W205

Rudolph Ferrara

Exercise 2

Architecture

Application Folder

The application is a storm application in the EXTwoTweetwordcount directory. On my machine I ran it in /root/ EXTwoTweetwordcount.

Topology

The topology file is in EXTwoTweetwordcount/EXTwoTweetwordcount.clj

Spouts

There is one spout defined in EXTwoTweetwordcount/src/spouts/tweets.py. This file was not changed from the file provided by the assignment other than changing the twitter credentials.

Bolts

There are two bolts defined in EXTwoTweetwordcount/src/bolts 1) parse.py, and 2) wordcount.py. parse.py was not modified.

Tweetwordcount table

Wordcount.py was modified to create the "ex2" database. If the database already exists it will not error and will not drop and recreate the database. In the "ex2" database it creates the Tweetwordcount table. If the table already exists it will not error and will not drop and recreate the table. This way word counts from previous runs can be kept, and a new run can continue counting. The wordcount.py script was modified to keep a count of the words in the Tweetwordcount table. It first tries to increment the count for a word. If no rows are updated, meaning that the word was not yet inserted, it inserts a row for the current count, with the value for the count currently kept track of in the script. This should be the equivalent of inserting a count of 1). Incrementing the count instead of updating it to the current count tracked by the script works better since there might be more than one instance of the bolt running. The update/insert is done within one transaction, and the isolation level is read committed, so there should be no conflicts between multiple instance of the bolt that may be running.

Serving Scripts

The finalresults.py and histogram.py scripts are in the EXTwoTweetwordcount folder – the same folder as the storm application.

Plot.png

The Plot.png bar chart showing the top 20 words was done by running the following query, and plotting the results in Excel:

select * from Tweetwordcount order by count desc fetch first 20 rows only;