Project A: Twitter dataset Analysis

Name : Radhika Yogesh Kulkarni

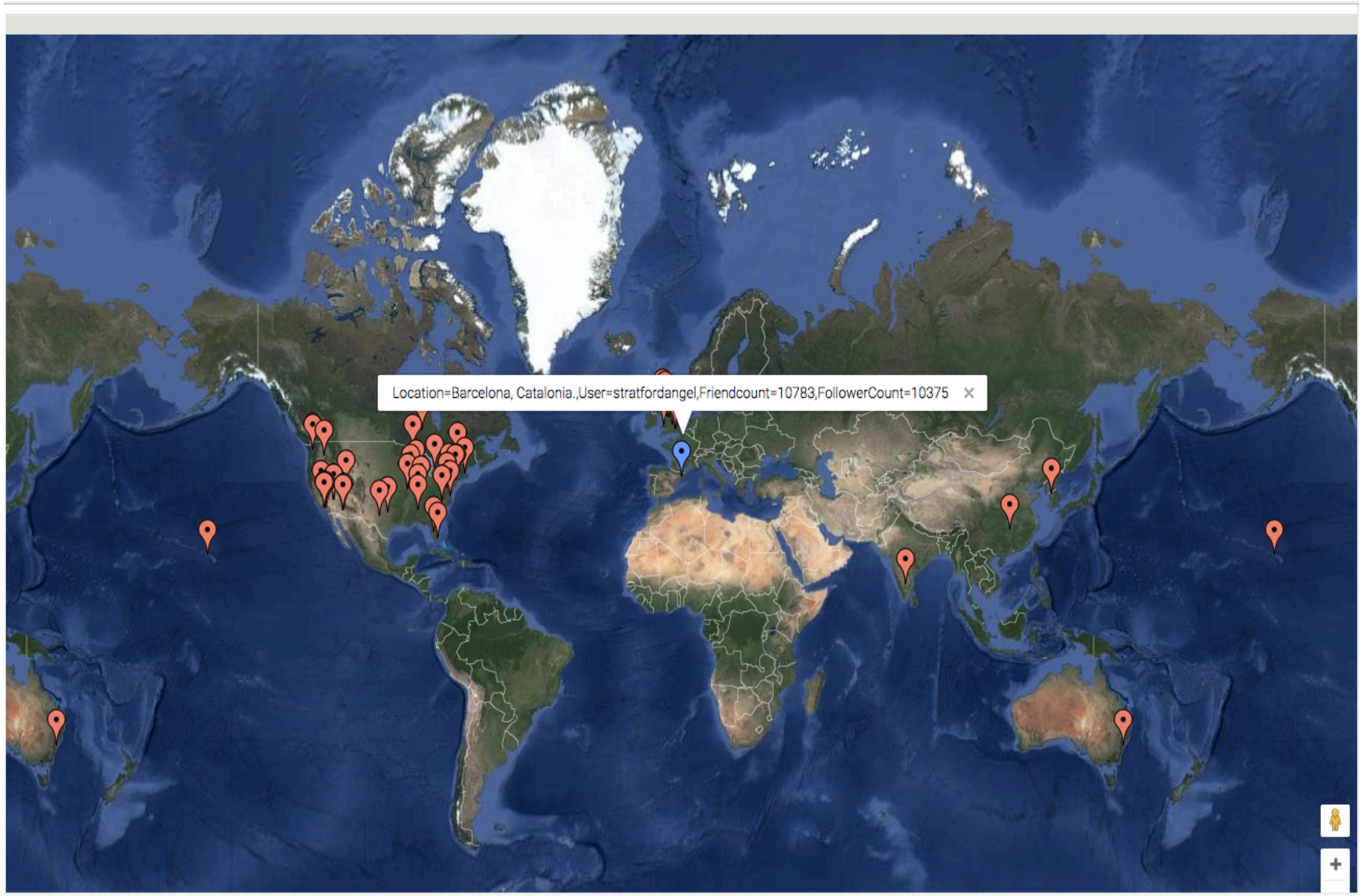
UID: 2000156979

Jetstream ID : rykulkar

VM Name : UID: 2000156979

Criteria selected to plot around 50 valid geolocations

“Twitter users which are Active with minimum 5000 followers and minimum 5000 friends all over the world”



