TDT4117 Information Retrieval - Autumn 2021
Assignment 2 - Solution

November 19, 2021

Task 1 - Relevance Feedback

1. Explain the difference between automatic local analysis and automatic global analysis.
   Answer: Global methods are techniques for expanding or reformulating query terms independent of the query and results returned from it. Therefore they expand the query using information from the whole set of documents.
   Local methods adjust a query relative to the documents that are initially retrieved for the query.

2. What is the purpose of relevance feedback? Explain the terms Query Expansion and Term Re-weighting. What separates the two?
   Answer: Relevance feedback refers to a feedback cycle in which documents that are known to be relevant to the current query q are used to transform it into a modified query q_m. Then, it is expected that the query q_m returns a higher number of documents relevant to q.
   Query expansion and term re-weighting are two methods of relevance feedback. Query expansion is based on adding new terms to the query extracted from relevant documents. Term re-weighting is based on user relevance judgements, where it increase weight of terms in relevant documents and decrease weight of terms in irrelevant documents.

Task 2 - Language Model

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

   Answer:

   Language modeling approach builds a probabilistic language model M_d from each document d, and ranks documents based on the probability of the model generating the query: $P(q|M_d)$.

Language models provides effective retrieval and they are conceptually simple and explanatory. Without an explicit notion of relevance, relevance feedback is difficult to integrate into the model. Current LM approaches use very simple models of language, usually unigram models.The assumption of equivalence between document and information need representation is unrealistic.

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

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

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

Answer:

-	q_1	q_2	q_3
d_1	0.099	0.0	0.0098
d_2	0.078	0.0	0.0028
d_3	0.037	0.0	0.0040

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

Answer: Smoothing is the process of estimating the distribution of query terms that are not in the document ($P_{\not\in}(k_i|M_j)$). Smoothing is important for fine tuning the language model ranking and avoids assigning zero probability to the terms that are not in the document. In the above example, smoothing avoids that $P(q_1|d_3)$ becomes zero.

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.

Answer:

Precision (P) is the fraction of retrieved documents that are relevant. Recall (R) is the fraction of relevant documents that are retrieved.

$$\text{Precision} = \frac{\text{relevant items retrieved}}{\text{retrieved items}} = P(\text{relevant}|\text{retrieved})$$

$$\text{Recall} = \frac{\text{relevant items retrieved}}{\text{relevant items}} = P(\text{retrieved}|\text{relevant})$$

The two quantities clearly trade off against one another: one can always get a recall of 1 (but very low precision) by retrieving all documents for all queries. Recall is a non-decreasing function of the number of documents retrieved. On the other hand, in a good system, precision usually decreases as the number of documents retrieved is increased.

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

Answer:

MRR is short for mean reciprocal rank. It is also known as average reciprocal hit ratio (ARHR) which tries to measure Where is the first relevant item?. It is closely linked to the binary relevance family of metrics. The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries Q:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}.$$

where rank i refers to the rank position of the first relevant document for the i-th query.

The reciprocal value of the mean reciprocal rank corresponds to the harmonic mean of the ranks.

MRR Pros: 1)This method is simple to compute and is easy to interpret. 2)This method puts a high focus on the first relevant element of the list. It is best suited for targeted searches such as users asking for the best item for me. 3)Good for known-item search such as navigational queries or looking for a fact.

MRR cons: 1)The MRR metric does not evaluate the rest of the list of recommended items. It focuses on a single item from the list. 2)It gives a list with a single relevant item just a much weight as a list with many relevant items. It is fine if that is the target of the evaluation. 3)This might not be a good evaluation metric for users that want a list of related items to browse. The goal of the users might be to compare multiple related items.

MAP:

Mean average precision (MAP) for a set of queries is the mean of the average precision scores for each query.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

where Q is the number of queries.

MAP pros:

1)Gives a single metric that represents the complex Area under the Precision-Recall curve. This provides the average precision per list. 2)Handles the ranking of lists recommended items naturally. This is in contrast to metrics that considering the retrieved items as sets. 3)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.

MAP cons:

1)This metrics shines for binary (relevant/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. 2)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.

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.

   Answer:

Ret	Rel	Recall	Precision
55	-	0.0	0.0
500	yes	0.1	0.5
2	-	0.1	0.33
23	yes	0.2	0.5
72	yes	0.3	0.6
79	-	0.3	0.5
82	yes	0.4	0.57
215	-	0.4	0.5

Task 4 - Interpolated Precision

1. What is interpolated precision?

   Answer:

   Interpolated precision p_{interp} at a certain recall level r is defined as the highest precision found for any recall level $r' \geq r$:

$$p_{interp}(r) = \max_{r' \geq r} p(r').$$

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

Recall	i-Precision
0.0	0.6
0.1	0.6
0.2	0.6
0.3	0.6
0.4	0.57
0.5	0.0
0.6	0.0
0.7	0.0
0.8	0.0
0.9	0.0
1.0	0.0