Information Retrieval

Chiel Kooijman (5743028), Michael Cabot (6047262)

November 11, 2012

1 Introduction & Method

The goal of this paper is to describe our familiarisation process with Information Retrieval techniques. To this purpose we have used *Indri* for indexing and running queries. *trec_eval* was used to evaluate the results. We used two document sets containing English language news papers used at earlier editions of CLEF[1]. The Porter stemmer[1] and a small list of stop words have been used to index the documents in different combinations. Four different preprocessing methods were used:

no preprocessing All words in the documents are indexed.

stopper The following three stop words are not indexed and are removed from the queries: *a, the, of.*

stemmer The documents and queries are stemmed with the porter stemmer.

stopper & stemmer Both a stopper and stemmer are used.

It is important to notice that a preprocessing method applied when indexing should always also be applied to the queries. Queries that only contain words that are stemmed in the documents but are not stemmed in the queries will not give any results since they were not indexed. Similarly, a query containing stop-words will not give any results if these stop-words are not indexed. Preprocessing only applied when indexing whould thus result in a much lower performance. Indri automatically applies the preprocessing methods to the queries that were used when indexing.

trec_eval was used to evaluate the performance for the four different preprocessing methods mentioned above. The performance is measured with the Mean Average Precision (MAP) because it takes into account the order in which the results are returned, i.e. the MAP will be higher if the first result is a relevant result than if this result if placed on a lower position. MAP is also known for its stability and counts as a standard in Information Retrieval literature. Figure 1 shows the MAP scores for the four different preprocessing methods. Table 1 shows the values. The stopper slightly decreases the performance of the indexer both with and without stemming. The stemmer increases the MAP performance.

	Without Stemming	With Stemming
Without Stopping	0.2311	0.2855
With Stopping	0.2308	0.2849

Table 1: MAP Scores

[1] Cross Language Evaluation Forum

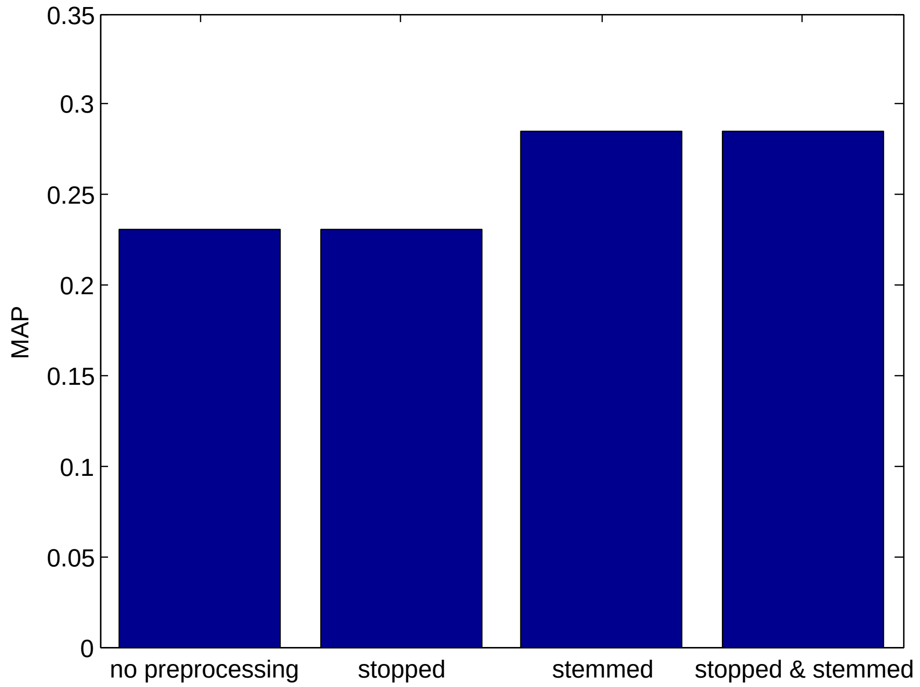

Figure 1: The MAPscores of four different preprocessing methods.

2 Conclusion

We familiarized ourselves with Information Retrieval by indexing documents and retreiving them with queries. The MAP scores of four different preprocessing methods were compared with each other. The results show that stemming increases the performance and removing stop-words slightly decreases the performance.

References

[1] M.F. Porter et al. An algorithm for suffix stripping, 1980.