Location=Barcelona, Catalonia.,User=stratfordangel,Friendcount=10783,FollowerCount=10375

a. *References used ~*

- ✓ https://wiki.cites.illinois.edu/wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012
- ✓ http://docs.mongodb.org/manual/reference/program/mongodump
- ✓ https://docs.mongodb.com/manual/reference/program/mongoexport/
- ✓ https://developers.google.com/chart/interactive/docs/gallery/map
- ✓ https://developers.google.com/maps/documentation/javascript/tutorialVisualization
- ✓ https://developers.google.com/chart/interactive/docs/reference
- ✓ http://docs.MongoDB.org/manual/core/crud-introduction/#query
- ✓ http://docs.mongodb.org/manual/reference/command
- ✓ http://codebeautify.org/htmlviewer/#

b. Answers the following questions:

i. How many locations were you able to validate (i.e., geolocate)? What is the remaining number? Give suggestions for resolving those that you were not able to resolve.

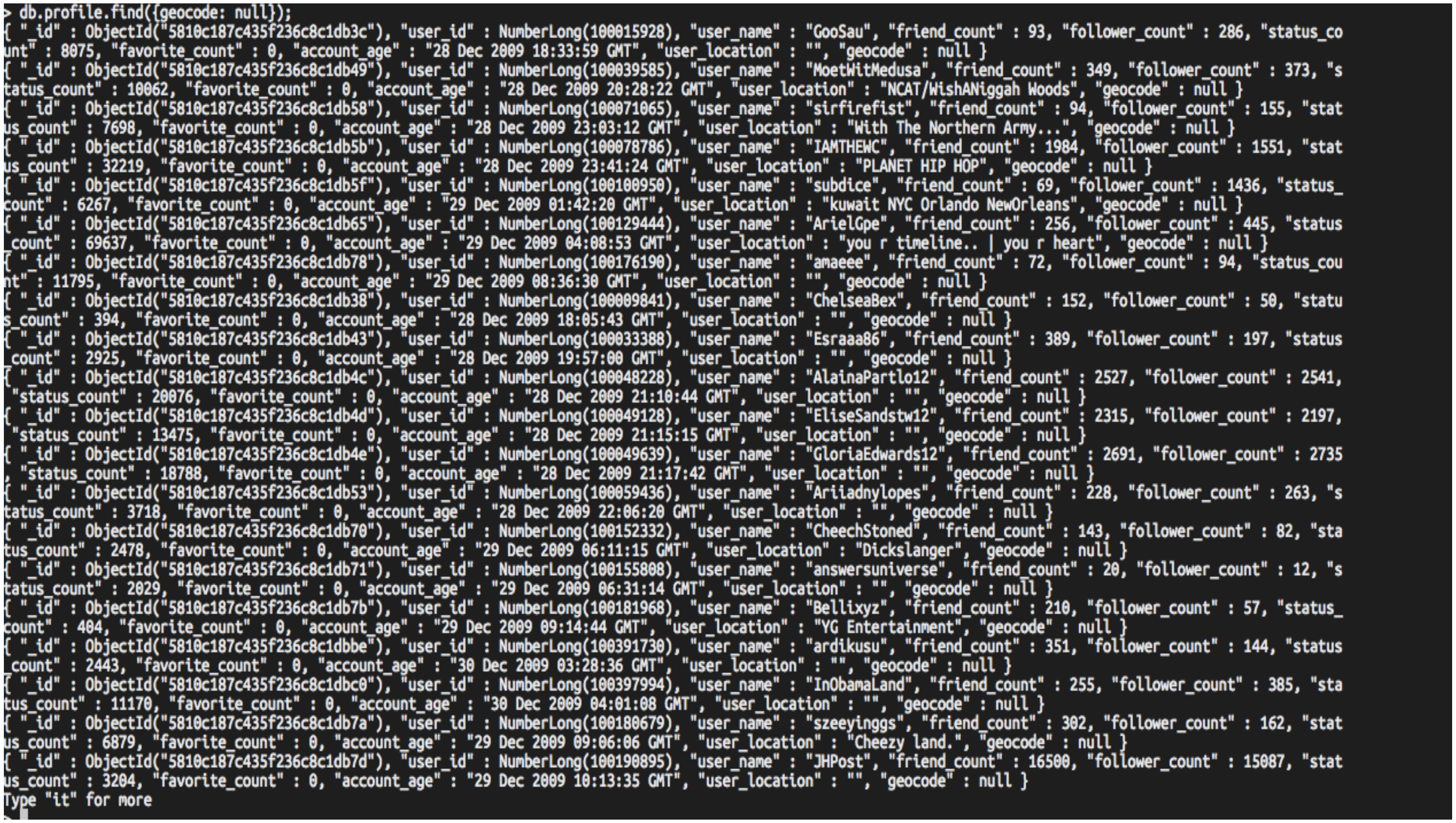
- **I was able to validate 7229 locations by using the script provided in the source code.**

```
> db.profile.count({geocode: {$ne : null} });
7229
```

