Information Retrieval 8130

Assignment 6

Submitted by: Hasan Mashrique

**Due Date:** 11/07/2022

**Problem 1 and problem 2:**

The code for the homework combines the work of problem 1 and problem 2 to produce the final 10,000 vocabulary files output.

The code does the following tasks:

1. Crawls the URL “www.memphis.edu” and gets all the links from the home page.
2. Crawls each link from the homepage and adds all the pages from those links. This continues until total link count > 10500.
3. For each link, crawls the link page, tokenizes the vocabulary from the page, preprocess (stopword removal, lemmitize) the vocabulary and generate the cleaned vocabulary for that link.
4. Saves the vocabulary into a file with the link name in it.

The link to all the saved vocabulary files is here:

(

<https://www.dropbox.com/scl/fo/ygu1g2zyi9clmgfrt71o5/h?dl=0&rlkey=hkyz3ggg0wphmfdozv8k8n2g3>

)