Elizabeth Chabot
W205-Ex2
Architecture

Overview:

Ex2 ingests live tweets from Twitter using Apache Storm. The tweets are filtered, cleaned and parsed into a Postgres database called tcount. In this application we will only look at the frequency of words within the tweets, but please note that the code can be modified to measure a number of different aspects of the incoming tweets.

Topology:

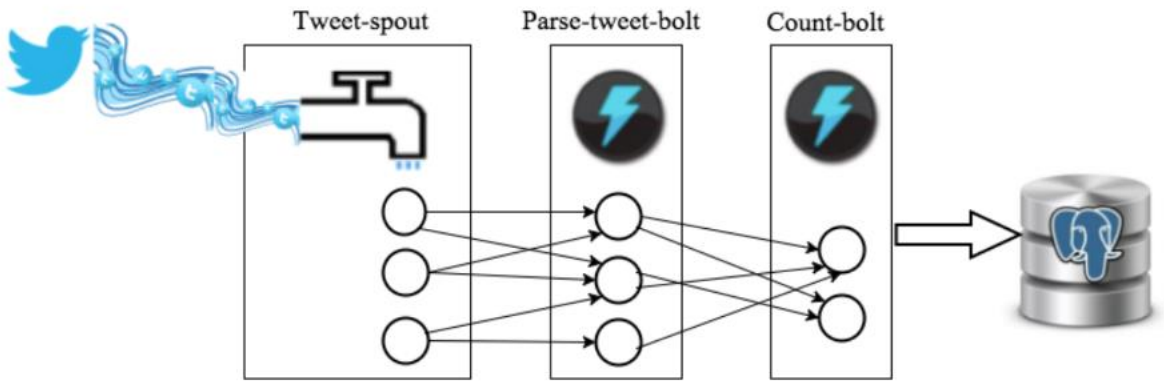

Figure 1: Application Topology

The tweetwordcount.clj file details the above topology. The Tweet-spout feeds English tweets into the parse-bolt. The parse-bolt parses out each tweet into words and the count-bolt takes the parsed words and counts the frequency. The count-bolt then inserts the word count into a postgres table.

File Structure:

Elizabeth Chabot
W205-Ex2
Architecture

Name of the program	Location	Description
tweets.py	ex2/src/spouts/	tweet-spout (will need to be updated with credentials)
parse.py	ex2/src/bolts/	parse-tweet-bolt
wordcount.py	ex2/src/bolts	count-bolt
Twittercredentials.py	ex2/	Twitter API Keys (will need to be made)
tweetwordcount.clj	ex2/topologies/	Topology for the application

Ingest

Live tweets are ingested from Twitter servers using 1. Twitter Application, 2. Tweepy API, 3. Psycop2 (PostgresSQL adapter). See tweets.py to review how spout ingests tweets.

Parse

Tweets are parsed into words through parse.py. In this bolt tweets are cleaned and filtered so that the count bolt will receive only English words.

Count

In count bolt all incoming words are compared to a dictionary and new words are added and withstanding words are given +1.

Store/Save

The count bolt will also store and save the current stream of data in postgres database.

Serve

Python scripts finalres.py and hist.py are examples of how the data can be served. Additionally the Plot.png is another way to serve information.

Elizabeth Chabot
W205-Ex2
Architecture

Dependencies:

You will need access to:

1. Amazon Web Services
2. AWS account and Key
3. Amazon Web Services Machine Image (AMI)
4. Twitter account
5. Twitter application
6. Github Access

Download:

1. Tweepy
2. Psycopg2
3. Python 3.0
4. Text Editor/Terminal (depending on computer may vary)

Basic understanding of:

1. Python
2. SQL
3. Databases
4. AWS
5. Twitter
6. English Language