John Sullivan
Exercise 2

Application and Architecture Description

This application connects to Twitter and downloads tweets, parses them into their constituent words, counts the number of times each word has appeared and writes that information to a data table in Postgres. It does this in several steps, first it connects to the Twitter streaming API with three spouts collecting tweets. It then has two stages of bolts. In the first stage tweets are parsed into individual words, filtering out the contents of tweets that are not words like hashtags, retweets, mentions and URLs. These bolts then emit all valid words. The valid words are passed to the count bolts, which count the number of times each word has appeared and write that information to a database in Postgres. If a word is not in Postgres a new row is created for the word, if the word is in the database its count is incremented by 1.

File Dependencies

There are several file dependencies in the architecture. The tweetwordcount.clj file which controls the running of the spouts and bolts requires the tweets.py spout and the parse.py and wordcount.py bolts to run and those files have to be in the correct directory for the tweetwordcount.clj file to be able to find them.

The script is also dependent on a database named tcount and a data table named tweetwordcount already existing in Postgres.

Running the Application

To run the application make your current working directory w205_2017_spring/
exercise_2/extweetwordcount and type “sparse run” into the terminal. The process will run, collecting and parsing tweets, then writing the words and their counts to the data table, until it is killed by the user.

To run the two serving scripts change your directory to /w205_2017_spring/exercise_2/
extweetwordcount/serving_scripts. The two serving scripts are there. Each program should be run like any other Python program, with arguments after the program name.

File Structure

The git repo is called w205_2017_spring and all application information is in the exercise_2 directory. In that directory there is the readme.txt which describes how to run the application, plog.png which has the bar chart, a screenshots directory which has screenshots to show various stages of the application working, a python file that stores my twitter credentials and an extweetwordcount directory which contains all the program files. There is also a folder called serving_scripts which has the two Python serving scripts in it.

John Sullivan
Exercise 2

The extweetwordcount directory is substantially the same as the version I downloaded from the class Git repo. The topologies directory contains the modified application topology and the src directory contains the modified bolt and spout scripts. Files for each bolt and spout are in their respective directories. The rest of the files in the directory are either what was in the original repo or log files automatically created by the application in the logs directory.