CS172 Project: Twitter Crawler Report

Team Members - Jackie Chan (861158919)
Abe Hu (861148832)
Daniel Nguyen (861156246)

Collaboration Details:

Jackie Chan:

- Installed python package manager and necessary packages
- Set up software versioning
- Set up tweepy API to stream tweets

Abe Hu:

- Planned and Designed architecture
- Reformatted data from stream
- Bug fixes and manual verification of system

Daniel Nguyen:

- Complete 4-5 page report
- Retrieved and parsed tweet urls
- JSON object manipulation for titles

Overall Collaboration Splits:

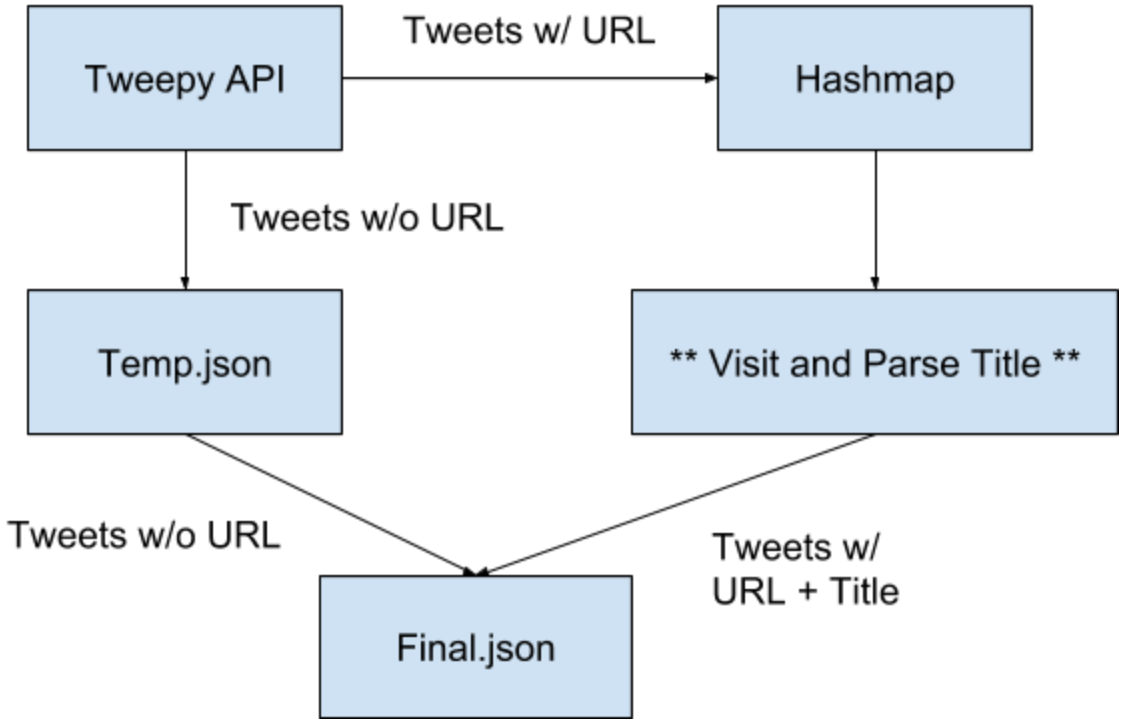
Jackie Chan - 33.33%

Abe Hu - 33.33%

Daniel Nguyen - 33.33%

Overview of systems:

a. Architecture:



b. Data Collection Strategy:

We used the Tweepy API to grab a stream of tweets. The user is able to give command line arguments for how many tweets to grab before the program terminates the stream. As we retrieve data from the stream, we immediately input the raw data into either a hashmap or a json file.

c. Data Structures and Architecture Explained:

The only data structures we used were a hashmap and temporary arrays for small computations and utilities. As explained in part b, we immediately store the stream of tweets into either a hashmap or temporary json file. We do this because we do not want to perform any expensive computations. If expensive computations are performed, we might have skipped data. As a result, we leave data manipulation until the stream terminates. If a tweet has any URLs, we store them into a hashmap.

Otherwise, we store all other tweets into a regular json file. With this structure, we have a clear separation between tweets with URLs and no URLs.

