Masha Lifshits and Amrita Reddy
Lab 3 for CS 341

Report

Introduction

This report describes the implementation of three algorithms for information retrieval and determines which one is the best algorithm by comparing the mean average precision of the algorithms. The best algorithm is named ***modified-tfidf*** and it has been implemented with the removal of stop words and references, and the use of letter n-grams.

Description of the overlap and tfidf algorithms

We implemented the ***overlap*** algorithm by creating a dictionary whose keys are the query ids and whose values are the list of tokens of each query. We did the same for the documents. We then looked to see if the tokens in each query were found in each of the documents and computed the score for each query-document pair.

We implemented the ***tfidf*** algorithm by creating a dictionary whose keys are the types in the corpus and values are dictionaries where the document id is mapped to the raw tf. We used this to calculate the tfidf of documents to queries. We initially tried to use a dictionary of lists but realized that a dictionary of dictionaries was more efficient.

Description of the modified-tfidf algorithm (best)

We modified the tfidf algorithm in the following ways:

- Removed stop words from both the documents and the queries (we used a list of stop words that Heather had given us for NLP and modified it by adding words that we thought would qualify as stop words for this particular corpus).
- Removed the references at the end of the documents (they did not carry any relevant information) and removed IDs from queries and documents - both of them improved the scoring because that increased the relevance of the query and document
- Used n-grams (n=14) - ended up giving the best average result, potentially, because the collection is mostly scientific and the long words matter more
- Normalize the stop words - helped, because when we normalized the stopwords were able to match the tokens better
- Round the score - helped, because averaged out the difference

We also tried using different numbers for letter n-grams, as well as manipulating the evaluation function and using normalized tf on queries terms, however, that resulted in a decrease of the mean average precision. We also tried removing tokens that have a length of three or less under the assumption that these tokens were not important. We found that this was not true.

Results:

Algorithm	MAP(%)
overlap	15.16
tfidf	30.90
modified-tfidf	35.51

Table 1. Mean average precision for the three algorithms that were implemented

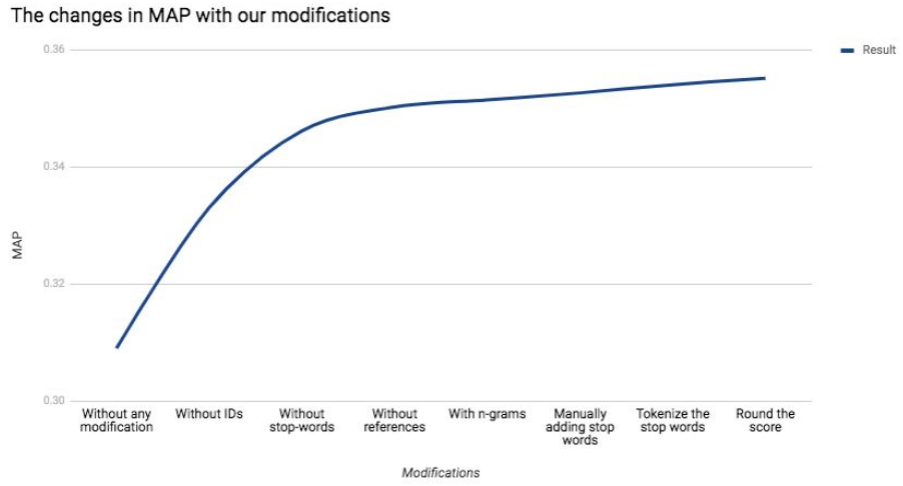

Figure 1. The changes in MAP for tfidf with each modification that we made

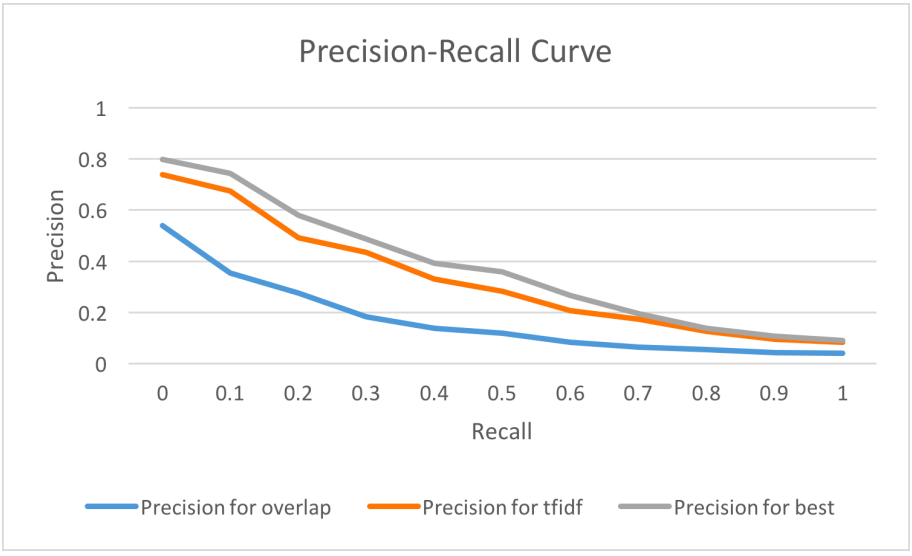

Figure 2. Precision-Recall plot for the three algorithms that were implemented