**CS5540**

**PRINCIPLES OF BIG DATA MANAGEMENT**

**INTERIM PROJECT REPORT**

**Team Member:**

**Anh Tuan Nguyen (18142335)**

**Mohit Sriram Tirumala (16292490)**

**Sai Haneesh Tanneru (16290739)**

**OBJECTIVE:**

* Collect tweets.
* Write at least 10 analytic queries to explore and understand the collected tweets.
* Create Visualisations on the results.

**QUERIES:**

* Query to get the top 10 sources.
* Query to get the 10 most popular Hashtags.
* Query to get the 10 most mentioned accounts.
* Query to get number of Twitter users created in years.
* Query to get the 5 most used languages.
* Query to get the top 10 users with the most firends.
* Query to get the top 10 users with the most followers.
* Query to get number of tweets and users collected.

**VISUALISATIONS:**

* Create visualisations (Table, Bar chart) of the above queries using Apache Zeppelin.

**SOFTWARES PLANNING TO USE:**

* Apache Spark
  + Spark Streaming to collect tweets.
  + SparkSQL to store and execute queries.
* Apache Zeppelin to execute code and create visualisations.

**CODE TESTING:**

* We can test the code by passing in the actual collected tweets to the program.

**CONTRIBUTIONS:**

* Anh Tuan Nguyen – Collect data, the first 4 queries and visualization.
* Mohit Sriram Tirumala - Collect data, 5th and 6th queries and visualization.
* Sai Haneesh Tanneru - Collect data, remaining queries and visualization.