-Aware Evaluation Metrics And When To Use Them

Reporting small improvements on inadequate metrics is a well known Machine Learning trap. Understanding the pros and cons of machine learning (ML) metrics helps build personal credibility for ML practitioners. This is done to avoid the trap of prematurely

https://medium.com/swlh/rank-aware-recsys-evaluation-metrics-5191bba16832

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/8a96d80f-0341-4d52-a430-a96299baf767/Informasjonsgjenfinning_ving2.ggb