Classifying Buffalo 311 Data

Submitted by Madhur Gupta, Mohit Arora

December 10, 2015

Problem Statement

Buffalo 311 is a service where residents call this service and register their complaints. These complains can be from any area in the city of Buffalo and can be of any type e.g. broken curb in front of an address, damaged street light, problems with the sewer etc. In this project we are categorizing and classifying Buffalo 311 data to identify pattern of the complaints in different areas of Buffalo. Based on this classification, it can be easily inferred about the prevalence of a particular type of complaint in particular area which can be taken care by Buffalo authorities by taking corrective measures in advance so as to reduce the future complaints.

Data

The City of Buffalo 311 Call and Resolution Center provides citizens with fast, centralized access to city services which can be of type, register complaints, get information, and access non-emergency police services. When appropriate, a Reference Number is assigned to the request so that its progress can be tracked. Data we got consists of complaint description and queue to which this complaint is categorized. Complaint description is a minimal English text of the verbal conversation between the caller and the operator. Queue is the category in which that particular complaint falls. This queue has predefined categories.

Methodology

Complaint description provided, was first cleaned to remove UTF-16 characters. Then each queue for the complaint category was given a unique integer value for the final prediction. Then each complaint that is taken as a text as an input was tockenized and various stop words were removed from this tocken stream as well as it was stemmed by using "NLTK" library of python. This cleaned and tockenized data is converted into a feature vector which will be fed into prediction algorithm. For getting the data as a feature vector, each word is hashed and a particular value(integer) is assigned to each unique word in the text so as to produce feature vectors of integers. Data is then divided into training and testing data sets as 70% for training and 30% testing. Using this training data, using Naive Bayes classifier, we built a model for prediction. Then this model was used to predict the testing data set values to find out the

accuracy. For this case, accuracy is defined as the number of correct queues model was able to predict for the testing data set. The maximum accuracy we achieved is 82.7% when feature vector size is kept at 5000.

Results

We tested the data based on various feature vector sizes to see how it affected the overall accuracy achievable by this model. We found that the accuracy initially increases with the feature vector size but after a certain value it starts decreasing, because of the following reasons:-

1. Complaint description text is small, so there are not many different words in the text.
2. As we increase the size of the feature vector, number of fill-in values(0,1) increases leading to redundant data in the vectors.

Following graph was plotted to see the variation of accuracy with the feature vector size :-