Midterm Exam – Python & NLTK Programming

CSCI 426 Fall 2024

1. [18 points] Open the book "austen-emma.txt"

1). How many tokens and distinct words do you have after tokenization?

2). How many tokens and distinct words do you have after removing the stop-words? (use the NLTK stop-word list here)

3) How many times that the token "situation" appears?

4) Is there any common contexts that are shared by the token "judgment" and "situation"? If yes, list all of the common context you found.

5). Generate the frequency dictionary after the tokenization & stop-word removal, write the most common 50 tokens with their frequency to the output file “fdict1.txt”.

2. [18 points] Extract & parsing the text data from

" https://www.nyit.edu/news/articles/guiliano-global-fellows-glacier-saviors-exoplanets-and-more/"

1) Apply word tokenization to the file.

2) Normalize all tokens into lowercase.

3) Find out all tokens that end with "ing", and then write those tokens to the text file "words.txt"

Please submit your notebook with output results, and the output file “fdict1.txt” & "words.txt" to CANVAS.