**CS 490 – Web Information Retrieval**

**Assignment 1**

Matt McCormick

[mbmccorm@purdue.edu](mailto:mbmccorm@purdue.edu)

September 25, 2012

**Part 1**

The results obtained using the RawTF retrieval algorithm for the text data through stemming and stopwords removed were more precise. The output from the trec\_eval command shows that the R-Precision for this set compared to the non-stemmed, stopwords included set is 0.1404, nearly twice as precise. Stemming includes similar relevant queries such as the plural or verb forms of a word. Removal of stopwords helps remove common words such as “where”, “how”, etc. These two refinements help improve the result set.

**Part 2**

The results obtained using the Okapi retrieval algorithm for the text data through stemming and stopwords removed were more precise. The output from the trec\_eval command shows that the R-Precision for this set compared to the others is 0.3813, much higher the other sets. The results obtained using the RawTFIDF retrieval algorithm for the text data through stemming and stopwords removed was the next most precise. The output from the trec\_eval command shows that the R-Precision for this set compared to the RawTF retrieval algorithm is 0.2386, just under twice as precise.

**Part 3**

The results obtained using the Okapi retrieval algorithm for the text data from the Purdue CS website through stemming and stopwords removed appeared to be more precise. Looking at the data generated by the ParseToFile command, you can see that the RawTF and RawTFIDF results are not very useful as it pertains to the query. They both start with a survival guide and then proceed to rank several publications from different home pages at the top. However, the Okapi results list the curriculum information and prominent news listings towards the top of the set.

**Part 4**

I implemented the Glasgow weighting scheme for my custom retrieval method. The results obtained using the Glasgow retrieval algorithm for the text data through stemming and stopwords removed were roughly the same precision as the Okapi retrieval method. The output from the trec\_eval command shows that the R-Precision for this set compared to the others is 0.3766, much higher the other sets but on par with the Okapi retrieval method.

I implemented a custom adjusted score method for my custom retrieval method. My custom adjusted score takes into account the number of unique queries and factors this in based on the average length of the documents in the set. This is then multiplied by the original adjusted score value.