Challenge 7

Sentiment analysis, also known as opinion mining, refers to the use of text analysis, computational linguistics and natural language processing to systematically identify and extract subjective information. Association rule mining is a powerful method which can help in extracting such useful information and can thus contribute to sentiment analysis.

In this challenge, you are provided with a dataset of 95,488 different tweets, collected from April to August 2020, which was the peak lockdown period in India due to COVID-19.

Prepare a report and submit a PDF file by Tuesday (**11/24/2020 before 6 pm**) considering the below details:

1. Read all tweets line by line, and simultaneously, create a dictionary of unique words by splitting each tweet into words based on spaces.
    a. Discard all punctuation marks, if any, from the end of tweets.
    b. Convert all letters to lowercase when you're creating the dictionary, i.e., Lockdown would be the same as LockDown or lockdown.
2. Once all unique words are extracted, read the tweets dataset once again. At this time, create an output file (name it as "new_dataset.txt") with exactly the same lines of code and tweets as the original dataset but replace words with their corresponding numbers.

    As an example, if the original dataset had only 3 tweets as below:

    I hate staying at home.
    @lockdown should be over right now.
    I hate @Lockdown

    The unique words dictionary would be as follows:

    Words_Dic = {I: 1, hate: 2, staying: 3, at: 4, home: 5, @lockdown: 6, should: 7, be: 8, over: 9, right: 10, now: 11}

    Then, the new file syntax would be as below:

1 2 3 4 5
6 7 8 9 10 11
1 2 6

3. Apply the Apriori itemset mining algorithm to the created new dataset (i.e., the “new_data.txt”) and report 5 itemsets that have a support value of at least 30%.
4. Apply the FPGrowth algorithm to the same dataset and report 10 association rules with a minimum support of at least 50% and a minimum lift of 1.