COSC 1336
Spring 2023

Assignment 9: Files and Text Processing

[1] **Objectives**: This assignment continues with the last assignment with more string processing and file operations. Many AI applications require us to process words, but some words don't mean anything in isolation. So, we are removing these words from the last assignments.

[2] **Description**: Stop-words are commonly used words in any language, not just English. Stop-words are critical to many applications because if we remove the words commonly used in a given language, we can focus on the important words instead. (See: https://kavita-ganesan.com/what-are-stop-words/ for why stop-word removal is important). A short stop-word list will be provided for your use in this assignment (stopword.txt). A recent Wall Street Journal news article will be used for word processing (wsj.txt). The news article has been edited somewhat because (a) it contains some non-ASCII characters and (b) to simplify processing for you. For example, hyphenated words are separated with a space.

[3] **Requirements:** The following is a list of additional work required for this assignment. You should reuse the code you developed for the previous assignment, including all the functions.

- You will need a new function to save the result to a file (see output for format). The function takes a file object and the word-count list.
- You may have to modify one or more functions to process the words further. You should output the most frequently used (defined as words with a count greater than 2) words first. Then, print the list again after removing the stop-words.
- Finally, save the words and counts (without stop-words) into a text file (prog9out.txt). See the output at the end of the assignment for detail.

I estimate that you may have to write about 20 new lines of code. You should reuse as much of your code in Assignment 8.

[4] **Output**: A sample output is given below. The list is long, so I showed only the first 50 lines. You should test for all possible input. You should note that I added [ ] around the words to ensure I am not getting some white characters in the word. Note that the saved file looks slightly different.

[5] **Deadline**: 11:59 pm, Monday, April 24, 2023

```
Enter the text file name to process: wsj.txt
```

```
Length of the list: 69
Count | Word
-----   ---------------
   44 | [the]
   39 | [to]
   26 | [of]
   25 | [and]
   16 | [a]
   13 | [said]
   13 | [that]
    9 | [attack]
    9 | [citizen]
    9 | [for]
    9 | [lab]
    9 | [spyware]
    8 | [as]
    8 | [by]
    8 | [has]
    8 | [in]
    8 | [software]
    7 | [been]
    7 | [group]
    7 | [have]
    7 | [it]
    7 | [its]
    7 | [quadream]
    7 | [used]
    6 | [nso]
    5 | [administration]
    5 | [apple]
    5 | [company]
    5 | [iphone]
    5 | [is]
    5 | [marczak]
    5 | [us]
    5 | [use]
    5 | [was]
    5 | [with]
    4 | [be]
    4 | [hacking]
    4 | [journalists]
    4 | [microsoft]
    4 | [mr]
    4 | [new]
    4 | [or]
    4 | [they]
    4 | [tools]
    4 | [victims]
    3 | [according]
    3 | [activists]
    3 | [are]
    3 | [at]
    3 | [attacks]
    3 | [biden]
    3 | [but]
    3 | [candiru]
    3 | [companies]
    3 | [government]
    3 | [human]
    3 | [into]
    3 | [israel]
    3 | [israeli]
    3 | [known]
```

```
Length of the list: 41
Count | Word
-----   ---------------
    9 | [attack]
    9 | [citizen]
    9 | [lab]
    9 | [spyware]
    8 | [software]
    7 | [group]
    7 | [quadream]
    7 | [used]
    6 | [nso]
    5 | [administration]
    5 | [apple]
    5 | [company]
    5 | [iphone]
    5 | [marczak]
    5 | [us]
    5 | [use]
    4 | [hacking]
    4 | [journalists]
    4 | [microsoft]
    4 | [mr]
    4 | [new]
    4 | [tools]
    4 | [victims]
    3 | [according]
    3 | [activists]
    3 | [attacks]
    3 | [biden]
    3 | [candiru]
    3 | [companies]
    3 | [government]
    3 | [human]
    3 | [israel]
    3 | [israeli]
    3 | [known]
    3 | [linked]
    3 | [officials]
    3 | [politicians]
    3 | [research]
    3 | [rights]
    3 | [system]
    3 | [without]
```

```
    3 | [linked]
    3 | [officials]
    3 | [on]
    3 | [politicians]
    3 | [research]
    3 | [rights]
    3 | [system]
    3 | [which]
    3 | [without]
```

Prog9out.txt

```
attack | 9
citizen | 9
lab | 9
spyware | 9
software | 8
group | 7
quadream | 7
used | 7
nso | 6
administration | 5
apple | 5
company | 5
iphone | 5
marczak | 5
us | 5
use | 5
hacking | 4
journalists | 4
microsoft | 4
mr | 4
new | 4
tools | 4
victims | 4
according | 3
activists | 3
attacks | 3
biden | 3
candiru | 3
companies | 3
government | 3
human | 3
israel | 3
israeli | 3
known | 3
linked | 3
officials | 3
politicians | 3
research | 3
rights | 3
system | 3
without | 3
```