A Project Report on

Twitter User Data Analysis

By: Team Member Name: **Sneha Godbole**

Email id: **snegodbo@umail.iu.edu**

Table of Contents

[1. Problem Statement / Requirements 3](#_Toc438237065)

[1.1 Introduction 3](#_Toc438237066)

[1.2 Use Cases Description 3](#_Toc438237067)

[1.3 Dependencies 4](#_Toc438237068)

[1.4 Assumptions 4](#_Toc438237069)

[2. Goals, Objectives and Deliverables 4](#_Toc438237070)

[2.1 Goals and Objectives 4](#_Toc438237071)

[2.2 Deliverables 4](#_Toc438237072)

[2.2.1 Artifacts 4](#_Toc438237073)

[2.2.2 Documentation 4](#_Toc438237074)

[3. Design 5](#_Toc438237075)

[3.1 Solution Architecture / High Level Design 5](#_Toc438237076)

[3.2 Code Components 5](#_Toc438237077)

[4. Implementation Details 7](#_Toc438237078)

[4.1 Data Acquisition & Ingestion 7](#_Toc438237079)

[4.2 Data Transformation & Analysis 7](#_Toc438237080)

[1.1 Data Visualization 9](#_Toc438237081)

[1.2 Tools and Technologies 9](#_Toc438237082)

[2. User Interface / Data Visualization Screenshots 10](#_Toc438237083)

[2.1 Twitter Users across the World 10](#_Toc438237084)

[2.2 Influential Twitter Users across the World 10](#_Toc438237085)

[2.3 Influential Twitter Users in the USA 12](#_Toc438237086)

[2.4 Most Active Twitter Users across the World 13](#_Toc438237087)

[2.5 Most Active Twitter Users in USA 14](#_Toc438237088)

[2.6 Top 10 Influential Users across the World 15](#_Toc438237089)

[5.User Manual 16](#_Toc438237090)

# Problem Statement / Requirements

## Introduction

Given a Twitter user’s dataset for a finite set of users, we need to pre-process the raw input data, populate additional fields for all users (such as latitude / longitude) based on the available information in the provided fields, and analyze the dataset to get meaningful insights. The insights needs be then plotted on a dashboard using different visualization techniques.

The preprocessing to be done, number of insights to be drawn, and number of different dashboards to be visualized would be dependent on the information available in the dataset along with the implementation timelines.

## Use Cases Description

Following use cases would be implemented for analysis and visualization of dashboards (based on current visibility on the available dataset fields):

1. Process the input data set to populate **latitude** and **longitude** information of all the users, and produce a formatted output (JSON or CSV or TSV).
2. Process the input data set to populate additionalinformation from the latitude, longitude based on the use case.
3. Analyze the formatted input data generated in 1.2.2 for the following meaningful insights
   1. Locate and Plot Twitter users in given dataset across the World.
   2. Locate and Plot Influential Twitter Users in given dataset across the World.
   3. Locate and plot all users in United States
   4. Locate and Plot Influential Twitter Users in given dataset across the United States.
   5. Locate and Plot Most Active Twitter Users in given dataset across the World.
   6. Locate and Plot Most Active Twitter Users in given dataset across the United states.

## Dependencies

The implementation is dependent on given dataset, and the Google Geocode API calls to be made for populating latitude and longitude information for each user.

## Assumptions

* + 1. Twitter users’ Dataset includes but does not limits to following fields:

user\_id, user\_name, user\_location, user\_follower\_count, user\_status\_count, user\_friends\_count and more based on availability.

* + 1. Google GeoCode APIs have a limit of 3000 calls per day, so limited number of users would get processed per day.

# Goals, Objectives and Deliverables

This section describes goals, objectives and deliverables made based on the current information / visibility on the dataset.

## Goals and Objectives

The objective of this project is to collect, ingest, analyze the Twitter user data for getting meaningful insights.

