CSCI 3450
Fall 2022
Activity 02: Basic Text Processing
Assigned date: 09/01/2022
Due date: 09/07/2022 (11:59 PM, EST)
Total points: 40

Learning Goal: Getting used to Basic Python and Text Processing

(5 points) Preparation:

- Make sure you have Python 3.x installed on your machine.
- Go over the steps and examples from:
  http://www.nltk.org/book/ch01.html (section 1.2 and 1.4)

(35 points) Task:

1. Prepare a source code **activity02Lastname.py**. Lastname should be your last name. Implement steps (a~g) and save the outcomes accordingly.
2. Initially, test your code with the **shortScience.txt** file. Once the parsing is done properly, run the experiment on the **science.tx**t file. Submit the result you get for the **science.txt** file.

Steps: Write a module in python that accomplishes the following tasks.

a) Read a text file (e.g., science.txt or shortScience.txt)
b) Convert the text to lower case.
c) Tokenize the text (hint: use **nltk.word_tokenize(…)**)
d) Remove the punctuations (hint: **isalnum** or **string.punctuation** or **something else**)
e) Count the frequency of the tokenized words
f) Present/print the top 20 words (in decreasing order of frequency).
g) Draw/plot a bar diagram of word frequencies (you may want to install **NumPy** and **matplotlib**) for those 20 words.

Submission:

Run your code and save the produced results (sorted list and the graph) in a file named **activity02Lastname.pdf**. Lastname should be your last name.
Submit the source code and the pdf file.

** Write the code as clearly as possible.
** If necessary, add documentation.

Resource:
https://pythonspot.com/matplotlib-histogram/