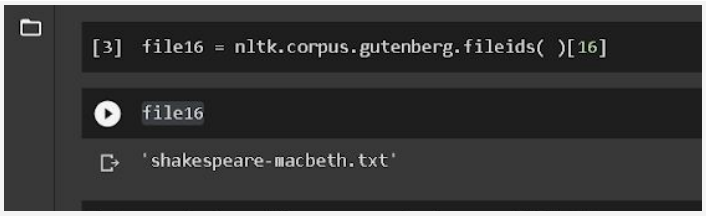
Name:Puja Bathija
SUID:692880674

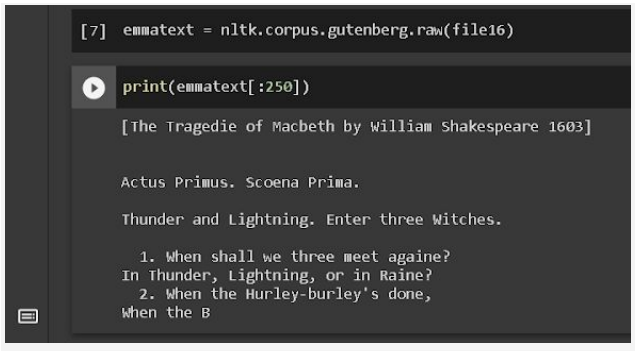
Title: Natural Language Processing lab installing nltk

Book title: Macbeth by Shakespeare

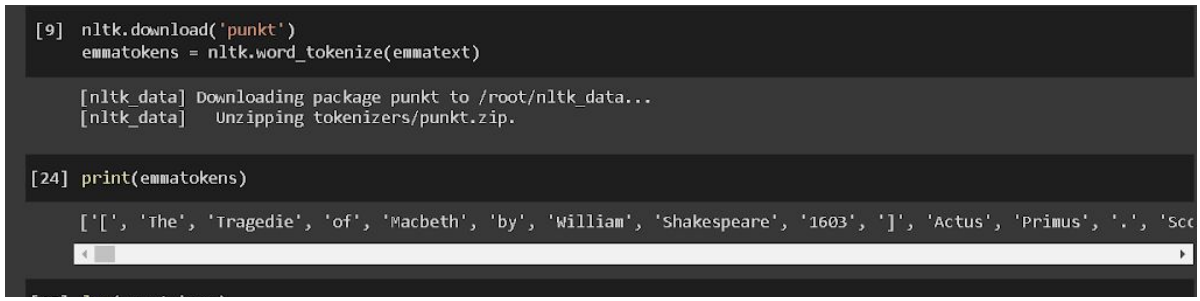


Steps :

1. Get the text with nltk.corpus.gutenberg.raw()
   raw() function is used to convert the information into a string[1].



2. Get the tokens with nltk.word_tokenize()
   This is the next step after getting the string to divide it into words this is done using the word_tokenize() function[2].



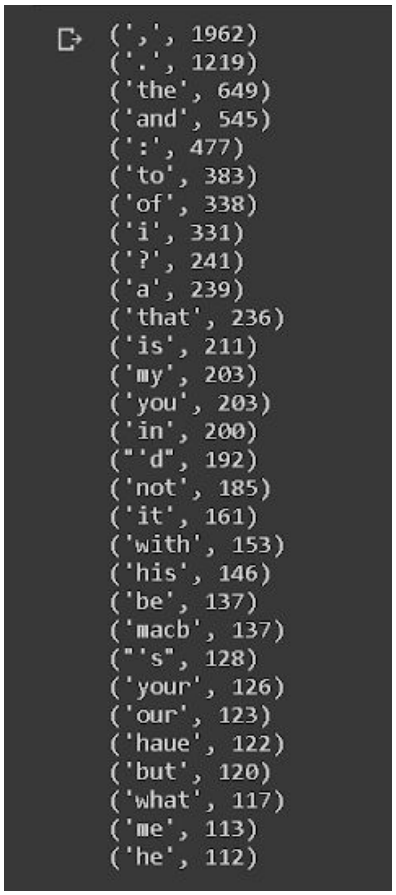
3. Get the words by using w.lower() to lowercase the tokens

   The words are converted into the lower case as a measure to prevent the same words but in a different form to be identified as different words. For example, Apple and apple would be identified as different words but once all words are converted into the lower case the text would be uniform[1].

4. Make the frequency distribution with FreqDist
   The FreqDist is a function in nltk library that takes the words and returns words and their frequency[3].

5. Get the 30 top frequency words with most_common(30) and print the word, frequency pairs.
   most_common() function returns the most frequent words and their count[3].

List of the top 30 frequent words



Observations:

Most of the top frequent words are either punctuations, pronounces, prepositions, and other filler words. These words have little to no meaning when it comes to analysis. To overcome this problem nltk has a set of stop words that are filtered out to get the words that have real meaning. Since in the experiment, these words weren't filtered out and are present in the result [4].

Lessons learned:

1. To convert the data in a string using raw() function.
2. To break a string into words using word_tokenize() function
3. To make a frequency distribution of the words using FreqDist and checking the most common words in the distribution using most_common() function.

References:

1. Accessing text corpora and Lexical Resources. (n.d.). Retrieved February 14, 2021, from https://www.nltk.org/book/ch02.html
2. Kabitakumar. (2020, July 18). How to perform nltk on raw text. Retrieved February 14, 2021, from https://medium.com/analytics-vidhya/how-to-perform-nltk-on-raw-text-5dc5dbe99f0
3. Python nltk.freqdist() examples. (n.d.). Retrieved February 14, 2021, from https://www.programcreek.com/python/example/16275/nltk.FreqDist
4. Removing stop words with NLTK in Python. (2020, November 24). Retrieved February 14, 2021, from https://www.geeksforgeeks.org/removing-stop-words-nltk-python/