We use a hashmap because of the possibility of duplicate URLs. Avoiding duplicate URLs could save the expensive operation of retrieving the same page title from the same URL. The key to the hash would be the URL, and the value would be the raw JSON object. It would be much harder to achieve the same outcome with arrays.

After the stream ends, our python script traverses through the hashmap and fetches page titles from the URLs. Then, the script injects the page title as a new attribute into the JSON object. After all the data has been injected with page titles, we write all the JSON objects from both the temporary file and hashmap to a new JSON file. The end result is a combination of edited tweets with page title attributes and unedited tweets with no page title attributes.

Limitations of Systems:

- The user may define number of tweets, but can not define the memory size of tweets (5 MB).
- The system may not handle extremely large amounts of data (overflow)

Instructions for system deployment:

1. Install python
2. Install pip (python package manager)
3. Install all python packages
    a. Tweepy
    b. BeautifulSoup
    c. Requests
4. Open command line and execute python script

a. '*python project.py* **[**output-file.json**] [**# of Tweets to Stream**] [**Search Term**]**'
b. Example: '*python project.py final.json 50 cars*'
    i. Gets 50 tweets that include keyword: *cars* and writes them to file: *final.json*

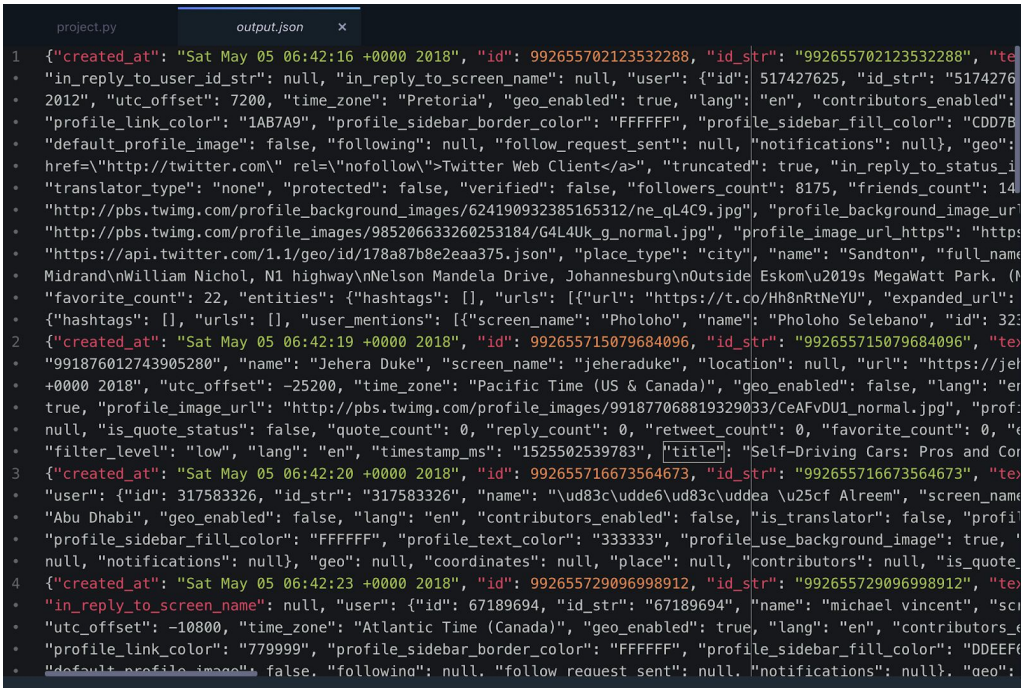
System in action:

Execute script with appropriate command-line arguments

```
Jackies-MacBook-Pro:CS172Project JackieC$ python3 project.py output.json 10 cars
output.json contains the tweets with a title attribute
cars are the keywords
1
2
3
4
5
6
7
8
9
10
No URLs in tweet
Self-Driving Cars: Pros and Cons for the Public's Health – helplibrary familysaurus com
No URLs in tweet
No URLs in tweet
No URLs in tweet
365 Days of Motoring on Twitter: "BL Cars Ltd reintroduced the MG marque as a non-sporting variant of its Metro on this day in 1982.
https://t.co/EvLrDAmK3a… https://t.co/PztRexHkot"
Brandon Ayash on Twitter: "@JerrySeinfeld  You asked someone, in Comedians In Cars. "If you would rather be smart, or funny "

I think that alot of funny people are smart.. My last name is Jewish, and im left handed.  Nice to meet you Jerry."
No URLs in tweet
No URLs in tweet
No URLs in tweet
Jackies-MacBook-Pro:CS172Project JackieC$
```

