**PDS ASSIGNMENT 3**

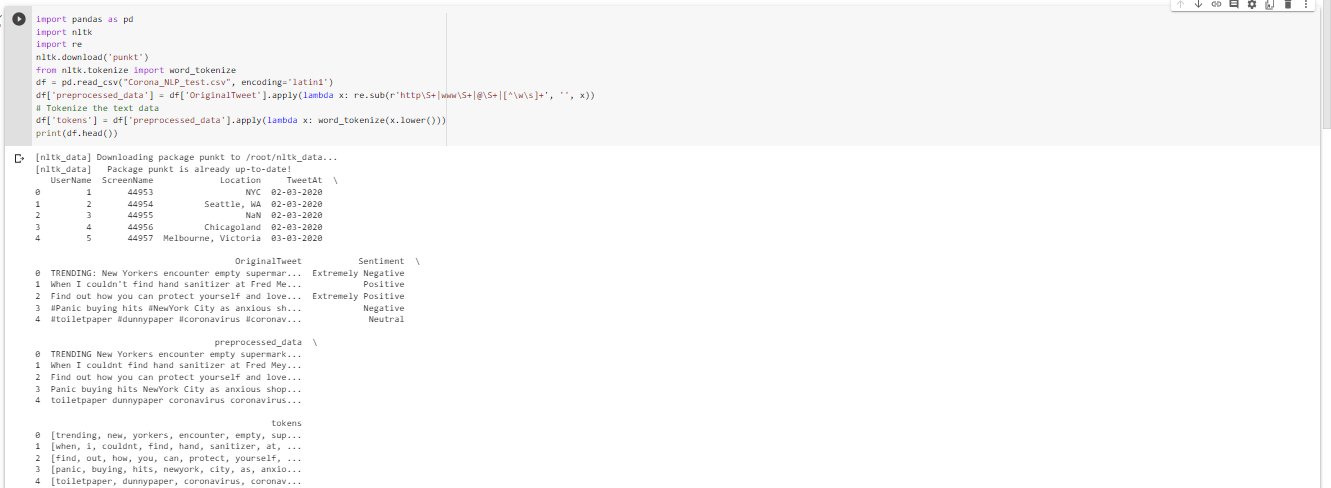
**NAMBURI SWETHA**

1) (20 points) The data file contains tweets that have been pulled from Twitter. In this dataset  
use the text data in the “OriginalTweet” column and perform the following:  
a) Convert the text corpus into tokens.  
b) Perform stop word removal.  
c) Count Word frequencies  
d) Create word clouds.

The goal of this report is to analyse a dataset of tweets pulled from Twitter. In particular, I carried out the following operations on the text data in the Original Tweet column.

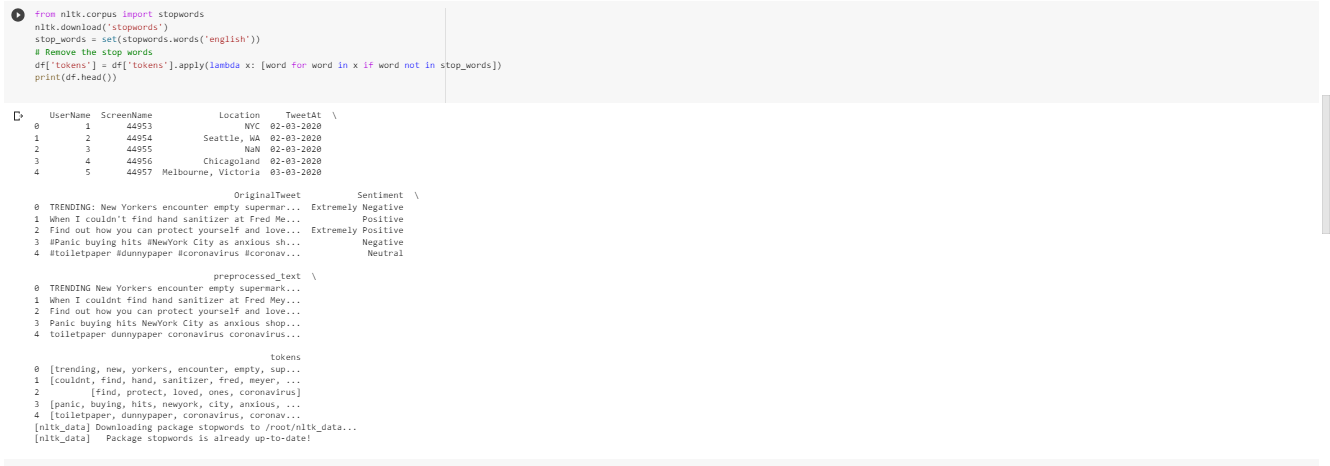
1) Convert the text corpus into tokens.

Tokenizing text data is the initial step, which entails dividing the text into individual words or tokens. With Python's nltk package, we can accomplish this. We first import the data into a panda’s data frame, after which we pre-process the dataset tokenize each tweet in the preprocessesed\_data column using the word\_tokenize function. A list of tokens for each tweet is then created in a new column called "tokens" as a result.



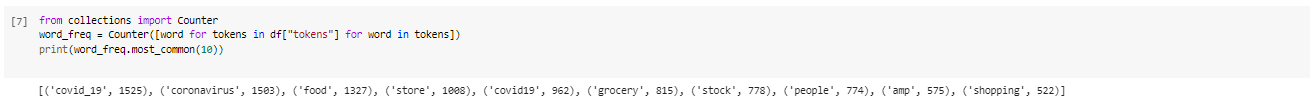
2) Perform stop word removal.

Stop words, which are common terms with little significance, can be eliminated from text data to increase performance and lower noise. With the help of the Python nltk module, stop words can be eliminated. We first import the stop words from the nltk corpus before using a lambda function to delete them from each tweet in the "tokens" column.



3) Count Word frequencies

We can now use the Python collections package to count the frequency of each word after tokenizing the text input and removing the stop words. The list of tokens is first flattened using a list comprehension, and then the Counter function is used to tally the frequency of each word.



4) Create word clouds.

To illustrate the most prevalent words in the tweet data, we can generate word clouds. For this, we can use the Python wordcloud package.

