Assignment 4 — Natural Language Processing with NLTK

HLTH 453 / 619 – Fall 2025

In this assignment, you will create a function that applies natural language processing to a Wikipedia page. Your function should work for any search terms, but for the purposes of testing, you should at a minimum make sure that your function works for the Health informatics page — it will be the first test on Marmoset.

1. Using the `wikipedia` library in Python, extract the page's text, contained in the `content` attribute of the `WikipediaPage` object.
2. NLP Steps:
   (a) Tokenize the text using `word_tokenize()` from NLTK.
   (b) Use a regular expression to exclude terms that have numbers and/or symbols. In other words, all included tokens should consist only of letters from the alphabet.
   (c) Make all included tokens lowercase.
   (d) Use the NLTK package's `stopwords` list to remove common words
   (e) Use the Porter stemmer from NLTK to normalize each token.
   (f) Note that to successfully use the NLTK package, you may need to install some of its sub-modules. See the code snippet below for examples of how this works.
3. Create a `DataFrame` object to hold your results
   (a) Fill it with the 30 most frequent terms in *descending* order of frequency. In cases where there are words with the same frequency, *sort them in alphabetical order*.
   (b) The index of the DataFrame should be terms. The name of the index should be "Term".
   (c) The DataFrame has only one column which contains frequencies. The name of the column should be "Frequency".
   (d) Recall that you've solved this part of the problem twice before on Assignments 1 and 2, you may re-use part or all of your code from those assignments
4. Your function should end by `return`ing the DataFrame object

Your Python solution should be contained in a file named `assignment4.py`. You may use the following function definition as a starting point:

```
1
2  import wikipedia
3
4  import nltk
5  nltk.download('punkt')
6  nltk.download('stopwords')
7
8  from pandas import Series, DataFrame
9
10 def getFrequencies(wiki_page): # Returns a DataFrame object with the 50 most frequent terms
11
12 # Create an empty dictionary object
13
14 # Get the Wikiepdia Object
15
16 # Letters only, convert to lower case
17
18 # Remove Stopwords from the Dict
19
20 # Use the porter stemmer
21
22 # Convert to a frequency table
23
24 # Convert to a DataFrame
25
26
27 return df[0:30]
28
29
30
```

1 Resources

You may find the following resources helpful in completing this assignment:

1. Documentation for the wikipedia module
2. The NLTK book provides some useful examples for natural language processing in Python
3. Python's documentation for regular expressions, in the `re` module
4. See Chapter 5 of the textbook regarding pandas DataFrame manipulations.

2 Grading

This assignment is worth 10% of your final grade in the course.

This assignment will be automatically graded on Marmoset by comparing its output to a solution's for 5 different Wikipedia pages, each worth 2%. The first test will be for the Health informatics page, the other tests will be against a variety of different pages. So you should make sure to test your method against a variety of Wikipedia pages before submitting your final solution. Submitted code must be error-free. If your code results in errors and does not execute to completion, a grade of zero will be given.