Final json file after script execution (See 'title' attribute in second object')



```
{"created_at": "Sat May 05 06:42:16 +0000 2018", "id": 992655702123532288, "id_str": "992655702123532288", "text": "RT @Pholoho: -Johannesb
{"created_at": "Sat May 05 06:42:19 +0000 2018", "id": 992655715079684096, "id_str": "992655715079684096", "text": "Self-Driving Cars: Pros
{"created_at": "Sat May 05 06:42:20 +0000 2018", "id": 992655716673564673, "id_str": "992655716673564673", "text": "\u201cYou\u2019re the c
{"created_at": "Sat May 05 06:42:23 +0000 2018", "id": 992655729096998912, "id_str": "992655729096998912", "text": "San Francisco culture i
{"created_at": "Sat May 05 06:42:23 +0000 2018", "id": 992655729076121600, "id_str": "992655729076121600", "text": "RT @JajaPhD: He also st
{"created_at": "Sat May 05 06:42:26 +0000 2018", "id": 992655741159952385, "id_str": "992655741159952385", "text": "BL Cars Ltd reintroduce
{"created_at": "Sat May 05 06:42:27 +0000 2018", "id": 992655748625846272, "id_str": "992655748625846272", "text": "@JerrySeinfeld  You ask
{"created_at": "Sat May 05 06:42:28 +0000 2018", "id": 992655751561785345, "id_str": "992655751561785345", "text": "RT @Woyshnis17: My bigg
{"created_at": "Sat May 05 06:42:31 +0000 2018", "id": 992655765738418177, "id_str": "992655765738418177", "text": "RT @BrooklynSpoke: \"Ca
{"created_at": "Sat May 05 06:42:34 +0000 2018", "id": 992655775263698944, "id_str": "992655775263698944", "text": "@ColleenB123 Gasoline a
{"created_at": "Sat May 05 06:42:16 +0000 2018", "id": 992655702123532288, "id_str": "992655702123532288", "text": "RT @Pholoho: -Johannesb
{"created_at": "Sat May 05 06:42:19 +0000 2018", "id": 992655715079684096, "id_str": "992655715079684096", "text": "Self-Driving Cars: Pros
{"created_at": "Sat May 05 06:42:20 +0000 2018", "id": 992655716673564673, "id_str": "992655716673564673", "text": "\u201cYou\u2019re the c
{"created_at": "Sat May 05 06:42:23 +0000 2018", "id": 992655729096998912, "id_str": "992655729096998912", "text": "San Francisco culture i
{"created_at": "Sat May 05 06:42:23 +0000 2018", "id": 992655729076121600, "id_str": "992655729076121600", "text": "RT @JajaPhD: He also st
{"created_at": "Sat May 05 06:42:26 +0000 2018", "id": 992655741159952385, "id_str": "992655741159952385", "text": "BL Cars Ltd reintroduce
{"created_at": "Sat May 05 06:42:27 +0000 2018", "id": 992655748625846272, "id_str": "992655748625846272", "text": "@JerrySeinfeld  You ask
{"created_at": "Sat May 05 06:42:28 +0000 2018", "id": 992655751561785345, "id_str": "992655751561785345", "text": "RT @Woyshnis17: My bigg
{"created_at": "Sat May 05 06:42:31 +0000 2018", "id": 992655765738418177, "id_str": "992655765738418177", "text": "RT @BrooklynSpoke: \"Ca
{"created_at": "Sat May 05 06:42:34 +0000 2018", "id": 992655775263698944, "id_str": "992655775263698944", "text": "@ColleenB123 Gasoline a
{"created_at": "Sat May 05 06:46:02 +0000 2018", "id": 992656650489204736, "id_str": "992656650489204736", "text": "RT @ifntinfo: [VIDEO] 1
{"created_at": "Sat May 05 06:46:05 +0000 2018", "id": 992656662430437376, "id_str": "992656662430437376", "text": "RT @Hannasuewilson: We\
```