**Checkpoint 1:**

Analysis of Twitter Data for Recommendation of Universities.

|  |  |  |
| --- | --- | --- |
| DP,Namratha,  Masters in Computer Science,  UAlbany,  Email:ndoddabadigereprasad@albany.edu | Manjunath,Srividhya  Masters in Computer Science,  UAlbany,  Email: smanjunath@albany.edu | Shah,Harnish  Masters in Computer Science,  UAlbany,  Email: hshah2@albany.edu |

**Under the guidance of:**

**Dr.** **Feng Chen**

Contents

[1. Project Title 3](#_Toc511853893)

[2. The names of all team members 3](#_Toc511853894)

[3. Abstract 3](#_Toc511853895)

[4. Implementation 3](#_Toc511853896)

[5. Challenges encountered. 5](#_Toc511853897)

[6. Next Checkpoint work that can be implemented. 5](#_Toc511853898)

# 1. Project Title

Analysis of Twitter Data for Recommendation of Universities.

# 2. The names of all team members

DP,Namratha

Manjunath,Srividhya

Shah,Harnish

# 3. Abstract

The system that is proposed is a recommendation system of university which recommends a student by collecting tweets and retrieving them. Once the tweets are retrieved, it is analyzed and predicted from the Actual live twitter in accordance to the all tweets including the advertisements.

We are planning to use the classifier model, prediction scheme and logistical regressive techniques to predict for the purpose to be met. Once its predicted using these different techniques, we shall compare as which suits the best for the prediction.

# 4. Implementation

For implementation, we have chosen a step wise approach. The steps we are following are described below.

Step 1. Data Collection

We have **already implemented** this step. This is a very crucial step, as every data or any prediction will be totally dependent on this step. We collected the data using several know keywords like “#University #rank #Computer Science #BestUniversity #studies #academics” and in the second set we collected using more related keywords like “#Admit #rejected #Games #InterUniversitySports #excited #Allset “etc. These set of keywords and query has returned us the major Dataset.

Step 2. Data Preprocessing.

The tweets from another language other than English were first removed. All the tweets were particularly saved in a format removing the special characters, colons and all other semi-colons. we also used a lot of stop words during this procedure like above, after, next, before and output and such words.

The following steps are still in progress.

Step 3. Clustering

Clustering is the process of making a group of abstract objects into classes of similar objects.

* A cluster of data objects can be treated as one group.
* While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
* The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Step 4. Classification

Following are the examples of cases where the data analysis task is Classification −

* A bank loan officer wants to analyze the data to know which customer (loan applicant) are risky or which are safe.
* A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.

Step 5. Application of SVM

The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high dimensional feature spaces.

Step 6. Implementation of Logistic Regression Model

The logistic distribution constrains the estimated probabilities to lie between 0 and 1. The regression coefficients measure the predictive capability of the independent variables. MLE is an iterative algorithm which starts with initial arbitrary numbers of what the coefficients should be

Step 7. Prediction of the outcomes.

# 5. Challenges encountered.

* While retrieving the tweets, it was hard to select the write query to make the Dataset the best fit for the system to work correctly.
* Also understanding the techniques and algorithm, and to understand its necessity.
* While implementing the SVM and linear regression, properly identifying the required parameters and the data to it.
* Also identifying the required keywords and resolving the errors like 🡪 “invalid value encountered in less x = np.where(x < 1.0, x, 1.0) # if x > 1 then return 1.0”

# 6. Next Checkpoint work that can be implemented.

Since we have already understood the basic algorithm, we have implemented the SVM and Logistic Regression Model and you can expect the step 7 to be completed by the next checkpoint. By this we can compare the 2 models for the Recommendation.