MIDS W205, Fall 2016
Exercise 2
Architecture documentation
Laura Williams

Summary

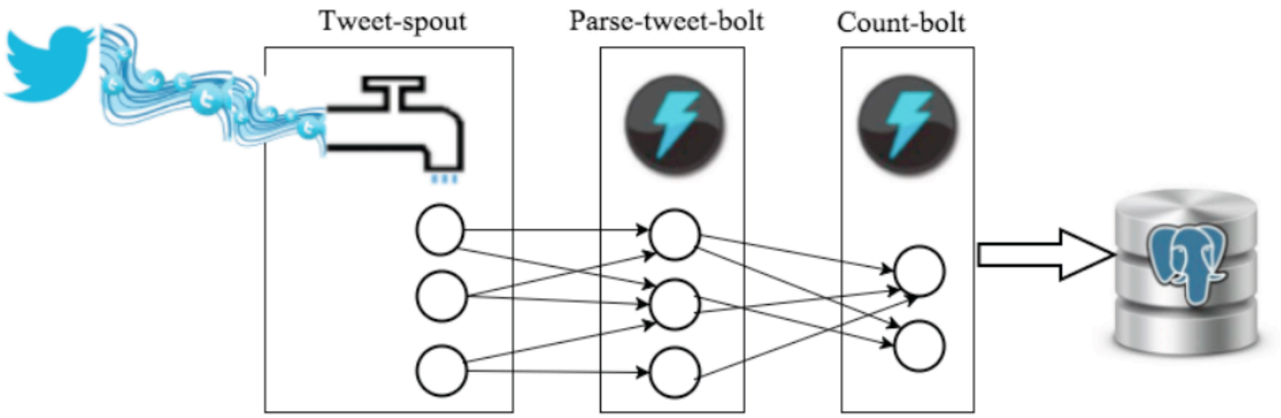
This project receives tweets from a Twitter stream, outputs a relational database for words and word counts from that stream, and supplies two pre-written queries to that database.

Architecture Concept and Technology Used

The project uses Apache Storm topology, including spouts and bolts, to route a Twitter stream into a relational database in Postgres, and includes pre-written queries using python and the psycopg2 library for analyzing the Postgres database.

The project is executed on an Amazon Web Services EC2 instance with Postgres, Python and Streamparse already installed.

The following figure visualizes the architecture for this project; it is taken from the instructions for this exercise:



Streamparse is used to integrate Python with the Apache Storm topology of spouts and bolts.

An application was created in Twitter to authenticate connecting to the Twitter stream. The name of this Twitter application is W205_Exercise2_LW.

Description of each step in the Apache Storm Topology

Tweet-spout
The python Tweepy library is used in the Streamparse spout to connect the stream from the Twitter API into the Apache Storm spout. The architecture creates three parallel processes of the spout, which all feed tweets into the next bolt in the architecture, the Parse-tweet-bolt.

Parse-tweet-bolt
This bolt receives tweets from the Tweet-spout, strips out hashtags, retweets, user mentions, urls, and leading and trailing punctuation, and splits each tweet into single words. The architecture creates three parallel processes of this bolt, which all receive tweets from the Tweet-spout, and which all feed single words into the final bolt of the architecture, the Count-bolt.

Count-bolt
This bolt receives words from the Parse-tweet-bolt, counts the number of words, and then uses the python psycopg2 library to update a Postgres SQL database with the words received and the updated word counts. The architecture creates two parallel processes of this bolt, which both update the Postgres database. This bolt is intended to be run with an empty database to count the words from a single Twitter stream. This bolt is not currently written such that it could update a database already containing data from a previous stream, but the code in this bolt could be adapted to work in that way.

