Rohit, VARUN

Project Milestone – Twitter Word Embeddings

**Below milestones have been achieved:**

**Data Cleaning:** OriginalData is structured in json format. Following steps are taken in process of cleaning the data and format in a data frame structure:

1. Flattening of Json Data: We flattened the json data for each key object.
2. Filtering of English Tweets: Filtered the data based on language indicator as ‘en’.
3. Extraction of all the hashtags in a tweet and tweet text.
4. Removing all hyperlinks, special characters, retweets.
5. We have used a custom Twitter stop words list to remove stop words from our corpus.
6. Stemming and Lemmatization have been implemented from nltk library to improve our results.

Below is an image which shows the text after cleaning process.



**Features Extraction**: The final cleaned dataset included 700,000 tweets with two columns. One column consists of hashtags and the other consists of the tweets text.  
Features are extracted from the corpus consisting of tweets text. An approximately 10,000 words dictionary is considered for our model. Each of these words is converted into a 32 length vector.

**Modeling Approaches:**

**Approach 1**: Combining hashtag and Tweet texts(only once)

In this approach our corpus consists of combination of hashtags and the tweets, where the hashtags are present only once in a tweet. We implemented our word2vec model in this corpus.

Example: Original data



Modified corpus:



**Approach 2**: Adding hashtag after each word in the tweet text.

In this approach we are adding our hashtag after each word in the tweet so that our hashtag will be in the context window size of each word in tweet text.

Example: Original Data



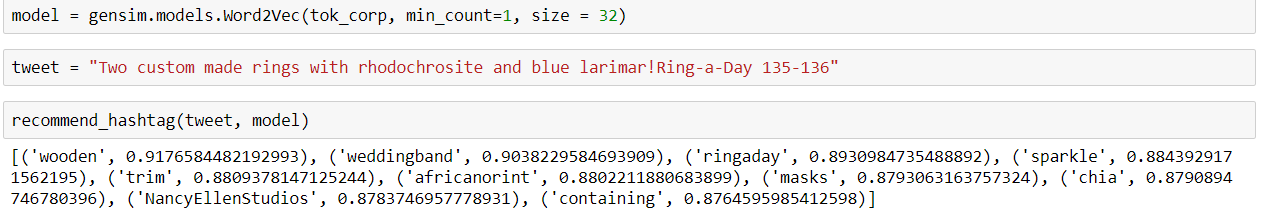
Modified Corpus



**Machine Learning Algorithm(Word2vec) and Results**

* We are able to predict the hashtag using CBOW method for any tweet text provided as input
* We have used Gensim library implementation of word2vec model with multiple context window size.
* Each word vector has a size of 32 words. The minimum count is taken as 1 which means takes all those words into consideration which appears at least once.
* In order to make use of current word2vec API, we are embedding the hashtag text in between every tweet word in order to maintain the context and target structure for word2vec model.
* Taking into account that our final aim is not to recommend original hashtag for training set (which would be a case of overfitting) but to predict relevant hashtags for the new tweet text, current results are fulfilling the objective

Below is a snippet from our Output for one of the tweets.



From the above results, we are able to predict the words/hashtags which is most suited to the tweet with the probability.

* Further improvements planned:
  + Instead of embedding hashtag in between all the tweet text tokens, we will be trying to “randomly sprinkle” the hashtag in between some of the tokens in order to avoid overfitting
  + Another plan to improve the model is to implement stemming and lemmatization for the tweet tokens.
  + Use of separate vocabulary subset for prediction of hashtag. We will use this list to make recommendation using cosine similarity so as to avoid recommendation of non-hashtag token as hashtag from tweets.
  + We need to create a evaluation criteria for our tweets hashtags prediction accuracy.