## Deliverables

Following artifacts / reports would be submitted as a part of deliverables of the project –

### Artifacts

1. Twitter Dataset being used.
2. A dataset exported from MongoDB having records populated with latitude and longitude information.
3. A JavaScript/D3.js / HTML 5.0 implementation module for implementing and visualizing dashboards.

### Documentation

A project report describing requirements, problem statement, project scope, design and implementation details would be submitted along with the artifacts.

# Design

## Solution Architecture / High Level Design

The high level design / solution architecture of a solution can be diagrammatically represented as follows –

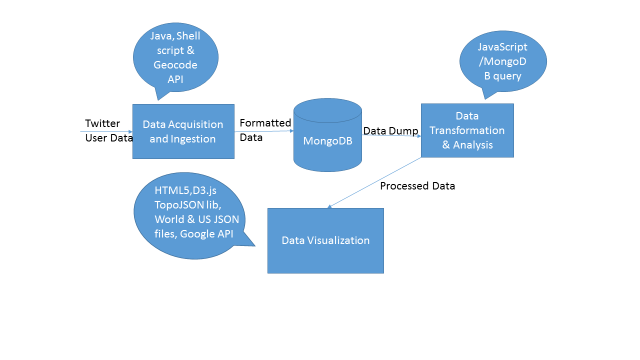


Figure 1: Solution Architecture

## Code Components

Twitter User Data Analysis Project├── Index.html (This is executable file)

├──Useres\_10000.txt (This is Twitter Dataset before formatting)

├──D3.min.js (This library is used to draw all the dashboards)

├──Topojson.min.js (This is extension of GeoJSON which encodes topology)

├── world.json & us.JSON (This files contains topology and other geographic features)

├── Top10Users.js (This file contains array of top 10 influential users worldwide)

# Implementation Details

## Data Acquisition & Ingestion

This step was implemented on 10,000 twitter user dataset, using a utility providing features to import the data into MongoDB, and apply GeoCode APIs to get latitude and longitude based on user location. In this phase, we performed the following steps –

* Preprocessing of Twitter user data
  + I had raw text file of user profiles containing 10,000 records, I had to store this big data into database, as we know MongoDB don’t accept txt file so I converted that file to UTF format using reformat utility. Then I added header line to this formatted data so that mongo DB can understand the data format.
* Ingestion of data into MongoDB
  + In this step I have used shell script to insert this twitter user data into MongoDB.
* Update latitude/longitude using geocode API
  + Twitter user data which I got doesn’t had latitude/longitude so I have used update script to update all the records. So now I had latitude/longitude for most of the records because some of the records because some of records didn’t have address to calculate latitude/long.

## Data Transformation & Analysis

This section describes the transformation and analysis done on the dataset to get meaningful insights.

* **Twitter data users across the world** dashboard is implemented based on latitude longitude information of all the twitter users provided. I exported the data processed in step 4.1 above from MongoDB in JSON format (to be used in HTML 5). The command used to do this is as follows –

mongoexport --db tw\_db --collection tw\_collection --fields geocode.location.lat,geocode.location.lng,user\_name,geocode.formatted\_address,follower\_count ,friend\_count --out processedData.json

* **The Influential users across the world** dashboard is implemented using a heat map in D3.js on the world map, each bubble representing a Twitter user and size of bubble representing the user influence (#follower count).
* **The Influential users across the USA** dashboardrepresents influential users in USA, with same definitions for bubble and bubble size as mentioned above.
* I define a **Most Active User** by computing “user activeness score” for each user as weighted sum of **friend\_count** and **follower\_count,** and then sorting users in ascending order based on this score. The mathematical formula to compute this score for a user is as follows (both friend and follower counts are given equal weights) –

(0.5 \* friend\_count) + (0.5 \* follower\_count)

The **Most Active Users in World** and **Most Active Users in USA** dashboards are implemented by using heat map with each bubble representing a user, and the bubble size representing user activeness.

The latitude / longitude corresponding to all users from the above generated data was later plotted on world map using HTML5 and D3.js visualizations.

