**CSC​ ​591​ ​-​ ​Data​ ​Intensive​ ​Computing**

**Title:** ​ HashTag Recommendation System for Twitter

**Team:**

**Name Student​ ​ID Unity​ ​ID**

Aneesh Gupta 200111508 agupta27

Chinmoy Baruah 200153979 cbaruah

Sanket Shahane 200153333 svshahan

**Overview:**

HashTags are cool to use and a trend in today’s communication. Use of proper hashTags help in popularizing your content in the right channels, and to the people of similar interests. However, many times, users are not aware of the popular hashtags relevant to their content. Our system will analyze a corpus of recent tweets and filter the most popular hashtags and the tweets corresponding to them. Finally, we will recommend popular hashtags to the users based on content similarity. Since, we will analyze fresh tweets at a fixed interval our recommendations will not only be relevant but also trendy.

**Claims:**

**1. Data​ ​ingestion​ ​layer:** ​

​The pipeline will fetch 50GB archived twitter monthly streams through Twitter APIs and archive data stores. Preprocessing will be done while ingesting the data to bring the data in a format usable in the project.

**2. Processing:**

Our models will be updated with new trends and new data on a regular basis. We need these updates and aggregations within reasonable amount of time. Thus, we will leverage distributed processing using Spark cluster (inherently Map-Reduce).

a. Aggregation Layer: Spark cluster jobs will pull data from the data store and perform aggregations on the hashtag usage frequency to analyze the trending topics in a distributed fashion.

i. Number of mentions of the hashtag at different time granularity weekly, biweekly, monthly etc.

ii. Get the frequency distribution of the hashtags and filter the top trending ones along with the tweets into Cassandra database. This will be our training data.

b. KNN matching of the the user text to the samples in training data will extract relevant and popular hashtags.

**3. Database​ ​Layer:**

Our system will be scalable and scalability is necessary since we might experience sudden spikes in tweets.We will be using Cassandra database which is a column based NoSql database and is highly scalable and also supports SQL.

**4. Presentation​ ​Layer:** ​

Our hashtag recommendations will be given to the user at runtime in a timely manner. ​In order to present the results in the form of dashboard, we will use FLASK and MEAN Stack.