Information Retrieval, Exercise 3

- Florian Eckerstorfer, florian@eckerstorfer.co
- Haichao Miao, hmiao87@gmail.com

Setup

We started the evaluation process by writing a script (`eval_all`) that executes `trac_eval` for every run (that is, every query executes with small, medium and large threshold indexes) we created during exercise 1 and 2.

In the next step we created a script to generate a CSV file (`build_csv`) that contains all measures from all runs. That way we could use standard software like Excel or Numbers to analyze our retrieval engine.

Due to a bug in the code we submitted for exercise 1 we had some `NULL` values as document ID in some of the results from exercise 1. We removed lines with `NULL` values from the result files, therefore some runs in exercise 1 contain less than 10 results.

Analysis

After we opened the data we could see that the number of retrieved and related documents is most of the time `0` for our retrieval engine from exercise 1. In fact, only for topic 1 our retrieval engine from exercise 1 returned any relative documents. The measures for this one topic are not very good for all three index sizes. We will therefore focus on our results from exercise 2 in the rest of our analysis.

From the runs executed in exercise 2 `18` retrieved no related documents and `33` retrieved at least one relevant document. In this case we can also see a difference regarding the index size. The runs executed against the large and medium indexes could not retrieve relevant documents for 5 queries, while the runs executes against the small index could not retrieve relevant documents for 8 queries.

The next value we looked at was Mean Average Precision (MAP). MAP basically tells you how good the relevant documents are ranked in your result. That is, if `map = 1.0` then all relevant documents are ranked on top in the result. One run produces a `map` of `1.0` and that is for topic 10 in the large index. Both relevant documents are ranked in our result in the same order as they are ranked in the golden standard evaluation.

However, the MAP measure falls quickly to a value below `0.5` . In this example it is a little bit hard to measure how good the system is based on MAP, because often two or three as

relevant evaluated exist in the collection. A MAP value of `0.5` with two relevant document is probably not statistically significant.

When we look at the Reciprocal Rank (`recip_rank`) measure our algorithm is quite good. In 23 runs our algorithm ranked the most relevant document first and only in 4 runs the most relevant document is not in the top 3.

Another measure we analyzed is R-Precision, that is, the precision after `r` documents (where `r` is the number of relevant documents in the collection) are retrieved. For our system this value is very similar to MAP, mostly because our system is good at retrieving the most relevant document as top result, but is not as good in retrieving other relevant documents.

At the end we also took a close look at the effect of the threshold on the quality of the results. In exercise 2 we retrieved relevant documents in 36 runs. When the sort these runs by MAP we can see that runs on the large or medium index are always a lot better than runs on the small index. The number of retrieved relevant documents is only different for three topics when we compare the large and medium index. However, we can see higher MAP values for the large index compared to the MAP values of the medium index for some topics. Our algorithm therefore can calculate a better ranking due to the larger amount of available terms in the index. But there is also one topic where the medium index performs best (topic 14) and one where the small index performs best (topic 11, all three runs are not very good).

The analysis we did was comparing the different evaluations from the human evaluators. For our retrieval system we could not see any differences. For every topic the measures of all evaluations were exectly the same.

Conclusion

We can see that our naive implementation in exercise 1 did not perform very well. Our retrieval system returned relevant (as defined in the golden standard) for one topic and it performs bad. However, we can see a great improvement in the performance of our system since we started using BM25 in exercise 2.

While the performance is quite good when we look only at the top 3 results (which, for example, in web search are the ones that matter most), our system is not that good when we look at all retrieved documents. However, one possible reason for this is that are often only very few as relevant evaluated (by the golden standard) documents for each topic. Therefore the average precision of ten results when only two relevant topics exist is not very meaningful.