Objective

Objective of this task is to pull the days trending twitter and categorizing based on the sentimental attributes.

Limitation

- The *Days Trending twittesr* are not possible to pull using the Twitter API. As per Tweepy (other API) API, trending hash tags can be filtered only based on geolocation. Please find the more information here

  Resolution: As per Tweepy API, I am pulling the last 7 days trending topic instead of one day.

Feature Engineering

Apart from the normal data cleansing, Sentence formation from Hash tags is an key module of the code. Because this is where Hash tags are converting to meaningful words and used in classification.

As part of Sentence formation each hash tags compared with dictionary words and picking the available words.

- There are some circumstance same hashtag may give more than meaning correct dictionary words. But this algorithm will take the large text which is covering all text in the sentence.
- Also some unidentifiable words present in the hashtag. So it is is not possible to infer anything from that. So I architecture as removing the unidentifiable words.

Due to these word anomalies, hash tag may refer to the wrong context. In order to overcome this, either we need to focus on the particular domain or classify the each hash tags to domain & find the respective domain dictionaries. However this algorithm works well for most of the common words.

Naive Bayes algorithm for sentiment analysis

We have used the Naïve Bayes algorithm for our problem. This algorithm is based on the **Bayes theorem**. As part of Bayes theorem, it will predict based on the past events. Also each experiment is independent. So in our case each hash tags are independent one. So it will give the better result for us.

Accuracy

The current model is providing the 60 % Accuracy in test set (Hashtags). Since most of the hash tags contains a objects like name and place, we could not differentiate that is dictionary word or object name. Due to this some hash tags are misled to wrong context. It is affected accuracy.

Furthermore, if we choose any one specific domain trending hashtags (E.g. Games) and train the relevant positive, negative & neutral data, it will provide the good accuracy.

Project Flow Diagram