CS172 Project - Part 1

Nathaniel Sarkissian

Lin Wang

1 November 2016

**Collaboration Details**

Nathaniel Sarkissian

Developed python script for tweet streaming.

Lin Wang

Developed bash script for interfacing with the python script.

**Overview**

Architecture

We used the python library *tweepy* to interact with Twitter’s API. This simplifies the application authentication and stream initiation.

Storage Strategy

The storage of the data is based on a simple chunking strategy. Small chunks are filled with new tweets. Once the current chunk reaches the maximum chunk size, it is written to a file. Once a certain number of chunks have been written to one file, a new file is created and the process begins again.

Data Structures

The only data structure we use is a python dictionary, which is essentially a hash with strings as keys and values. We use this because python has a very straightforward module that converts json to a python dictionary.

**Limitations**

If the user’s computer is unable to keep up with the number of tweets arriving, twitter will cut the connection. The python script does detect this and attempt to reset the connection, but tweets will be missed during the reinitialization of the stream.

**Instructions**

To run the twitter crawler: *./****twitterStream.sh*** *<****chunk size****> <****chunk count****>*

**Chunk size** corresponds to the size of the chunks being written to the file.

**Chunk count** corresponds to the number of chunks per file.

Smaller chunk size will result in more file writes, while larger chunk size will require more RAM use.

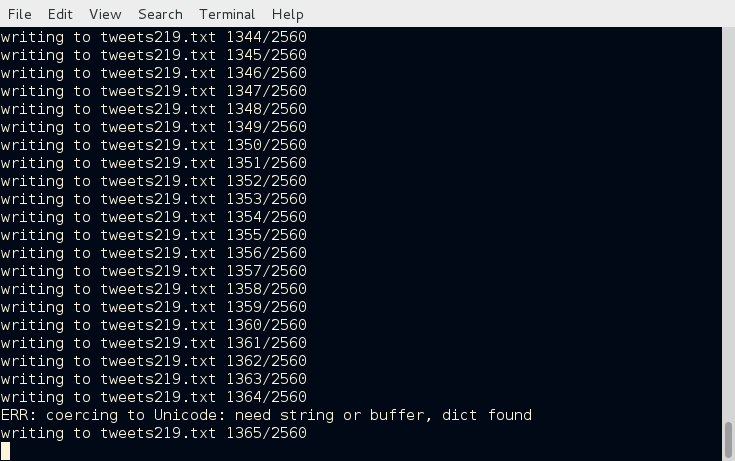
The size of each file can be estimating using the following equation:

file size = (chunk count)\*(chunk size)\*(average bytes per tweet)

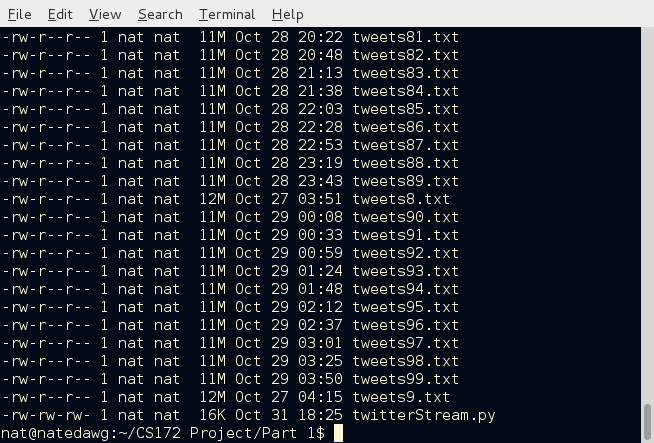
So if the average tweet is 100 bytes, chunk size is 1024, and chunk count is 10240, the files will all be about 10MB. This will vary slightly because tweets are not always the same length.

Requires the python module *tweepy*, and python 2.7.

**Screenshots**



This screenshot displays the output of the streamer as it is collecting tweets. It is currently writing to the 219th file and is working on chunk 1366 out of 2560. Every now and then there will be a small error with an unknown character(usually an emoji). For the sake of speed, it is easier to throw out the whole tweet and move on.



This screenshot displays the files that have accumulated in the directory. Each one is about 10MB, give or take a few MB, as expected.