Information Retrieval 8130

Assignment 4

Submitted by: Hasan Mashrique

**Due Date:** 10/12/2022

**Problem 1:**

**The code:**

* Extracts 10 articles from the news.yahoo.com website.
* Reads the content of the articles and writes them into 10 text files.
* Reads the text files into a string.
* Does the processing of the text by tokenizing, removing stopwords and lemmatizing.
* Outputs the final processed words.

**Input:**

URL: <https://news.yahoo.com> is given as input inside the code.

**Functions used:**

* tokenize\_words() : converts the text file contents into tokens
* remove\_stopwords(): removes the stopwords from the tokens
* lemmatize\_words(): lemmatizes the remaining words

**Output:**

* HTML links to parse
* Word token size after initial parsing
* unique token size
* token size after removing the stopwords
* token size after lemmatizing
* Final tokens after processing