**Vincent Nkawu**

**Assignment #5**

**Due Date: October 16, 2014 (Excused for late Submission)**

**Information Retrieval**

**Dr. Rus**

Problem 1 [40 points].

Automatically collect from memphis.edu

**10,000 unique documents**.

The documents should be proper (>50 valid tokens   
after saved as text) .html, .txt, and .pdf files.

That is, your output should be a set of 10,000 text files (not html, txt, or pdf docs).

You must write your own code to collect the documents - DO NOT use an existing crawler.  
  
Store for each proper file the original URL as you will need it later   
when displaying the results to the user.  
  
Problem 2 [20 points]. Preprocess all the files using assignment #4. Save all  
preprocessed documents in a single directory which will be the input to  
the next assignment, index construction.

For this assignment, firstly I extract the links from the memphis.edu domain using HTML::LinkExtor, then I retrieve all the stop words from the page given to us in class

"<http://www.cs.memphis.edu/~vrus/teaching/ir-websearch/papers/english.stopwords.txt>".

All the stop words are stored in a hash to make it easier or rather faster to remove the stop words when processing the documents. Last home work I did not use a hash and had to loop through all stop words each time I was checking. But this time I simply check to see if a word is part of the hash using a hash of hash and assinging value to one to each word (key) than is a stop word (that is the words in the hash) So I simple check the the hash feeding it a key and checking the value. If the key is not one then I know it is not part of the hash so it is not a stopword.

So I used a subroutine to facilitate the deletion of stopwords using the mechanism explained above, then another subroutine to manage all links so that we can keep track of links visited and links to be visited. Another subroutine is used to preprocess the file, depending on the whether it is a txt file, pdf file or html file. The html file are processed similar to previous home work. And finally I stemm the document using porter stemmer just like I did in previous home work.

Last thing to note. It may take you forever trying to sit and wait for the entire program to run, crawl and preprocess documents. But If you would like to, change the directory location for variable newDir. So that the preprocessed documents will be stored there.