Julian Waksberg 000801926
Dor Hanegby 204009633

Home Assignment No. 3

Part 3: The report

1. When we started, we decided to do a mix between the stop words defined in lucene and the ones we found in the corpus.
    a. "an are be if into it no not or such their then there these they this will" ***In lucene but not in corpus most frequent***
    b. "the of a to in and for that with was on last by but as at his from week is" ***most frequent in corpus***

2. We made the list of stop words static, once we calculated we saw no need to calculate it in each run of the search
3. When we started to write the evaluation of the system, since we didn't have enough information about the corpus or the users of the system we decided to set α to *0.5* so it will evenly consider recall and precision.

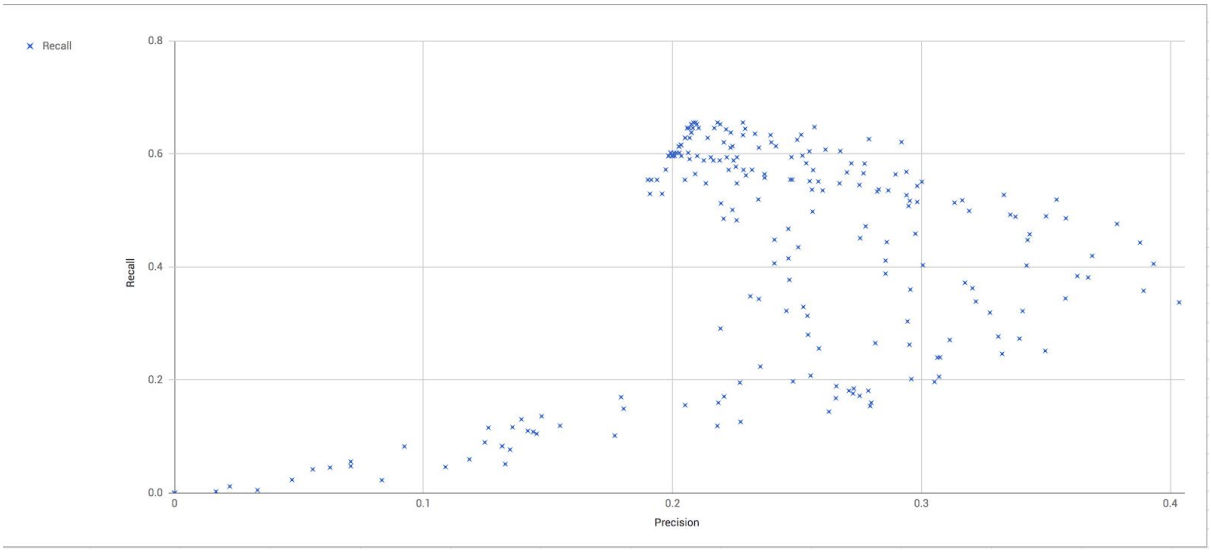
Improved Algorithm

1. We decided to encode all the known formulas for **tf** and for **idf** and after that try to maximize the F-Score
2. After running several iterations and doing analysis on the values we found that the best values where (you can find in Report analysis all the combinations) :
    a. TF: Log normalization
    b. IDF: Probabilistic
    c. Threshold: 8.699999999999985
    d. F-Score: 0.40191233895103873
3. Then we applied to this configuration a Porter stemmer to see if it would improve the performance of the search engine. The F-score was : 0.37707521593424237 (yielding a difference of 0.02483712302 against our best F-Score)
4. We also tried to change the stop words to see if it would have any effect on the system, no improvement was found out of this experiment.

Julian Waksberg 000801926
Dor Hanegby 204009633

Graphs:

Precision - Recall comparison



Threshold - F