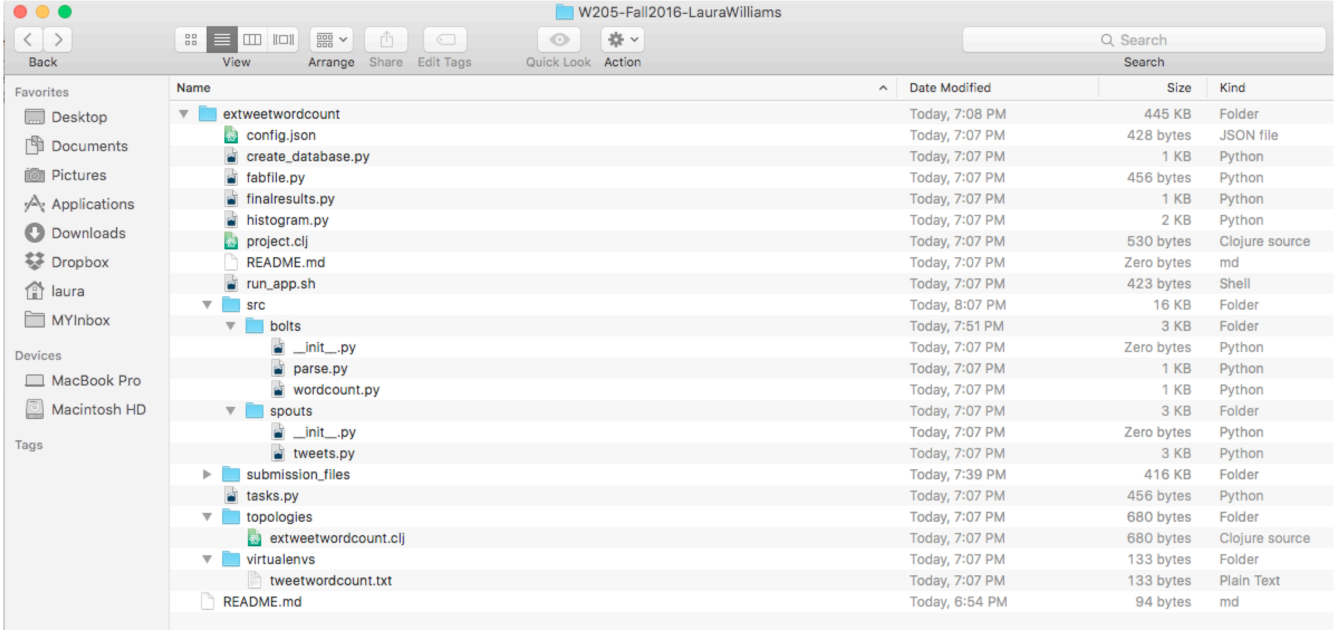
Tcount database
The postgres database is called tcount, and the table inside the database is called tweetwordcount. The database and table are created in advance of running the Streamparse project, in a file called create-database.py. This file must be run immediately before running each streamparse project that calls a Twitter steam, so that each time a Twitter stream is called, the project begins with an empty database.

Queries

Two pre-written queries are included in the project. One will return the number of occurrences in the Twitter stream of any word. The other will search for words that have word counts between, and including, two integers.

Required files

All necessary files and the file structure of the project are outlined below.



Name	Date Modified	Size	Kind
extweetwordcount	Today, 7:08 PM	445 KB	Folder
config.json	Today, 7:07 PM	428 bytes	JSON file
create_database.py	Today, 7:07 PM	1 KB	Python
fabfile.py	Today, 7:07 PM	456 bytes	Python
finalresults.py	Today, 7:07 PM	1 KB	Python
histogram.py	Today, 7:07 PM	2 KB	Python
project.clj	Today, 7:07 PM	530 bytes	Clojure source
README.md	Today, 7:07 PM	Zero bytes	md
run_app.sh	Today, 7:07 PM	423 bytes	Shell
src	Today, 8:07 PM	16 KB	Folder
bolts	Today, 7:51 PM	3 KB	Folder
__init__.py	Today, 7:07 PM	Zero bytes	Python
parse.py	Today, 7:07 PM	1 KB	Python
wordcount.py	Today, 7:07 PM	1 KB	Python
spouts	Today, 7:07 PM	3 KB	Folder
__init__.py	Today, 7:07 PM	Zero bytes	Python
tweets.py	Today, 7:07 PM	3 KB	Python
submission_files	Today, 7:39 PM	416 KB	Folder
tasks.py	Today, 7:07 PM	456 bytes	Python
topologies	Today, 7:07 PM	680 bytes	Folder
extweetwordcount.clj	Today, 7:07 PM	680 bytes	Clojure source
virtualenvs	Today, 7:07 PM	133 bytes	Folder
tweetwordcount.txt	Today, 7:07 PM	133 bytes	Plain Text
README.md	Today, 6:54 PM	94 bytes	md

Exceptions:

Two README.md files are git files and are not necessary for the project to run.

The submission_files folder holds files required for this file submission, and are not necessary for this project to run.

Executing the Application

Detailed instructions for carrying out each step of this project are included in the readme.txt file that accompanies this architecture. I will not duplicate that information here. Please proceed to that document for carrying out the application.