CIS/STA 9665: Assignment 1

Applied Natural Language Processing

Due Date: 11:59 pm Oct 1

Guidelines:

- Use Python as a programming language and finish this assignment in Jupyter Notebook
- Work is to be done individually for this assignment
- Students handing in similar work will both receive a grade of 0 and will face disciplinary actions.

Chapter 1. Language Processing and Python

You should install NLTK, import NLTK and download the NLTK Book Collection at the beginning.

1.1 Find the collocations in text2 from NLTK Book Collection.

1.2 Review the discussion of conditionals in Section 4 in Chapter 1. Find all unique words in the Chat Corpus (text5) ending with the letter “l”. Show the first 20 words in alphabetical order.

1.3 What is the difference between the following two lines of codes? Which one will give a larger value and Why? Will this be the case for other texts?

```
sorted(set(w.lower() for w in text2))
sorted(w.lower() for w in set(text2))
```

1.4 Find all the Twelve-letter words in the Chat Corpus (text5). With the help of a frequency distribution (FreqDist), show these words in decreasing order of frequency.

1.5 Review the discussion of looping with conditions in Section 4 in Chapter 1. Use a combination of for and if statements to loop over the tokens of text2 and print all the distinct numbers, one per line.

Chapter 2. Accessing Text Corpora and Lexical Resources

2.1 Use Gutenberg Corpus Module to explore “shakespeare-caesar.txt”. Please answer the following questions:

a. How many word tokens does this book have?

b. How many word types?

c. What is the lexical richness?

2.2 Explore the Section "Hobbies" in Brown Corpus. Count the number of all "wh-" words (words starting with "wh"). Please make sure we are not double-counting words like “What” and “what”, which differ only in capitalization.

2.3 Explore Movie Reviews Corpus which contains 2k movie reviews with sentiment polarity classification (positive or negative reviews). Please find out the 20 most common words in the negative reviews and positive reviews separately. Please get rid of stop words, numbers and punctuations from reviews. Please make sure we are not double-counting words like “This” and “this”, which differ only in capitalization.
(Hint: from nltk.corpus import movie_reviews
from nltk.corpus import stopwords)

2.4 Explore Names Corpus, which initial letters are more frequent for males vs. females? (Hint: Plot of a Conditional Frequency Distribution will be useful for answering this question)

2.5 Use Gutenberg Corpus Module to explore “austen-persuasion.txt”. Write a function to find out the 20 most frequent occurring words of this text file that are not stopwords. Please get rid of numbers and punctuation. Please make sure we are not double-counting words like “This” and “this”, which differ only in capitalization.

2.6 Use one of the path similarity measures to score the similarity of each of the following pairs of words. Rank the pairs in order of decreasing similarity: car-automobile, journey-voyage, boy-lad, coast-shore, midday-noon, furnace-stove, food-fruit, bird-cock, bird-crane, tool-implement, journey-car, cemetery-woodland, food-rooster, coast-hill, forest-graveyard, shore-woodland, coast-forest, lad-wizard, chord-smile, glass-magician, noon-string.

What to Submit

a. Use Python as a programming language and finish this assignment in Jupyter Notebook
b. I have created an ipynb file with questions. Please add your code and answers in this ipynb file
c. After completion, please save your finalized ipynb file as a PDF file
d. Submit both **PDF file** and **ipynb file** to Blackboard
e. Please answers questions clearly, concisely, and completely. To answer some questions, the code is not sufficient. You should complement your answers in words by using **comments (#)**
f. The assignment will be graded on the correctness of the answers, comprehensiveness of the analysis, clarity of results' presentation and neatness of the report.