CS 4364/5364: INFORMATION RETRIEVAL AND INFORMATION VISUALIZATION
DEPARTMENT OF COMPUTER SCIENCE
THE UNIVERSITY OF TEXAS AT EL PASO
FALL 2021

Assignment# 2: Pre-processing

Due: October 30, 2021, in Blackboard

Given a text file (input.txt), write a program (in your language of choice) to pre-process the file. The program should

a. Tokenize the lines of the file based on white-spaces.
b. Case-fold each token to lowercase.
c. Identify the stop-words. Use the stop words from the stop_words.txt file.
d. Remove all punctuation.
e. Use Porter stemmer* to stem the tokens.
f. Write each token on a separate line in the output file.
g. Write the type of token - for all non-empty token, next to the token, in the same line of the output file, write 'article' if the token is an article (i.e., a, an, the), 'stop word' if the token is a stop word, otherwise, write the *stemmed* version of the token.

Sample input file:

The less there is to justify a traditional custom, the harder it is to get rid of it.

Sample output file:

the	- article
less	- less
there	- stop word
is	- stop word
to	- stop word
justify	- justifi
a	- article
traditional	- tradit
custom	- custom
the	- article
harder	- harder
it	- stop word
is	- stop word
to	- stop word
get	- get
rid	- rid
of	- stop word
it	- stop word

Deliverables: 2 files

1. The source code of the program. The name of the file should be lastname_firstname.xyz (replace xyz with proper extension).
2. The output txt file. The filename should be lastname_firstname.txt

*https://tartarus.org/martin/PorterStemmer/