- **I was not able to validate 2771 locations using script provided**

```
> db.profile.count({geocode: null});
2771
```

Few of the records which were not resolved are as in below example:



```
> db.profile.find({geocode: null});
{ "_id" : ObjectId("5810c187c435f236c8c1db3c"), "user_id" : NumberLong(100015928), "user_name" : "GooSau", "friend_count" : 93, "follower_count" : 286, "status_co
unt" : 8075, "favorite_count" : 0, "account_age" : "28 Dec 2009 18:33:59 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db49"), "user_id" : NumberLong(100039585), "user_name" : "MoetWitMedusa", "friend_count" : 349, "follower_count" : 373, "s
tatus_count" : 10062, "favorite_count" : 0, "account_age" : "28 Dec 2009 20:28:22 GMT", "user_location" : "NCAT/WishANiggah Woods", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db58"), "user_id" : NumberLong(100071065), "user_name" : "sirfirefist", "friend_count" : 94, "follower_count" : 155, "stat
us_count" : 7698, "favorite_count" : 0, "account_age" : "28 Dec 2009 23:03:12 GMT", "user_location" : "With The Northern Army...", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db5b"), "user_id" : NumberLong(100078786), "user_name" : "IAMTHEWC", "friend_count" : 1984, "follower_count" : 1551, "stat
us_count" : 32219, "favorite_count" : 0, "account_age" : "28 Dec 2009 23:41:24 GMT", "user_location" : "PLANET HIP HOP", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db5f"), "user_id" : NumberLong(100100950), "user_name" : "subdice", "friend_count" : 69, "follower_count" : 1436, "status_
count" : 6267, "favorite_count" : 0, "account_age" : "29 Dec 2009 01:42:20 GMT", "user_location" : "kuwait NYC Orlando NewOrleans", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db65"), "user_id" : NumberLong(100129444), "user_name" : "ArielGpe", "friend_count" : 256, "follower_count" : 445, "status
_count" : 69637, "favorite_count" : 0, "account_age" : "29 Dec 2009 04:08:53 GMT", "user_location" : "you r timeline.. | you r heart", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db78"), "user_id" : NumberLong(100176190), "user_name" : "amaeee", "friend_count" : 72, "follower_count" : 94, "status_cou
nt" : 11795, "favorite_count" : 0, "account_age" : "29 Dec 2009 08:36:30 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db38"), "user_id" : NumberLong(100009841), "user_name" : "ChelseaBex", "friend_count" : 152, "follower_count" : 50, "statu
s_count" : 394, "favorite_count" : 0, "account_age" : "28 Dec 2009 18:05:43 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db43"), "user_id" : NumberLong(100033388), "user_name" : "Esraaa86", "friend_count" : 389, "follower_count" : 197, "status
_count" : 2925, "favorite_count" : 0, "account_age" : "28 Dec 2009 19:57:00 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db4c"), "user_id" : NumberLong(100048228), "user_name" : "AlainaPartlo12", "friend_count" : 2527, "follower_count" : 2541,
 "status_count" : 20076, "favorite_count" : 0, "account_age" : "28 Dec 2009 21:10:44 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db4d"), "user_id" : NumberLong(100049128), "user_name" : "EliseSandstw12", "friend_count" : 2315, "follower_count" : 2197,
 "status_count" : 13475, "favorite_count" : 0, "account_age" : "28 Dec 2009 21:15:15 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db4e"), "user_id" : NumberLong(100049639), "user_name" : "GloriaEdwards12", "friend_count" : 2691, "follower_count" : 2735
