W205 Summer 2017
Exercise 2

Himal Suthar

Overview:

In this exercise, we bring together the multiple technologies we have learned in class to build an end to end system that is able to access a live Twitter stream, recognize individual words, then place these words as well as their frequency of occurrence in the Twitter stream into a database for analysis.

Technologies:

Below is a list of the technologies used in this system as well as their function:

1. Twitter – a popular social media platform from which we receive streaming data
2. Tweepy – a Twitter API that allows us to access the streaming data
3. Apache Storm - computation system that allows for the processing of real-time streaming data. This is the tool that is reading the Twitter stream and counting each word
4. Postgres – a relational database into which we store words and their frequency
5. Python – the programming language which we use to bring the above technologies together so that they may communicate with one another.
6. AWS – a cloud computing system on wich we will be running all of the above technologies.

Architecture:

Below is a high level architecture of the system.

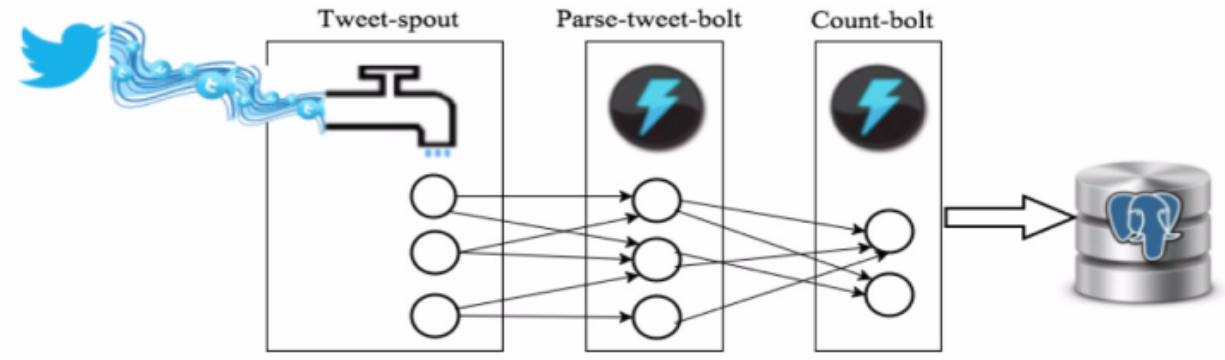

Figure 1: Application Topology

Folder Structure:

exercise_2
- exttweetwordcount
  - src
    - bolts
      - __init__.py
      - parse.py
      - wordcount.py
    - spouts
      - __init__.py
      - tweets.py
  - topologies
    - tweetwordcount.clj
  - virtualenvs
    - wordcount.txt
  - .gitignore
  - README.md
  - config.json
  - fabfile.py
  - project.clj
  - tasks.py
- screenshots
  - README.txt
  - screenshot_database.png
  - screenshot_finalresult.png
  - screenshot_tweetstream.png
- Plot.png
- README.txt
- Twittercredentials.py
- Twittercredentials.pyc
- finalresults.py
- hello-stream-twitter.py
- histogram.py
- psycopg-sample.py
- psycopg-word-count.py