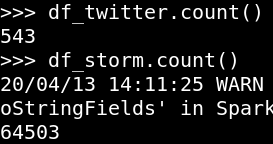
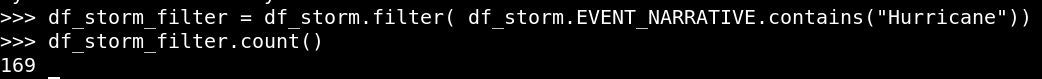
For this Assignment, I examined Hurricane Sandy occasions, which created on October 22, 2012 from a tropical wave and got one of the strongest and most ruining typhoons of that season. In the US it influenced 24 states, from Florida to Maine and over the Appalachian Mountains to Michigan and Wisconsin, with especially causing extreme harm in New Jersey and New York. (according to the portrayal in the Overview).

The following section outlines the steps that I took on the VM to process and analyze the data:

Creating two dataframes for Storm and Twitter data:



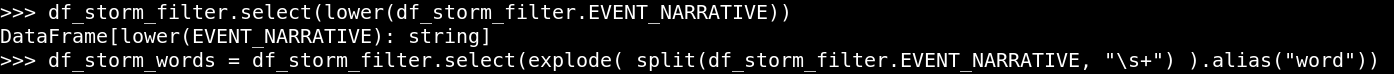
Filtering the storm events that contain the word hurricane in them:

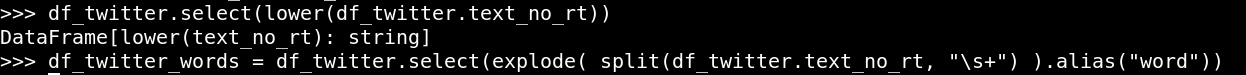


Converting all the alphabets of the strings to lower cases:



Splitting the words to create token and saving the words in a separate column using explode()

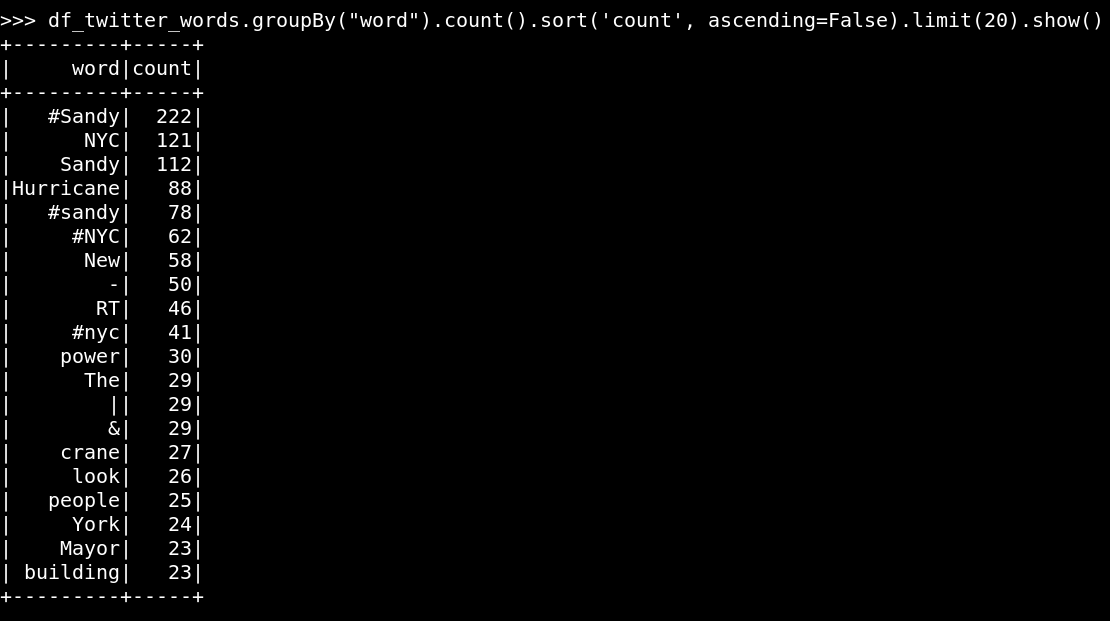




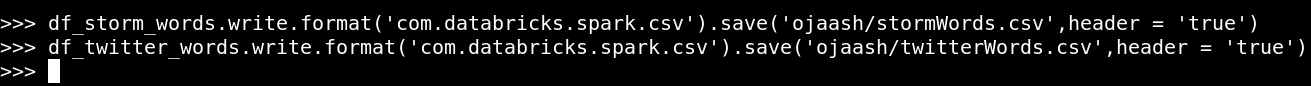
Adding the stopwords and to be removed:



Finding insights into the data to find the top 10 occurring words for Storm and Twitter Data:



Saving the data in a CSV in order to be used in the python notebook attached alongside:



Analysis from the Python Notebook:

I attempted to look at the contrasts among government and twitter stories based on the following parameters stated below:

1. Emotional versus Factual words:

By taking a look at the initial 10 words, we can say that the government information has progressively factual words like 'extreme', 'harm' while since twitter information has open accounts, it has increasingly sentimental words like 'power', 'individuals', 'look', and so on. To get better insights on what sort of words are utilized in both, I attempted POS tagging using nltk library.

2. Parts of Speech labeling:

Since the information size of both the information outlines is unique, I determined the level of things utilized in every one of them, by taking the percentage metrics of the considered criterion. Prior to this step, I had an intuition that the government information ought to contain more things while twitter information ought to have more descriptive words. Government information contains the news and realities given by the legislature. The consequences of the POS labeling confirmed my examination. As per the calculation, twitter information contains 49.64% things and government information contains 65.72% things. To get more information, I performed further assessment via textblob library.

3. Sentiment Analysis:

I determined the assumption for each word and afterward took its mean to get last conclusion score. The examination indicated that the mean opinion score of government information is more than that of twitter information. This implies the legislature revealed a progressively positive review of the circumstance. As indicated by individuals, more annihilation occurred than revealed by the administration.

Every one of these investigations at long last infer that the government information introduced a progressively positive perspective on the circumstance. As per the tweets by individuals, the damage was more than what was expressed by the legislature. Additionally, the tweets demonstrated more sentimental words in the information.