* **The top 10 influential Twitter** **users** **in World** are found by sorting all twitter users in *descending* order of their “follower\_count”. This data is exported into a JSON file by using following command –

mongoexport -d tw\_db -c tw\_collection --sort '{follower\_count:-1}' --limit 10 --type=csv --fields geocode.location.lat,geocode.location.lng,user\_name,geocode.formatted\_address,follower\_count --out sorted.csv

This dashboard is implemented by using a bar chart in D3.js and HTML5.

* The data required for USA specific dashboards is captured by implementing a location filter on the entire dataset, and then processing it with the appropriate algorithm (influential users OR active users).

## Data Visualization

* The dashboard visualization is implemented in HTML and D3.js (at high level, detailed tools and technology stack is described in section below)
* A web server for deploying HTML and D3.js dashboards would be done in Python. The command for start web server is

python -m SimpleHTTPServer 8000

## Tools and Technologies

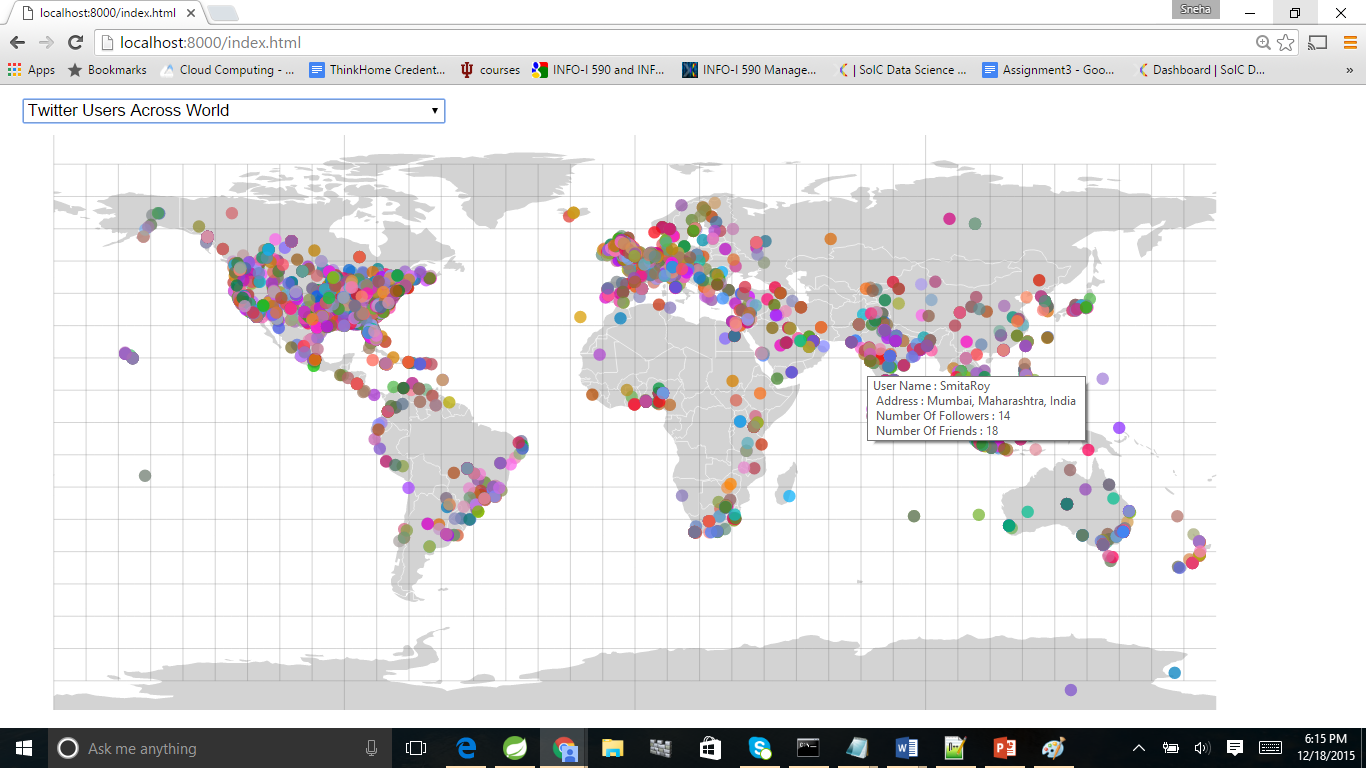
Following tools and technologies were used for implementation of the solution:

1. Processing and Analyzing Dataset
   1. Core Java
   2. Shell Script
   3. MongoDB
   4. Google Geocode APIs
2. Creating and Visualizing Dashboards
   1. Python
   2. D3.js
   3. HTML 5.0
   4. CSS
   5. Topojson.js
   6. Jquery.js
3. Documentation
   1. Microsoft Office

# User Interface / Data Visualization Screenshots

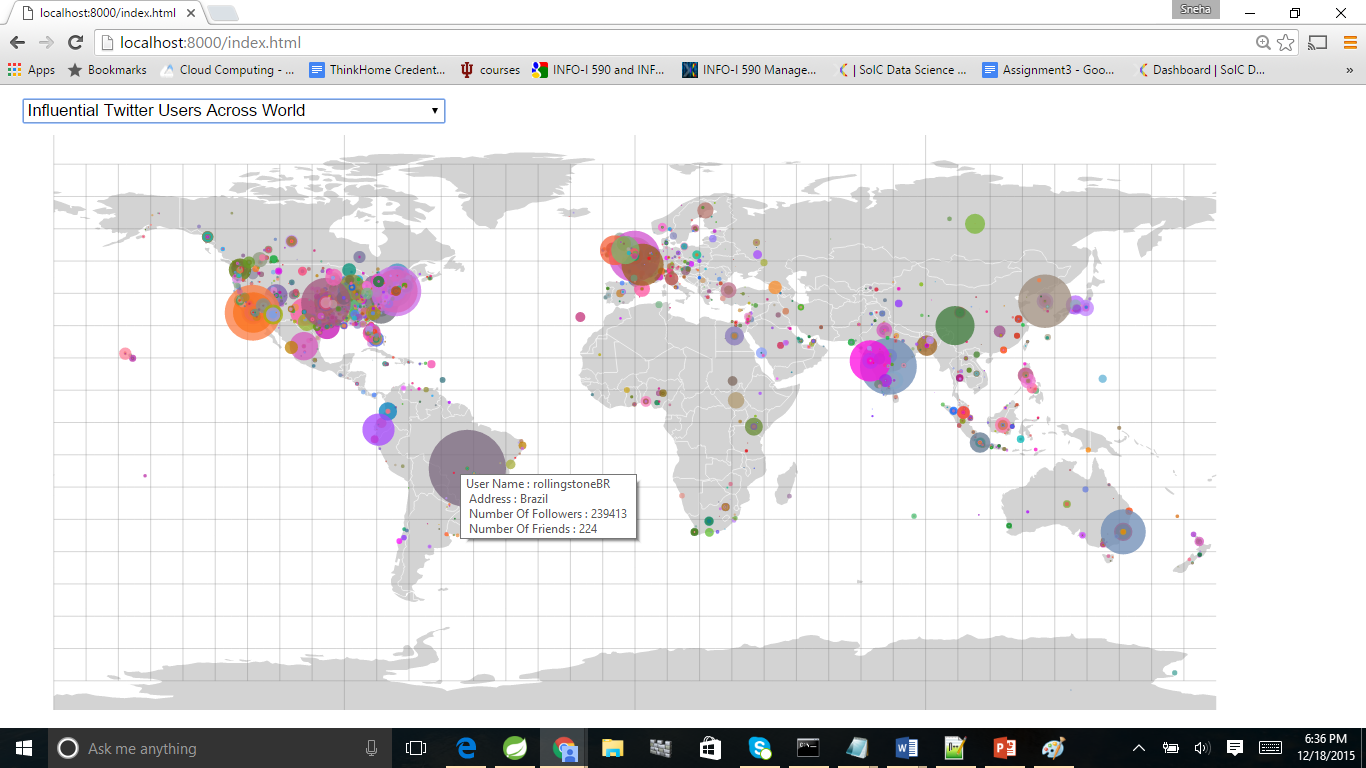
## Twitter Users across the World

In this dashboard I have displayed all the twitter users across the world using circles with different colors. If you mouse over on particular circle it will give you information related to that user.



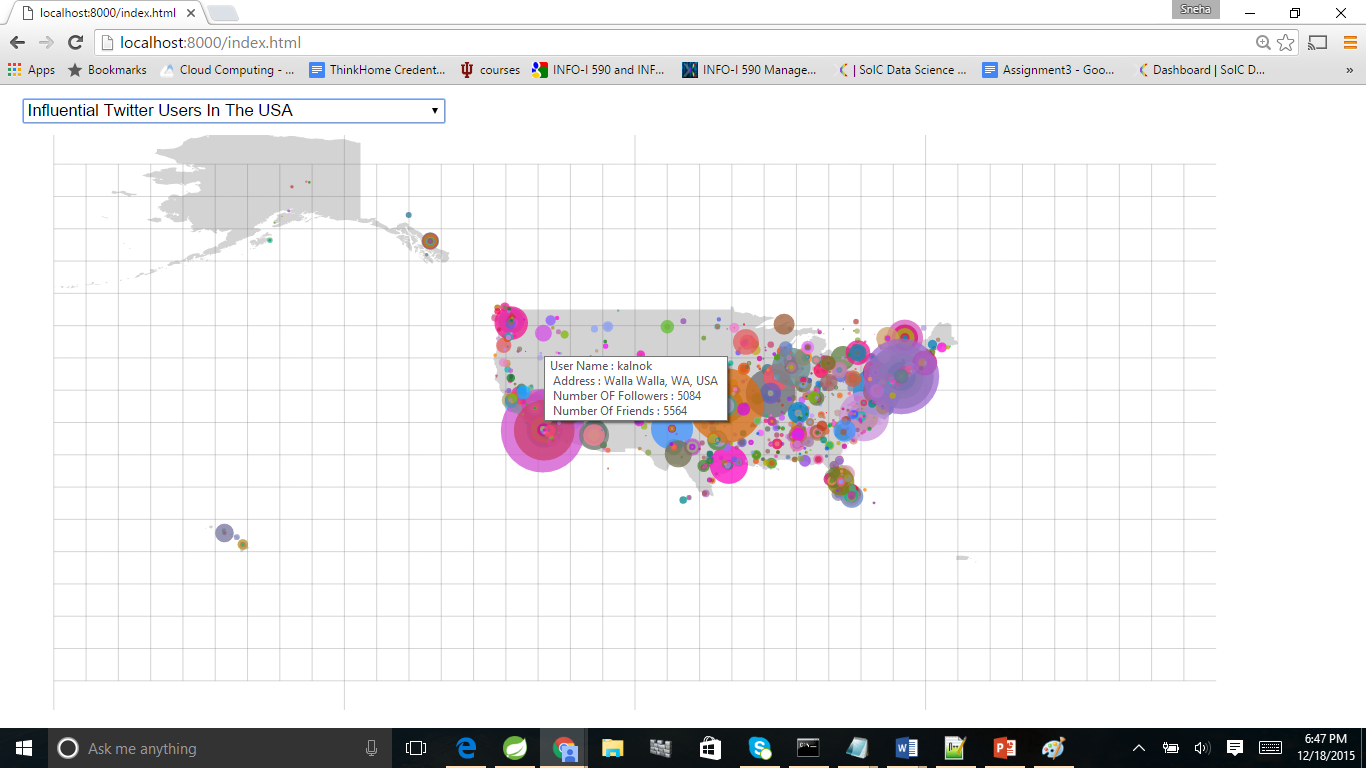
## Influential Twitter Users across the World

In this dashboard I have displayed Influential (users having lot of followers) the twitter users across the world using circles with different colors. So here Larger circle means user is most influential. If you mouse over on particular circle it will give you information related to that user.



