

Assessment 1

Tweets Cleanup and Analysis Using Regular Expressions

Objective: Use regular expressions to work with messy tweets data: clean up the data, extract hashtags, analyze the most popular hashtags that occur along with a target hashtag (#economy).

Problem Statement: Social media is a gold mine of information. Brands, governments, or anyone can leverage their business with the help of the information contained. It can be information on the sentiments for a brand, or the themes being spoken about, or the associated trends for a particular hashtag. In this project, we will work on the tweets on Twitter.

We will find other hashtags that occur frequently with our target hashtag. This will give us an understanding of which other topics people are associating this hashtag with.

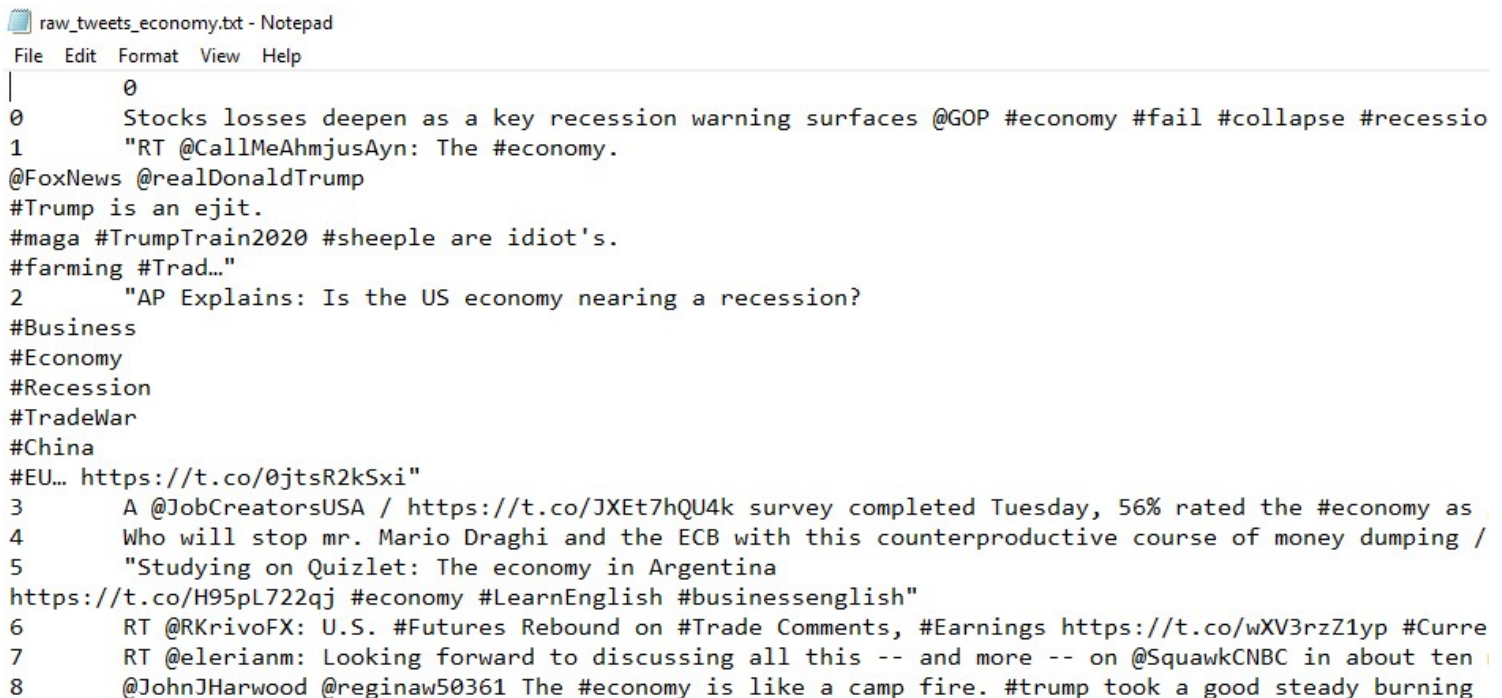
Domain: Social media

Analysis to be done: Cleanup tweets and analyze top hashtags

Content:

Dataset: 'raw_tweets_economy.txt'

Dataset has no header. For each row, it has an index and then the tweet text.



```
raw_tweets_economy.txt - Notepad
File Edit Format View Help
|
0      0
0      Stocks losses deepen as a key recession warning surfaces @GOP #economy #fail #collapse #recessio
1      "RT @CallMeAhmjusAyn: The #economy.
@FoxNews @realDonaldTrump
#Trump is an ejit.
#maga #TrumpTrain2020 #sheeple are idiot's.
#farming #Trad..."
2      "AP Explains: Is the US economy nearing a recession?
#Business
#Economy
#Recession
#TradeWar
#China
#EU... https://t.co/0jtsR2kSxi"
3      A @JobCreatorsUSA / https://t.co/JXEt7hQU4k survey completed Tuesday, 56% rated the #economy as
4      Who will stop mr. Mario Draghi and the ECB with this counterproductive course of money dumping /
5      "Studying on Quizlet: The economy in Argentina
https://t.co/H95pl722qj #economy #LearnEnglish #businessenglish"
6      RT @RKrivoFX: U.S. #Futures Rebound on #Trade Comments, #Earnings https://t.co/wXV3rzZ1yp #Curre
7      RT @elerianm: Looking forward to discussing all this -- and more -- on @SquawkCNBC in about ten
8      @JohnJHarwood @reginaw50361 The #economy is like a camp fire. #trump took a good steady burning
```

Steps to perform:

Tweets data is generally very ill-formatted and contains URLs, user handles, retweet markers, etc. The data is not clean and it is difficult to extract the desired information. We will use regular expressions to clean up the tweets and then, use it to extract hashtags from the data. We will eliminate the target hashtag #economy, which is effectively a contextual stop word in this case.

Tasks:

1. Load the tweets file using read_csv function from Pandas package. (Hint: provide the appropriate separator)
2. Drop the column 'Unnamed: 0' and rename the column containing the text to 'tweet'.
3. Get the tweets into a list for easy text cleanup and manipulation.
4. Normalize the case by converting all text into lower case and assigning new variables. See first five tweets to confirm if we got the desired result.
5. Using regular expressions, remove user handles. These begin with '@'.
 - a. First, try removing user handle from the test string '@Rahim this course rocks! <https://linkedin.com/in/rahim-baig>'.
 - b. Then, cleanup all the tweets once the function or pattern is decided. Check first five items to confirm we got the desired result.
6. Using regular expressions, remove URLs:
 - a. First, try removing URLs from the test string '@Rahim this course rocks! <https://linkedin.com/in/rahim-baig>'.
 - b. Then, clean up all the tweets once the function or pattern is decided. Check first five items to confirm we got the desired result.
7. Using regular expressions, extract only the hashtags from the tweets. Remember, we have to analyze the most common hashtags in tweets.
 - a. Extract hashtags from the sample string, '@Ram, #food is #love'.
 - b. Then, extract from all the tweets once the pattern is decided. Check the first five records to confirm the result.
8. Counting the most common hashtags:
 - a. First, collate all the tags into one single list to conveniently pass on to a counter.
 - b. Remove the **contextual stop word** i.e. '#economy'. This will be dominant, as this is the target hashtag. We need to analyze the other hashtags.
 - c. Use a counter to count the most common hashtags in the data.
 - d. Get top 10 hashtags.
9. Plot the top 10 hashtags in a horizontal bar chart.
10. Feature Engineering: Use Count Vectorizer to represent each document.
 - a. Instantiate Count vectorizer with 3000 vocabulary size
 - b. The vectorizer needs the strings, not vectors. Join the tokens into the string for each article
 - c. Apply Count vectorization on the articles and determine the shape of the matrix.