W205 - Data Storage
Exercise 2
Thong Bui

Description

This exercise is to familiarize yourself to elements of a streaming application, specifically application that analyzes the Twitter data using Apache Storm to parse data and count words streamed from Twitter and store the words counted in Postgres DB.

Technologies used here:

- Amazon EC2: to build the environment to run this exercise
- Twitter API: to access Twitter stream data
- Apache Storm: to parse data streamed from Twitter
- Python: the main programming language used for this exercise
- Streamparse: to create python Storm projects
- Postgres DB: used to store word counted from Twitter live stream data
- Psycopg: PostgreSQL adapter for python to connect to Postgres DB

Directory and file structure

Once you check out github, you can see that exercise_2 is the main directory containing all the files and subdirectories for this exercise:

1. *Twittercredentials.py*: this is the one contains all the Twitter credentials to connect to Twitter API used by hello-stream-twitter.py below
2. *hello-stream-twitter.py* : this is the application that print all the tweets in the stream containing “Hello” in a 1-min period. At the end it also count how many tweets it receives during this 1 min
3. ***tweetwordcount***: is the directory that contains all files and subdirectories created for this Storm application

    a. *topologies/tweetwordcount.clj*: implements the tweet word-count topology
    b. *src/spouts/tweets.py*: spout code that streams data from Twitter API and filter for English tweets only
    c. *tweetwordcount/src/bolts/parse.py*: first bolt that parses for valid words from the tweets receives from the spout (b)
    d. *tweetwordcount/src/bolts/wordcount.py*: second bolt that reads the words from the 1st bolt (c), count them and store the word count data in the local Postgres DB
    e. *tweetwordcount/README.md*: instructions of how to run this application
4. *finalresults.py*: this application reads data from Postgres to either returns the total number of word occurrences or all the words and their occurrences

Ex1:

```
$ python finalresults.py hello
$ Total number of occurences of “hello”: 10
```

Ex2:

```
$ python finalresults.py
$ (<word1>, 2), (<word2>, 8), (<word3>, 6), (<word4>, 1), ...
```

5. *histogram.py*: given 2 integers, this script will return all the words with occurences in this range
6. *screenshots* : this directory contains all the screenshots as the results of this exercise run
7. *Plot.png*: This is the bar chart showing the top 20 words in my twitter stream

Dependencies

tweetwordcount/README.md mentions the tweetwordcount Storm app’s dependency on the running of local Postgres DB. It must have:

- Database tcount created
- Table tweetwordcount created in tcount database