Assignment-2

Quesion 1: Download the real-time twitter stream using Python library called Tweepy to connect to Twitter API. You can write your own script or can use provided script (twitter-to-mongo.py) in the assignment folder. Please install the following libraries to run the script: (if not already installed)
(1) $ pip install tweepy
(2) $ pip install pymongo

To start with, you will need to have a Twitter developer account and obtain credentials (i.e. API key, API secret, Access token and Access token secret) on the to access the Twitter API, following these steps:

- Create a Twitter developer account if you do not already have one from : https://developer.twitter.com/
- Go to https://developer.twitter.com/en/apps and log in with your Twitter user account.
- Click "Create an app"
- Fill out the form, and click "Create"
- A pop up window will appear for reviewing Developer Terms. Click the "Create" button again.
- In the next page, click on "Keys and Access Tokens" tab, and copy your "API key" and "API secret" from the `Consumer API keys` section.
- Scroll down to `Access token & access token secret` section and click "Create". Then copy your "Access token" and "Access token secret".

Fill these credentials into script file and run the script:

```
$ python twitter-to-mongo.py
```

This script will save the downloaded tweets into "december4" collection under the "TwitterStream database".

If you are unable to download tweets, you can directly import the december4.json file provided in assignment folder into your TwitterStream database by using this command:

```
 $ sudo mongoimport --db TwitterStream --collection december4 --file
december4.json
```

To see the imported database, go to mongo terminal and type:

```
> show dbs
> use TwitterStream
> show collections
```

Question2: To analyze the Twitter data, please run the following queries using python.

```
$ pip install wordcloud
```

Required libraries:

```
import json
from pprint import pprint
import nltk
import numpy as np
import re
from datetime import datetime
from collections import OrderedDict
import pandas as pd
from os import path
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import matplotlib.pyplot as plt
import codecs
# For connecting MongoDB
from pymongo import MongoClient

client = MongoClient()
db = client['TwitterStream']
```

1. Show the collections in the TwitterStream database.
2. Count the total number of tweets in the particular collection(december4)
3. Print the top one tweet in structured format.
4. Print the fields of tweet.
5. Print the text of top 10 tweets with posting time.
6. Print the most popular 50 hashtags and plot the scattar, bar and wordcloud diagram to represent most popular hashtags.
7. Print the top 50 retweeted tweets.
8. Print the top 10 users based on user frequency.