**Principles of Data Science**

**Name: Srujan Nagothu**

**ID: 16341576**

**Assignment 3**

**Answers:**

The raw dataset is loaded initially and performed the preprocessing steps on the tweets data to transform the data into analysis ready form.

**Loading the raw data:**

Graphical user interface, text, application, email

Description automatically generated

**Preprocessing the data:**

Graphical user interface, text, application

Description automatically generated

1. **Converting the entire text corpus into tokens.**

**Loading the cleaned preprocessed data:**

Graphical user interface, text, application, email

Description automatically generated

**Converting the text corpus into tokens:**

**Graphical user interface, text, application

Description automatically generated**

**Background pattern

Description automatically generated with medium confidence**

1. **Removing all stop words in the data**

Removing all stop words from the tokenized vectors

**Graphical user interface, text, application

Description automatically generated**

**A picture containing text

Description automatically generated**

1. **Counting the words frequencies**

Frequency of occurrence of filtered tokens in the corpus:

Graphical user interface, text, application, chat or text message

Description automatically generated

Most frequently occurring words in the corpus:

**Graphical user interface, text

Description automatically generated**

Representing the top 30 frequent words over the plot:

**Graphical user interface, text, application

Description automatically generated**

**Chart, histogram

Description automatically generated**

1. **Creating the word cloud for visual representation of frequency of tokens.**

Generating a word cloud using the filtered tokenized vector of the text corpus:

**A picture containing text

Description automatically generated**

**Word Cloud:**

**Text

Description automatically generated**