**Data Science Challenge Report**

Course: DSCI 551 – Foundations of Data Management

Group members:

Rushi Shah

Dhruvil Anil Trivedi

**Task Description:**

The main tasks for the data science challenge provided by Group 35 – Shree Shanmukhaa Gorla was to structure the json dataset for the into proper table and perform basic analysis such as source based and language-based analysis.

**Solution:**

For the Data Science challenge, we were assigned the task created by Shree Shanmukhaa Gorla (Team 35). The task was to clean, transform and perform some basic analysis in the twitter dataset. The dataset provided was in json format and was very raw twitter data. Hence, the main focus was upon the transformation and cleaning rather than just the analytics part. The dataset was a set of json objects containing the data. The main key was data and thus the data within those keys was the actual. That data was entered into a data frame to visualize the data efficiently. The null values and subsets of dictionaries of certain entities were removed initially as they were not very consistent. As the data was very less (676 rows), there wasn’t very high analytics as compared to the preprocessing needed. Hence, we plotted a few graphs for analyzing which language and sources is used the most.