Project 3

Twitter API

Search API

To get the tweets related to a search term, you are supposed to send a GET request to https://api.twitter.com/1.1/search/tweets.json . (This is what happens generally, but we have done this for you.. Run the following command to mimic the process). To learn more about this, visit https://dev.twitter.com/docs/api/1.1/get/search/tweets

Run the following to get the tweets related to a term of your choice

```
$ python3 fetch_tweets.py -c fetch_by_terms -term "[your_chosen_term]" > search_output.txt
```

Some examples of search term are presidential elections, Game of Thrones, etc.

When working with the text data, most of the applications demand the usage of building feature vectors from the text documents. So, in this assignment, we will be using the following three methods for building feature vectors.

1. TFIDF Vectorizer:
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer

2. Count Vectorizer:
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.feature_extraction.text.CountVectorizer

3. Hashing Vectorizer:
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.HashingVectorizer.html#sklearn.feature_extraction.text.HashingVectorizer

Query:

You will use the tweets obtained from Twitter data to answer the following queries:

1. Given a query(some search topic of your choice), return top 10 similar tweets to the given query.
2. Return 5 clusters of similar tweets(Here, a cluster is a set of tweets that are similar).

FAQ:

1. Should the Data be cleaned?

A: Yes. Data Cleaning is a necessary step in the pipeline(Tokenization, stop word removal, etc).

2. There are lots of hyper parameters in each of the Vectorizers, which ones are relevant to the task at hand?

A: min_df, max_df are some parameters of interest. You are welcome to experiment other parameters.

Bonus Points :

Generally, unigrams are used as features. Experiment with other types of features like bigrams, trigrams, Part of Speech tags, combinations of these. Document whatever you have tried.