Peter Jough

IS392 Report

For this assignment, I used the provided HTML files. For each page, I would start by removing the script and style tags as they would not contain any information that's needed for this indexing task. Looking through a few of the Wikipedia pages, I noticed that the Wiki pages all had the content within the div tag with an id of content/relevant data. Inside that div tag, all of the actual content would be in paragraph tags. So I gathered all the inner HTML of the paragraph tags and combined them into one giant string. This allowed me to run some string manipulation commands to get rid of any extraneous information that isn't important. I removed most the symbols, "[edit]"s, and set all the letters to lowercase. Some of the symbols are left in there as well as the numbers, because they could be symbols from other languages and be somewhat important. Afterwards, I split the text into single word tokens, separated by any whitespace. Using a for loop, I was able to store the location of each token by the index of the token array. I used a similar for loop to store the frequency of each token. After those two for loops, I created two more for loops to transfer the data from the local dictionaries to the main dictionaries. When the most outer for loop finishes, then all 476 files will have been processed and the location of every word as well as their respective frequency will be stored in the main dictionaries. After that I moved the information into some output files. The final format of the information is as follows: term: website : location/frequency.

Statistics:

Total number of unique terms: 27974

Term that occurred in most files : ‘the’ at 469 files

Most frequently used word: ‘the’ at 50809 times

10994 terms occur exactly once