MIDS W205 Exercise 2: Twitter Streaming Application

Heather Koo

12/3/2017

Application Idea

The goal of this exercise was to create a Twitter streaming application, which reads a stream of tweets using the Twitter Streaming API, parses the tweets into individual words, and then counts the occurrences of each word in the tweet stream and stores the words and the counts in a Postgres database. There are two serving scripts also included to summarize results by pulling data from the database.

Description of Architecture

The application uses the Tweepy library to read the live stream of tweets from Twitter in the tweet-spout. The Parse-tweet-bolts parses each tweet into individual words and then emits the word into the count-bolt component. Count-bolt counts the number of each word in the received tuples, and updates the counts associated with the word in tweetwordcount table in the tcount Postgres database. The application topology has three instances of tweet-spout, 3 instances of Parse-tweet-bolt, which emits words into 2 instances of Count-bolt.

Directory & File Structure

The main directory is Exercise_2/ in the student GitHub repo. All code is runnable in the UCB MIDS W205 EX2-FULL AMI on Amazon Web Services. The AMI ID is ami-d4dd4ec3.

Program Name	Location	Description
tweets.py	exercise_2/extweetwordcount/src/spouts/	tweet-spout
parse.py	exercise_2/extweetwordcount/src/bolts/	parse-tweet-bolt
wordcount.py	exercise_2/extweetwordcount/src/bolts/	count-bolt
tweetwordcount.clj	exercise_2/extweetwordcount/topologies	topology for the application
create_tcount.py	exercise_2/	Program to create Postgres database tcount and table tweetwordcount
finalresults.py	exercise_2/	Serving Script - returns number of word occurrences in the stream
histogram.py	exercise_2/	Serving Script - returns all words with total number of occurrences equal to or between the arguments

Other Information

The EBS volume used in the class must be attached and mounted to the UCB MIDS W205 EX2-FULL EC2 instance to /data to run Postgres.

Psycopg (PostgreSQL database adapter for Python) must be installed onto your EC2 instance in order to run the scripts properly.
This package can be installed by running `$pip install psycopg2`

Tweepy (twitter package for Python) must be installed to run the tweet spout, to pull twitter streams from the Twitter API.
This package can be installed by running `$pip install tweepy`