, "status_count" : 18788, "favorite_count" : 0, "account_age" : "28 Dec 2009 21:17:42 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db53"), "user_id" : NumberLong(100059436), "user_name" : "Ariiadnylopes", "friend_count" : 228, "follower_count" : 263, "s
tatus_count" : 3718, "favorite_count" : 0, "account_age" : "28 Dec 2009 22:06:20 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db70"), "user_id" : NumberLong(100152332), "user_name" : "CheechStoned", "friend_count" : 143, "follower_count" : 82, "sta
tus_count" : 2478, "favorite_count" : 0, "account_age" : "29 Dec 2009 06:11:15 GMT", "user_location" : "Dickslanger", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db71"), "user_id" : NumberLong(100155808), "user_name" : "answersuniverse", "friend_count" : 20, "follower_count" : 12, "s
tatus_count" : 2029, "favorite_count" : 0, "account_age" : "29 Dec 2009 06:31:14 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db7b"), "user_id" : NumberLong(100181968), "user_name" : "Bellixyz", "friend_count" : 210, "follower_count" : 57, "status_
count" : 404, "favorite_count" : 0, "account_age" : "29 Dec 2009 09:14:44 GMT", "user_location" : "YG Entertainment", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1dbbe"), "user_id" : NumberLong(100391730), "user_name" : "ardikusu", "friend_count" : 351, "follower_count" : 144, "status
_count" : 2443, "favorite_count" : 0, "account_age" : "30 Dec 2009 03:28:36 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1dbc0"), "user_id" : NumberLong(100397994), "user_name" : "InObamaLand", "friend_count" : 255, "follower_count" : 385, "sta
tus_count" : 11170, "favorite_count" : 0, "account_age" : "30 Dec 2009 04:01:08 GMT", "user_location" : "", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db7a"), "user_id" : NumberLong(100180679), "user_name" : "szeeyinggs", "friend_count" : 302, "follower_count" : 162, "stat
us_count" : 6879, "favorite_count" : 0, "account_age" : "29 Dec 2009 09:06:06 GMT", "user_location" : "Cheezy land.", "geocode" : null }
{ "_id" : ObjectId("5810c187c435f236c8c1db7d"), "user_id" : NumberLong(100190895), "user_name" : "JHPost", "friend_count" : 16500, "follower_count" : 15087, "stat
us_count" : 3204, "favorite_count" : 0, "account_age" : "29 Dec 2009 10:13:35 GMT", "user_location" : "", "geocode" : null }
Type "it" for more
>
```

Here ,user location is either Blank or Invalid like "Toronto — Best City Ever!" or “Cheezy land”

I. Suggestions for resolving those that we were not able to resolve:

- Some of the locations still could be validated in case of Invalid locations:

Example: user.location as “Toronto — Best City Ever” where we need effective string parsing. So we could use string parsing scripts already present or write a new one to get the locations references more accurately so those can be geolocated.

References for scripts can be found at :
http://www.labware.com/limshelp/labstation/LabStation_v1_Users_Guide_10__nbsp__Pars.htm

- As a Twitter is allowing invalid values/Null values ,it become difficult to plot such locations so it could be improved at the input level itself by the application to allow only to put valid locations with City/Country combinations to make the data valid for further use in analysis/plotting or any future effective use of data.

ii. Ways in which this pipeline could be improved, including other tools that could be used.

A pipeline is a set of data processing elements connected in series, where the output of one element is the input of the next one. First I have listed down the pipeline tasks used for project as below.