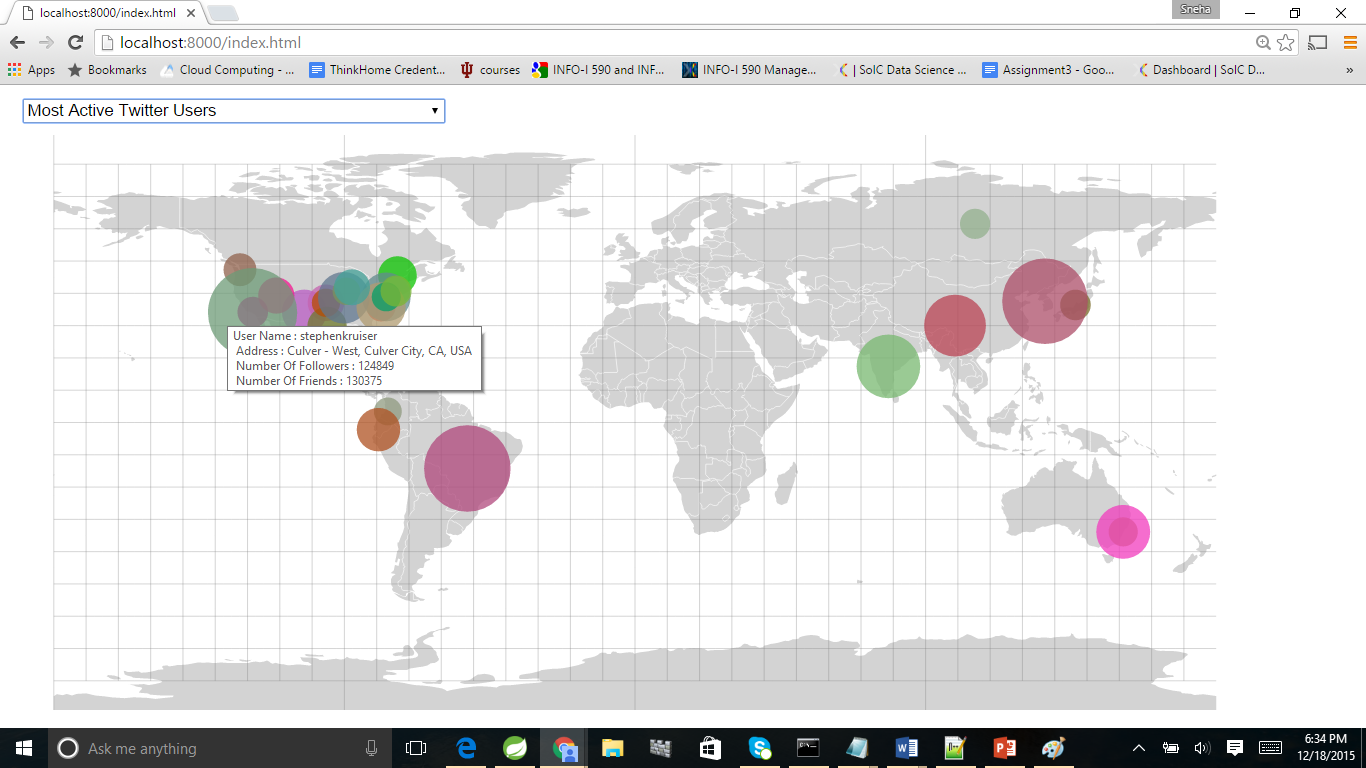
## Influential Twitter Users in the USA

In this dashboard I have displayed Influential (users having lot of followers) the twitter users in USA using circles with different colors. So here Larger circle means user is most influential. If you mouse over on particular circle it will give you information related to that user.



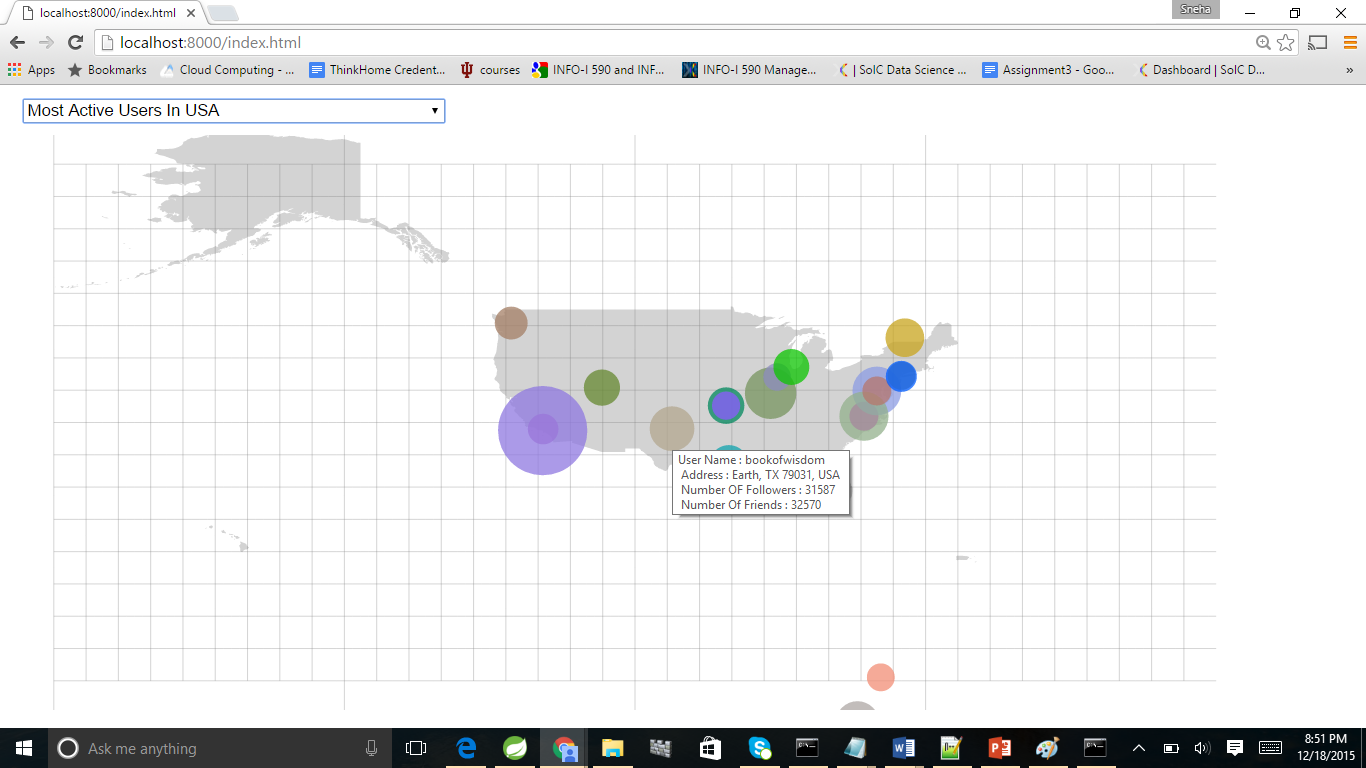
## Most Active Twitter Users across the World

In this dashboard I have displayed Most Active (users having lot of friends and followers) the twitter users across the world using circles with different colors. So here Larger circle means user is most influential. If you mouse over on particular circle it will give you information related to that user.



