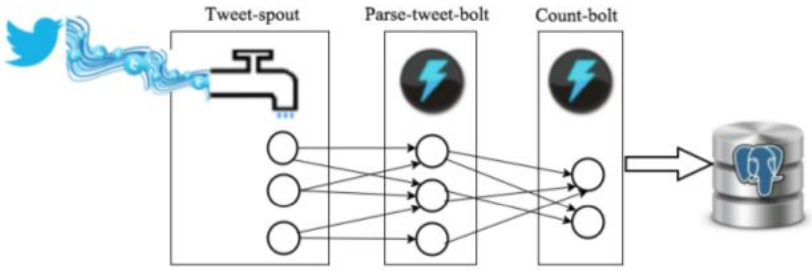
James King
W205-2 -- Spring 2015
Ex2 Architecture

Tweetography:
Counting Words. Really Fast.

Overview

Tweetography uses Apache Storm to ingest live tweets from Twitter, filter them, and store them in a database in near real time.
This has many possible uses including public sentiment awareness, language analysis, and “trending now” analysis.



Architecture

Tweetography is available from Github and can be cloned here.

There are five components to Tweetography: Ingest, Parse, Count, Save, Serve. The following sections give details of each component.

Ingest

The data is ingested directly from the Twitter servers using a Twitter application and the Tweepy API. Tweepy allows filtering results, so this code can be adapted to your needs.

Parse

When tweets are received by the ingest “spout” (data source), they’re directed to one of several parsing “bolts” (processing modules) which breaks the tweet up into it’s component words, filters out some junk, and sends them on to be counted.

Count

In the count bolt, all incoming words are compared to a running dictionary. If the dictionary has the word already, its count is incremented, otherwise it’s added to the dictionary as a new word.

Save

The save function is actually implemented inside the count bolt, but it can be thought of as a separate entity.
In the save function, any previous contents of the data base are cleared and the word count dictionary is used to update a Postgres database with the words and their counts

Serving

Python scripts **finalresults.py** and **histogram.py** are provided as examples of interacting with the collected data, though arbitrary queries against the database can be performed.

File Structure

File Name/Path	Description
EX2Tweetwordcount/src/bolts/wordcount.py	counting and db load bolt
EX2Tweetwordcount/src/bolts/parse.py	tweet parsing bolt
EX2Tweetwordcount/src/spouts/tweets.py	tweet spout
EX2Tweetwordcount/src/serves/finalresults.py	serving file #1
EX2Tweetwordcount/src/serves/.histogram.py	serving file #2
EX2Tweetwordcoung/topologies/tweetwordcount.clj	App topology
EX2Tweetwordcount/Readme.txt	Detailed Installation Instructions
EX2Tweetwordcount/setup.sh	setup script for the application

EX2Tweetwordcount/dbsetup.sql	database setup; run automatically by setup.sh
EX2Tweetwordcount/screenshots	screenshots from running the app
EX2Tweetwordcount/docs/Architecture.pdf	the required documentation
EX2Tweetwordcount/docs/Plot.png	the required bar chart

Dependencies

For this application you'll need three accesses:

1) Amazon Web Services account with keys
2) Amazon Web Services Machine Image (AMI) #ami-003f7f6a (it's in "Community AMIs")
3) Twitter account. You'll need to create a Twitter application here.
4) Github access. Code can be fetched here.