Exercise 2 Architecture

Younus Ahmed

Spring 2017

The following twitter application pulls tweets from the Twitter streaming API and uses Apache Storm to perform live processing and analysis of data. The Tweepy module is used to interface with twitter using python which pulls live tweets from Twitter. Streamparse is then used to split up the words in each of the tweets and then performs a wordcount. The output of the wordcount bolt stores into a table called tweetwordcount in postgres under database Tcount.

There are also two python executable .py files in the exercise 2 folder called histogram.py that list all of the words that are used with a frequency that is within a range of two numbers that are specified by the user and finalresults.py that either displays the count of a user-specified word or it lists all the words and their counts. If the word is not present in the list of words, the file will display "this word does not exist in this list".

The topology in Apache Storm used to execute the live streaming and analysis of data is described in Figure 1 below, which contains three tweet-spouts, three parse-tweet bolts, and two-count bolts:

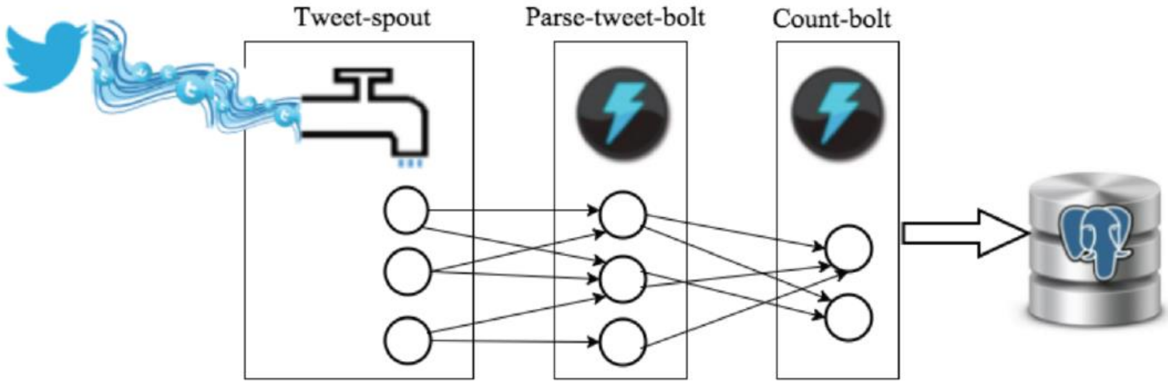

Figure 1: Application Topology

While logged in as w205 in the EC2 instance, the user must first clone github repository under username younusa and change directory into exercise_2. User must then open README.txt to review required initial steps to run the application. It is essential for user to place unique twitter application credentials into relevant fields in the tweets.py file. Please read README.txt for further directions.

File Directories:

README.txt, finalresults.py, histogram.py, Architecture.PDF can be found in

```
/home/w205/w205_2017_spring/exercise_2
```

Topology is stored as tweetwordcount.clj in:

```
/home/w205/w205_2017_spring/exercise_2/extweetwordcount/topologies
```

Spout is stored as tweets.py in:

```
/home/w205/w205_2017_spring/exercise_2/extweetwordcount/src/spouts
```

Parse bolt is stored as parse.py in:

```
/home/w205/w205_2017_spring/exercise_2/extweetwordcount/src/bolts
```

Count-bolt is stored as wordcount.py in:

```
/home/w205/w205_2017_spring/exercise_2/extweetwordcount/src/bolts
```

Three Required Screenshots located in

```
/home/w205/w205_2017_spring/exercise_2/Screenshots
Bar Graph of top 20 most frequently used words located in
/home/w205/w205_2017_spring/exercise_2/plot
```

Tweet-Spout

Uses Tweepy and twitter app tokens to collect tweets from twitter and emit tweets into bolts for further analysis.

Parse Tweet Bolt

Receives tweet from the Tweet spout, filters hash tags, URL's and certain punctuation characters, and then emits the words in each of the tweets into the word-count bolt.

Word-Count Bolt

Logs the count of each of the words that are emitted from the parse tweet bolt. The wordcount.py also contains application-specific code to log the count of each word in a table called tweetwordcount contained in a database called tcount in postgres. The User can log into postgres via "psql -U postgres" and then connect to database tcount to communicate with the tweetwordcount table. The two .py python executable files can also be used to query the count of specific words or all words. In order to execute the python files, user must be in following directory:

```
/home/w205/w205_2017_spring/exercise_2
```

Running the Extweetwordcount Application

The twitter application is dependent on the following applications: Apache Storm, Streamparse, Tweepy, Twitter API, Psycopg, Python, Postgres, and Amazon EC2 Instance.

In order to run the extweetwordcount application, user must be in the following directory:

```
/home/w205/w205_2017_spring/exercise_2/extweetwordcount
```

And must type "sparse run". Once the application has run for a sufficient amount of time, user should type "ctrl-c" to end the counting algorithm.