- **Data pipeline task 1: reformat the data**

The raw txt file of user profiles is encoded in ISO-8859-1 format. This is a format that the MongoDB NoSQL store does not accept so its converted to tsv file and added with data headings for Twitter fields.

```
Command :./bin/reformat.sh <input file> <output file>
```

- **Data pipeline task 2: Import the data into MongoDB**

The tab-separated values (tsv) file is imported directly into MongoDB

```
Command : ./bin/import_mongodb.sh projectA profile tsv user_10000.tsv
```

- **Data Pipeline Task 3: Query and Update the User Profile Collection**

Through the QueryAndUpdate.sh tool we accessed the Google geocoding API to validate user locations and extract valid Latitude/Longitude of the user locations

```
Command :./bin/QueryAndUpdate.sh config/config,properties projectA profileinput/query.json test1.log
```

- **Data Pipeline Step 4: Visualization**

The final step in the data pipeline is used to visualize atleast 50 selected user profiles and their geo-locations using Google Maps. Here we need to run export command and write script/manually convert csv format to Google chart JavaScript lib file support format.

Below are the ways in which pipeline can be improved

- **We should implement a good profanity filter**

  I have exported 85 records from ALL countries where Number of Followers and Number of Friends are greater than 5000 in number to show below criteria.

  "Twitter users which are Active with minimum 5000 followers and minimum 5000 friends all over the world"

  Query: I have plotted valid records out of 85 records. I needed to manually remove few records like below which caused trouble in plotting.

  Location=ÌÏT: 41.975518, 87.900742

  Location=Russia To Brooklyn you bitch

  This type of data need to go from further cleanup process before its been geolocated.

  More references can be found at:

  http://stackoverflow.com/questions/273516/how-do-you-implement-a-good-profanity-filter

➢ We can Automate few processes

Export of records, converting resulting output file to TSV file and converting file format to Geolocate API can be automated for below.

Commands used to export:

Command 1:

```
rykulkar@js-19-114:~/Project/I590-TwitterProjectCode$ mongoexport --db projectA --collection profile  -q ' { $and: [ { geocode : {$ne: null} }, { follower_count:

{ $gt: 5000 } },{ friend_count: {$gt: 5000} }] } '   --csv --fields user_id,user_name,friend_count,follower_count,status_count,favorite_count,account_age,user_location,geocode.formatted_address,geocode.location.lat,geocode.location.lng --out data.csv;

connected to: 127.0.0.1

exported 85 records
```

Command2:

```
cat data.csv | sed 's/,/\t/g' > data.tsv
```

To make the TSV format conversion to format of geolocation API ,we can write a script such that once we feed the CSV file to the code, it can plot for any locations .

- **We can use MongoDB Aggregation Pipeline**

The aggregation pipeline is a framework for data aggregation modeled on the concept of data processing pipelines.

Documents enter a multi-stage pipeline that transforms the documents into aggregated results.

AGGREGATION PIPELINE OPERATORS

Stage Operators

Expression Operators

Accumulators

Reference : https://docs.mongodb.com/manual/core/aggregation-pipeline/?_ga=1.27478078.380707914.1477726147

In this project, results can be aggregated for particular condition for which we plan to plot the graph.

Eg: It can be aggregation on Countrywide /No of followers of users in particular area.

Building Data Pipelines with Python and Luigi

In the early days of a prototype, the data pipeline often looks like this:

```
$ python get_some_data.py

$ python clean_some_data.py

$ python join_other_data.py

$ python do_stuff_with_data.py
```

Luigi is a Python tool for workflow management. It has been developed at Spotify, to help building complex data pipelines of batch jobs.

Some of the useful features of Luigi include:

- Dependency management
- Checkpoints / Failure recovery
- CLI integration / parameterisation
- Dependency Graph visualisation

Reference :https://marcobonzanini.com/2015/10/24/building-data-pipelines-with-python-and-luigi/

Deliverables submitted:

1. ***ProjectA-Records_Plotted.html***
2. ***ProjectA-Records_Plotted.tsv***
3. ***Project Report***