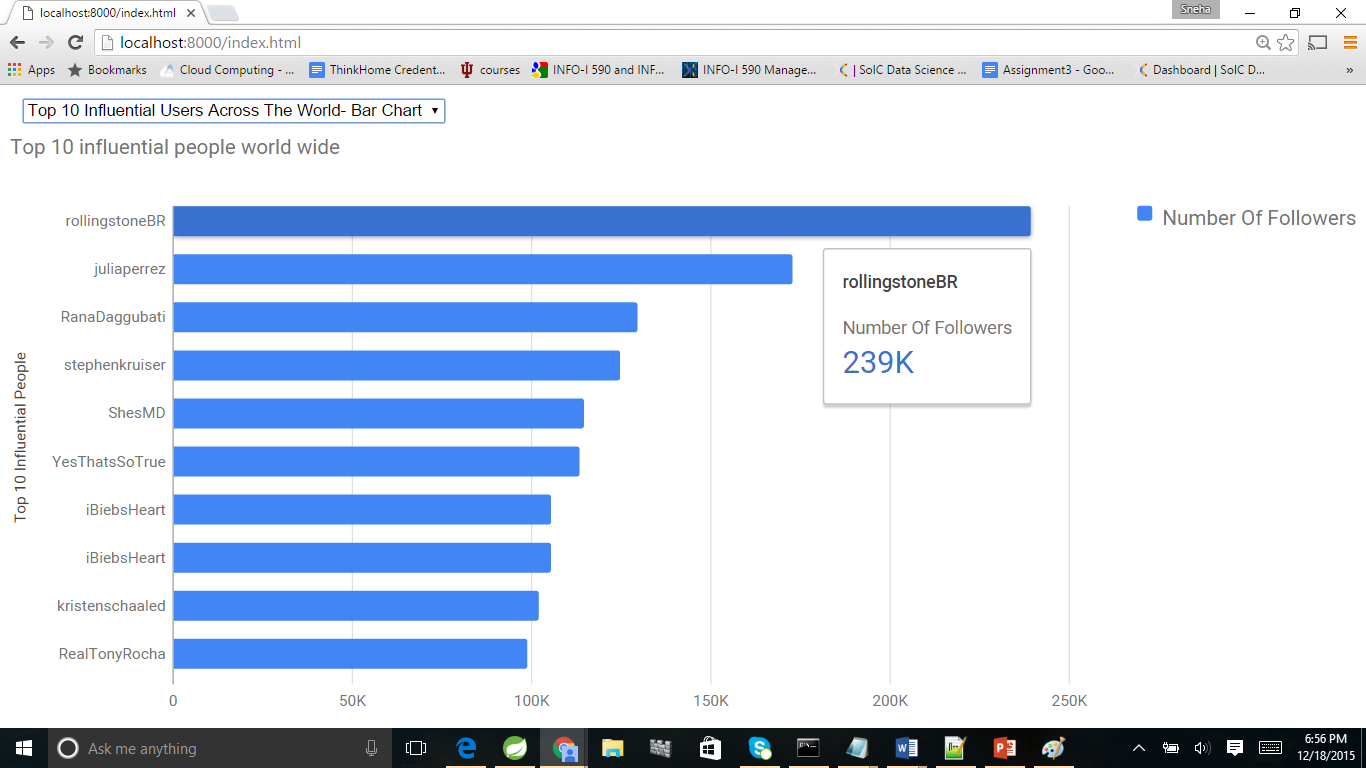
## Most Active Twitter Users in USA

In this dashboard I have displayed Most Active (users having lot of friends and followers) the twitter users in USA using circles with different colors. So here Larger circle means user is most influential. If you mouse over on particular circle it will give you information related to that user.



## Top 10 Influential Users across the World

In this dashboard I have displayed Top 10 the twitter users across the world using Bar chart. If you mouse over on particular circle it will give you information related to that user.



# 5.User Manual

Running the project is quite simple as I have kept all the dependent files in same folder. Follow the steps given below

* To run the project first, unzip project and open the project folder.
* Now open command prompt and go to project folder path

e.g

C:\IU>cd Project\_BDAA

When you are in project folder now start web server by using following command

C:\IU\Project\_BDAA>**python -m SimpleHTTPServer 8000**

Serving HTTP on 0.0.0.0 port 8000 ...

* Web server is up and running, so open any web browser and type in

**http://localhost:8000/index.html** and hit enter. You can see different dashboards by **using drop down**.