NLP

Portfolio Component: Text Analysis

Objective: Perform simple text analysis on a corpus of documents.
Turn in: Your Python .py files.

Instructions for Program 1:

- Download the inaugural address corpus found here: https://archive.org/details/Inaugural-Address-Corpus-1789-2009 (not the one in NLTK). Download inaugural.zip and expand.

- Define an Address Class with the following attributes and methods.
  - Attributes
    - year
    - president
    - original text
    - token length of original text
    - list of processed tokens
    - top 25 most common processed tokens
    - readability score
    - readability rating (string)
    - sentiment score
  - Methods
    - compute readability score (see below)
    - evaluate sentiment of each address using VADER (see this notebook: https://github.com/kjmazidi/NLP/blob/master/Part_2-Words/Chapter_07_WordNet/7.4_VADER.ipynb)
    - display method to print all attributes, but only first few lines of text

- Iterate through each file in the directory.
  - Read each file.
  - Preprocess the text to a list of tokens.
    - lower case
    - reduce to alpha tokens
    - remove stopwords
  - Create an Address Object for each file. Add to the Address dictionary (see below).
  - Invoke the readability method and store the readability measures as attributes
  - Invoke the sentiment method and store the sentiment as attributes
  - Create a dictionary with key = (president, year) and value = an Address object.
  - Pickle this dictionary and save it. Pickle examples in this notebook: https://github.com/kjmazidi/NLP/tree/master/Xtra_Python_Material

Instructions for Program 2:

1. Read in the pickled dictionary. Extract a set of President names from the first element in the key tuple.
2. In a forever loop, display a menu of your choice. Example:

```
Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
```

   a. If user selected 1, display the list of presidents
   b. If user selected 2, prompt the user for a president. Edit the user's input if they did not use Capital case for name. If there is no match, prompt again. Or see Extra credit below for another option. Display the keys for that president. Prompt the user for a year, or have a numbered menu. Once a match is found, invoke the display method on that address.
   c. If user selected 3, prompt the user for a two-word phrase. Return a collocation score and a statement whether or not the phrase is likely to be a collocation. Calculate pmi on the combined texts of all the addresses in a separate function.
   d. If the user selected 9, quit the program.

Extra credit: 10 points

If there is no match, try to match with a Levenshtein distance method. If you get a "fuzzy" match, ask the user if that's what they meant. Write a custom method using the algorithm here: https://en.wikipedia.org/wiki/Levenshtein_distance

Readability Scores

There are many readability formulas. Here is a simple one you can use, the Lasbarhetsindex (LIX), although you are free to use another.

readability = W / S + (L * 100)/W

W = The total number of words
S = The total number of sentences
L = The total number of long words (over 6 characters)

Use this scheme to set the readability score attribute:

<25 Children
25-30 Simple
30-40 Normal
40-50 Factual
50-60 Technical
>60 Difficult
Source: https://www.diva-portal.org/smash/get/diva2:721646/FULLTEXT01.pdf

Feel free to make your output and user interface friendlier than this. Your homework is graded only on the NLP aspects, not the user interface.

- Sample Run:

```
Welcome to the Inaugural Address Explorer!
Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
Please enter your selection: 2
Enter president name: Washington
[('Washington', '1793'), ('Washington', '1789')]
Please select a year: 1793

**********************************************************
Year: 1793
President: Washington
Length in words: 135
Readability =  61.16 which is level Difficult
Sentiment analysis: {'neg': 0.054, 'neu': 0.868, 'pos': 0.079,
'compound': 0.5719}

First few lines:
 Fellow citizens, I am again called upon by the voice of my country to
execute the functions of its Chief Magistrate. When the occasion
proper for it shall arrive, I shall endeavor to express the high sense
I entertain of this distinguished honor, and of the confidence which
has been reposed in me by the people of united America.

Previous to the execution of any official act of the President the
Constitution requires an oath of office. This oath I am now about to
take, and in your presence: That

Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
Please enter your selection: 9
```

Sample Run:

```
Welcome to the Inaugural Address Explorer!
Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
Please enter your selection: 3
Please enter a two-word phrase: chief magistrate
pmi= 2558.45
There is strong evidence that this is a collocation

Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
Please enter your selection: 3
Please enter a two-word phrase: random phrase
pmi= 0.00
This is not a collocation

Main menu:
1. See a list of Presidents
2. Look up addresses by President
3. Look for collocations in the inaugural corpus
9. Quit
Please enter your selection: 9
```