CS2108 Lab 1 report

He Yuchen
A0162473E

This lab used the methods of relevance feedback and query expansion to enhance the accuracy of text retrieval.

In the first function, where only relevance feedback is implemented, the query adds the most relevant document vectors times a constant alpha and subtracts the irrelevant ones times a constant beta. The relevance can be determined from the similarity matrix. Here we use alpha = 0.8 and beta = 0.2. Option of getting the input info is added to check the effects of different iterations on the result of accuracy. It turns out that more iterations are performed, more accurate results will be gotten. Yet a potential problem is that the original query will be heavily modified that it may no longer represent what the user originally means.

In the second function, query expansion is added so that the query not only takes the relevant vectors but also the top terms included in those vectors. As we get the top terms using the same argsort function in the similarity matrix, those terms can be looked up inside the inverted dictionary of document vocabularies. After getting the set of top terms for each query, the set of top terms are then transformed to a tfidf vector and then added back to the query.

The result is that, pure relevance feedback increased the accuracy from 0.518385904086 to 0.626723219321 when only 1 iteration is performed, and the query expansion further optimized this value to 0.629641195932.