**ASSIGNMENT 3**

**STUDENT ID: 16326673**

**NAME: ANJANI PRASAD GAJULA**

1. **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**

1. **Convert the text corpus into tokens.**

**Text

Description automatically generated**

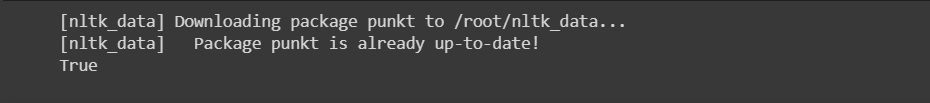
The code snippet downloads and updates the "punkt" package in the Natural Language Toolkit (nltk) library.

The "punkt" package contains data required for tokenization, which is the process of splitting text into smaller units called tokens.

The dataset contains 3798 tweets related to the COVID-19 pandemic.

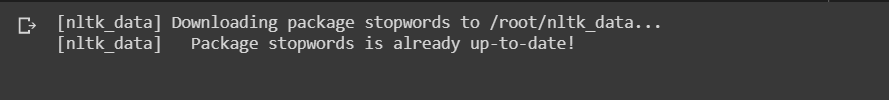
Each tweet has been tokenized into a list of individual words using the "punkt" package.

The tweets cover various aspects of the pandemic, including panic buying, protective measures, and the impact on daily life**.**

****

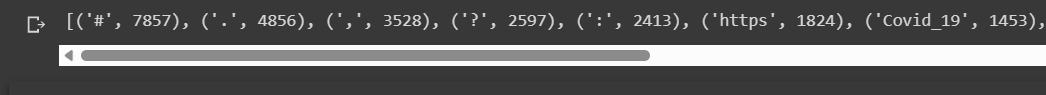
This code updates the "punkt" package in the nltk library, which is used for tokenizing text data into individual words or sentences. The output confirms that the package is up-to-date.

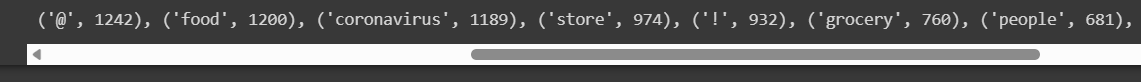
b. **Perform stop word removal.**

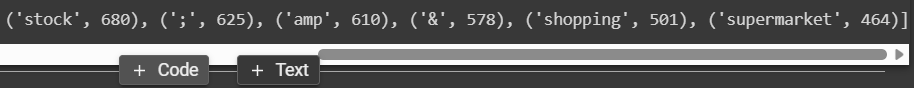


This code updates the "stopwords" package in the nltk library, which contains a list of common words often removed from text data during natural language processing. The package helps reduce noise in text data and improve analysis accuracy. The code output confirms the package is already up-to-date.

**c. Count Word frequencies**







The given data shows the frequency counts of tokens in a text corpus. The most frequent tokens are symbols, followed by words related to COVID-19 and grocery shopping. The presence of these words suggests that the corpus is related to the COVID-19 pandemic and grocery shopping.

**d. Create word clouds.**

A picture containing text

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence