

CS172 Crawler Project Part A

Group: Spencer Lee and Jeffrey Chen

Collaboration Details

Spencer handled development of the project due to his prior experience and familiarity with Java. Jeffrey handled the initial setup and documentation.

Overview

Libraries used:

Twitter4j-core: Provides a Java library that allows easy communication with the Twitter API.

Twitter4j-stream: Extends the core library to allow easy communication with the Twitter Streaming API.

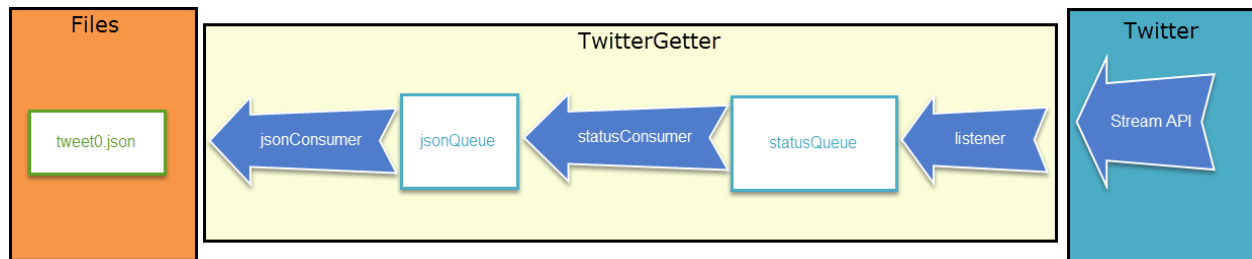
Gson: Google's JSON interpreter that is used to decode JSON that Twitter uses. It is also used to reencode strings into JSON format.

Jsoup: Provides a handful of classes that allow HTML to be downloaded and parsed into human-readable strings. The crawler uses it to get the title of the webpage and store it as part of Tweet information.

Crawler runtime high-level overview:

1. The crawler prompts the user for the number of files to be written to and the size of each file. The current default values are 1 10MB file.
2. The files are created, statusQueue and jsonQueue are initialized.
3. Multiple writers are created, one writer per file.
4. The full authentication key and token are initialized and validated with Twitter servers.
5. A Twitter4j TwitterStream then samples and returns only Tweets with a geolocation tag.
6. If the Tweet contains a link, that link's title is grabbed and stored.
7. Tweets returned by the TwitterStream producer are pushed onto statusQueue.
8. The Tweets are then processed by multiple statusConsumers converting the raw data into a Tweet object, then converting it again into a predefined JSON format. The information is then pushed into jsonQueue.
9. jsonQueue is accessed by multiple jsonConsumers which handle writing the JSON to the initialized files.
10. Steps 5-9 are repeated until each file reaches the defined file size.
11. The crawler stops all jsonConsumers, then stops the twitterStream producer, then statusConsumers.
12. The links to files are severed.

Architecture



Crawling strategy

Twitter allows only a single connection for each application. This denies the ability to create multiple connections to speed up data collection. However, due to the additional processing that needs to be done to convert stream data, steps have been taken to ensure no data is lost and each Tweet gets processed. All Tweets gathered are stored inside of queues that serve as buffers. From these queues, multiple consumers continually monitor and process data.

Data Structures employed

statusQueue: stores parsed raw stream data

jsonQueue: stores data ready to be written to file

class Tweet: class used to store all relevant Tweet information including screenname, hashtags, link title, etc.

class Link: class used to store a link's URL and title

Limitations

Steps have been taken to optimize the crawler data manipulation. Protection against program crashes has not been rigorously tested. Try/catch statements should handle the majority of potential issues. Additionally, only a low percentage of the Twitter population userbase uses the geolocation feature. The end result is that there are far less Tweets to process resulting in a reduced data accumulation rate of approximately 15 minutes for 10MB of Tweet data. This rate is variable dependent on the time of day.

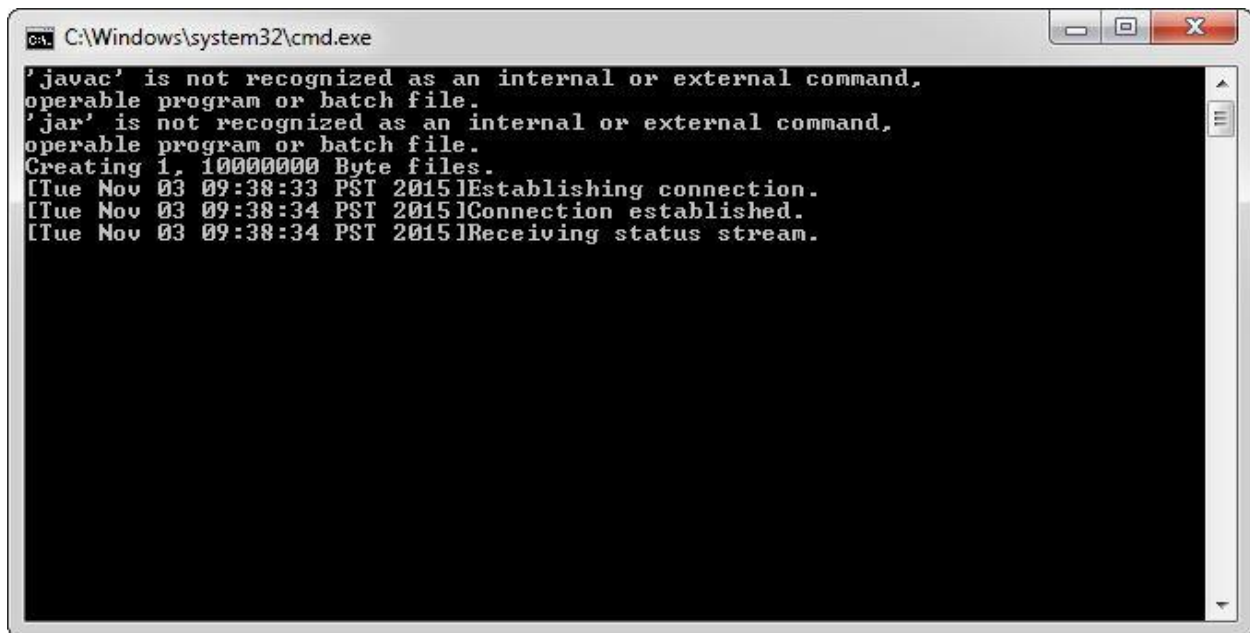
Usage instructions

Run crawler.bat. Tweets will be stored at ".\Tweets\tweet0.json". Default parameters are 1 10MB file.

Alternatively, run .\TwitterGetter.jar <num-files: 1> <bytesPerFile: 10000000>

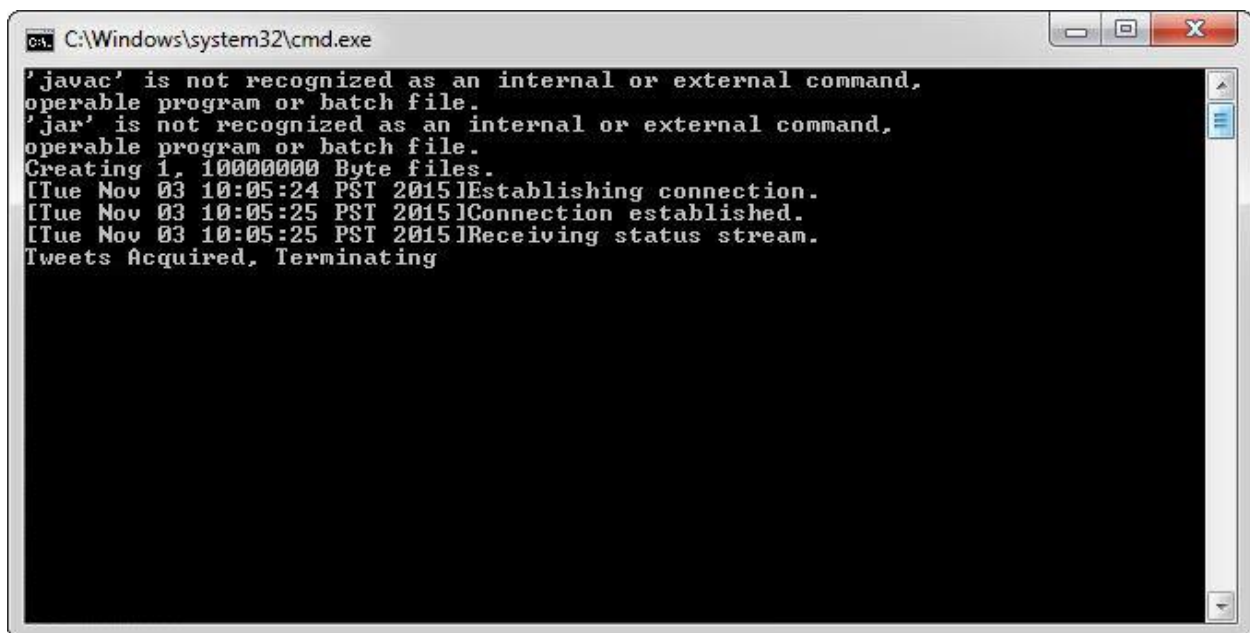
Screenshots

Running In-progress



```
C:\Windows\system32\cmd.exe
'javac' is not recognized as an internal or external command,
operable program or batch file.
'jar' is not recognized as an internal or external command,
operable program or batch file.
Creating 1, 100000000 Byte files.
[Tue Nov 03 09:38:33 PST 2015]Establishing connection.
[Tue Nov 03 09:38:34 PST 2015]Connection established.
[Tue Nov 03 09:38:34 PST 2015]Receiving status stream.
```

Completed



```
C:\Windows\system32\cmd.exe
'javac' is not recognized as an internal or external command,
operable program or batch file.
'jar' is not recognized as an internal or external command,
operable program or batch file.
Creating 1, 100000000 Byte files.
[Tue Nov 03 10:05:24 PST 2015]Establishing connection.
[Tue Nov 03 10:05:25 PST 2015]Connection established.
[Tue Nov 03 10:05:25 PST 2015]Receiving status stream.
Tweets Acquired, Terminating
```

10MB File

