Tweet Word Count Application

Summary

This application uses a Tweet API to capture tweet live data, process them in real time and store the aggregated results in a database. Users can execute psql query against the database to extract insights.

Application Topology

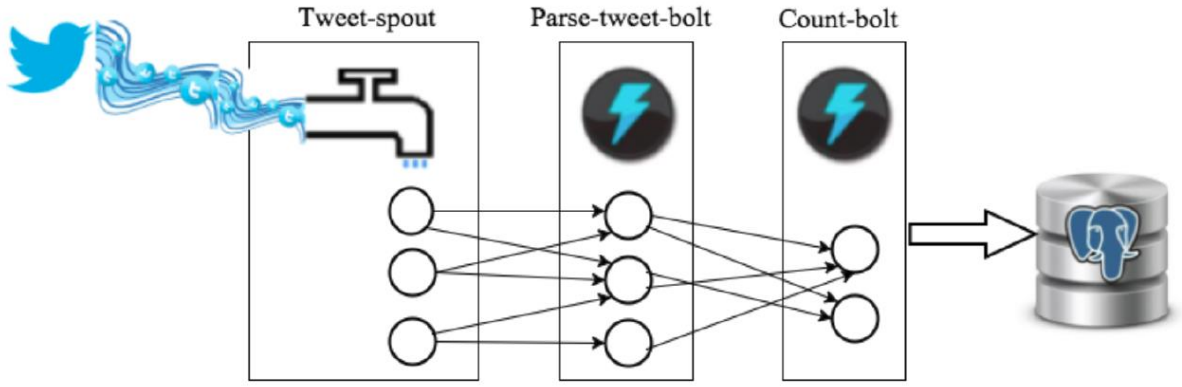
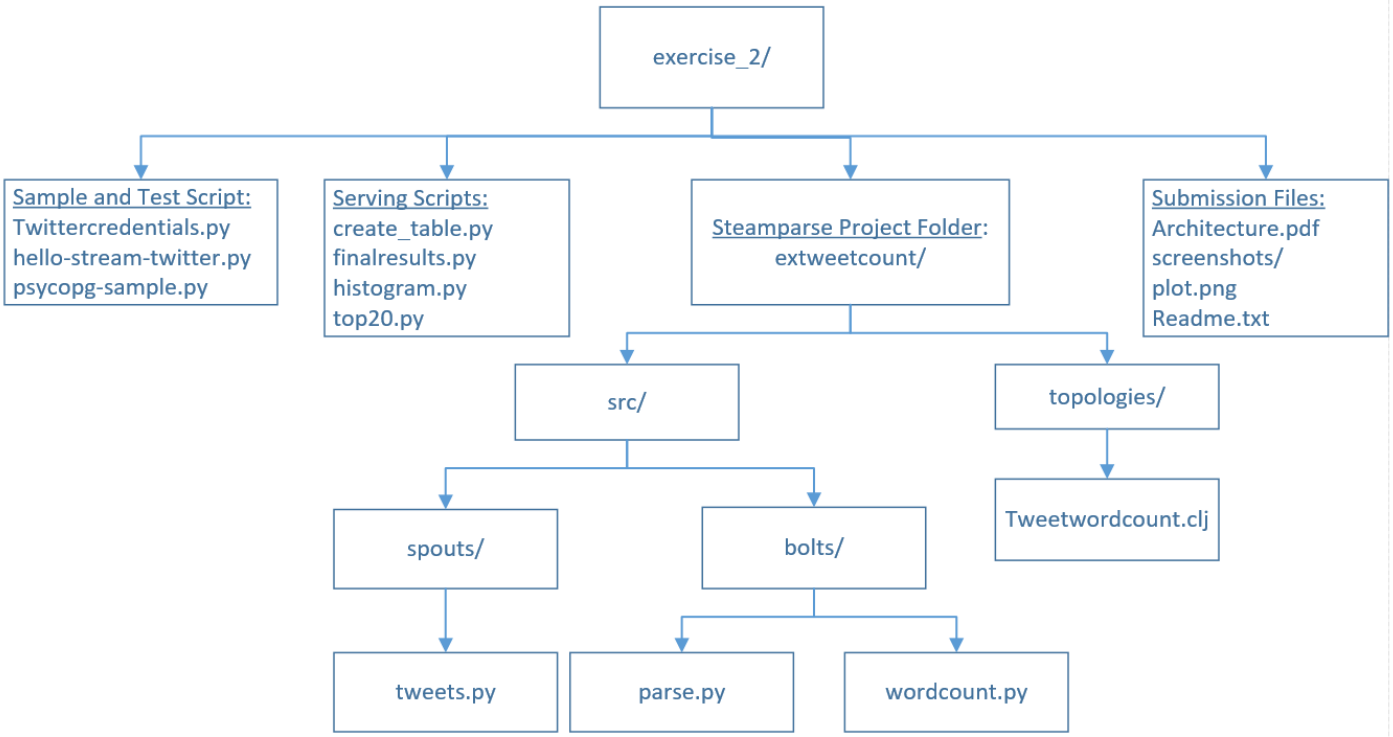


Figure 1: Application Topology

Using Tweepy library, the application reads live tweets data from Twitter in the **Tweet-spout** component. The **Parse-tweet-bolt** component parses the tweets, extracts words from each tweet and emits the words to Count-bolt component. The **Count-bolt** counts the number of each word in the received tuples and updates the counts in the **tweetwordcount** table in the **tcount** Postgres database.

Using psycopg library, the user can run python scripts to extract data and insights from the Postgres table.

File Structure



exercise_2 is the root folder for this application. There are some sample and test scripts to test the Twitter Apps connection. Serving scripts and submission files are saved under the root folder.

extweetcount is the Steamparse project folder. The most important two folders are src and topologies.

- **src** folder contains spouts and bolts components.
    - In spouts folder, tweets.py connects Twitter and read in live data and emit raw tweet data to parse.py which belongs to bolts components.
    - Parse.py split the tweet into word and emit the word to wordcount.py.
    - wordcount.py count the frequency of each word, not only print the log to the screen but also update the count in the postgres database.
- **topologies** folder contains Tweetwordcount.clj which connects spouts and all botls.

Run the Application

Please refer to Readme.txt to run the application.