Tweet Word Count

Overview

Create a process to stream tweets from Twitter. Each tweet will be broken down into word components and then stored in a database by count.

To setup

- Create an EC2 instance using the "UCB MIDS W205 EX2-FULL" AMI (AMI Id: ami-d4dd4ec3)
- Attach an EBS volume
- Connect to the instance
- Discover the path to your EBS volume (e.g. /dev/xvdf)
  `fdisk -l`
- Download the starting script
  `wget https://raw.github.com/jason-becker/MIDSw205_Exercise2/master/scripts/setup-tweet-word-count.sh`
- To grant permission to execute the script
  `chmod +x setup-tweet-word-count.sh`
- Replace the volume location with your EBS volume location
  `./setup-tweet-word-count.sh [/dev/xvdf]`

The setup script may take a couple minutes to complete.

To stream some tweets:

- `cd/ /root/data`
- `./start-tweet-word-count.sh`

To serve some information about the data:

- `cd/ root/data`
- `python finalresults.py` **OR** `python histogram.py`
  Final results - supply a single word to find out its frequency (optional)
  Histogram - supply a min and max number of occurrences to find out how which words fall into that count range

Source Files

All files and scripts are stored in https://github.com/jason-becker/MIDSw205_Exercise2

Scripts:

Contains .sh and .py files. If you download and execute "setup-tweet-word-count.sh", it will pull modify and place all files in the appropriate folders

Screenshots:

Images of the various steps of the process:

- screenshot-setup: run the setup script on your AMI.
- screenshot-startingstream: this is how you begin streaming tweets from Twitter
- screenshot-twitterstreaming: this is what it looks like if you are successfully taking in data from Twitter
- screenshot-results: this is an example of results

Documentation:

This folder contains a readme with explicit instructions, a visualization of the top 20 words pulled by a sample run of this script, and this architecture document.

Destination File Structure

/root/data

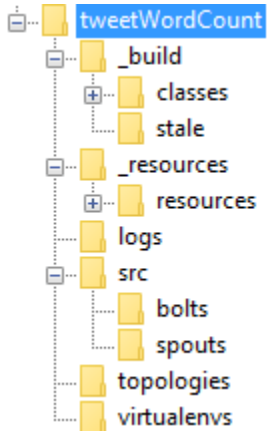
Filename	Filesize	Filetype
..		
finalresults.py	1,391	PY File
histogram.py	1,165	PY File
start-tweet-word-count.sh	48	Shell Script
start_postgres.sh	93	Shell Script
stop_postgres.sh	92	Shell Script

- start-tweet-word-count.sh (for beginning the streaming process)
- finalresults.py and histogram.py (for reporting)
- start_postgres.sh and stop_postgres.sh (for gracefully starting and stopping the database)

/data/MIDSw205_Exercise2

- A clone of the github directory. All content is stored here before unpackaging

/data/tweetWordCount



- Result of a standard sparse quickstart process
- /src/bolts contains parse.py and wordcount.py
- /src/spouts contains tweets.py
- /topologies contains tweetwordcount.clj