University of Cape Town

Information Retrieval

CSC4000W

Assignment 1b Report

Author:	*Student Number:*
Alon Bresler	BRSALO001
Andreas von Holy	VHLAND002
Osher Shuman	SHMOSH001

May 18, 2016

1 System Design

1.1 Blind Relevance Feedback

...

1.2 AND Search

...

1.3 Stop Word

Our system has the option to ignore stop words when performing the searching function. The stop words are ignored in two features of our search system – ignoring stop words in the query and ignoring stop words in the documents when doing blind relevance testing.

1.4 Thesaurus

A thesaurus is implemented, using a Natural Language Tool-kit (NLTK) with a WordNet package, to get synonyms for the search terms. The synonyms are stemmed using Porter's stemming algorithm, if the parameter is set true. The thesaurus results for the search term is used to recall index files which are synonymous to the search term. When calculating the $tr * idf$ value for the synonymous term, it a factored down by multiplying the result by $1/(numSynonyms)$.

1.5 Title

This feature was initially added, but removed as not all the test beds had documents with titles. This feature compared the query to the title of the document, and the ranking of the document was increased depending on the similarity of the query to the title.

2 Performance Results

2.1 Mean Average Precision (MAP)

2.2 Normalized Discounted Cumulative Gain (NDCG)