Real-Time Tweet Counts using Apache Storm

Description –
The goal of this project is to capture a stream of tweets from twitter in real-time, split up these tweets into individual words and calculate a running count of the number of words that are being used in tweets in real-time.

Components –

- Message processing framework – Apache Storm
- Data store – PostgreSQL database

Environment -

- Linux with Apache Storm installed (preferably on AWS)

Dependencies -

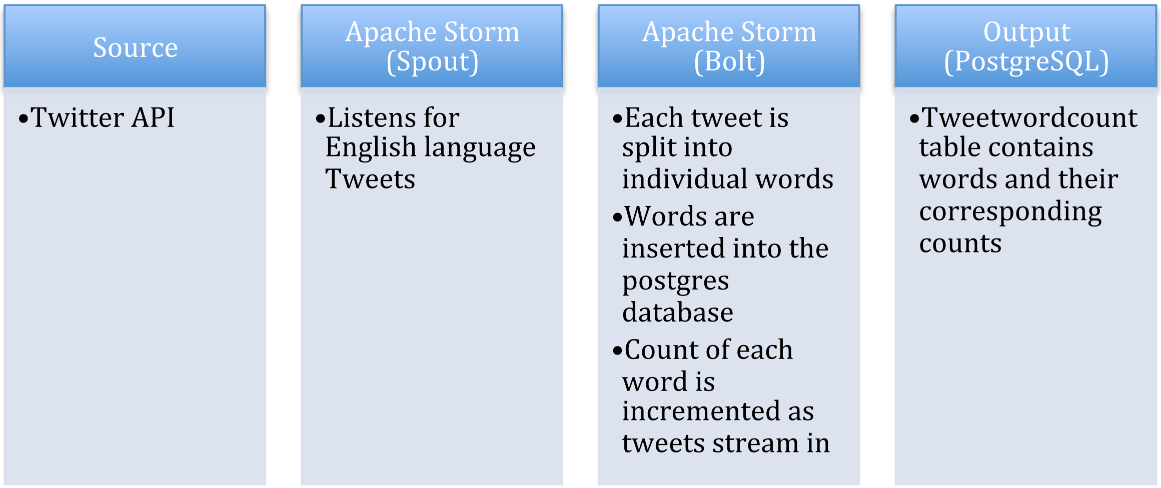
- Install the following python packages using pip
    - **a.** Tweepy
    - **b.** Psycopg2

Directory and File Structure –

Project Folder - EXTweetwordcount

File Name	Path
finalresults.py	ExTweetwordcount/
histogram.py	ExTweetwordcount/
wordcount.py	ExTweetwordcount/src/bolts/wordcount.py
tweets.py	ExTweetwordcount/src/spouts/tweets.py

Architecture



Source

- Twitter API

Apache Storm (Spout)

- Listens for English language Tweets

Apache Storm (Bolt)

- Each tweet is split into individual words
- Words are inserted into the postgres database
- Count of each word is incremented as tweets stream in

Output (PostgreSQL)

- Tweetwordcount table contains words and their corresponding counts

Instructions

- Clone the repository named “EXTweetwordcount” using the command
- Ensure that PostgreSQL is installed and running on the system
- Create a database named “tcount” with username “postgresql” and password “pass”.
- Navigate to the folder of the cloned repository
- Run the command “create_db.py” to create the database
- Run the command “sparse run” from the main project folder
- To view a list of all the words that were loaded into the database,run the script “finalresults.py” from the main project folder
- You may also type in “python finalresult.py wordname” to get the count for a specific word
- Lastly, the file “histogram.py” gives you a list of words whose counts fall within a certain range.
- To run this script, type in “python histogram.py start,end” from the main project folder. For example, “python histogram.py 3,7”