Semester Long Project:

Analysis of Users from Yelp Dataset

Ahmed Ulde: 1001090889

Rohan Patil: 1001052795

Description: Importance of Yelp Dataset is that they don’t import users from other social sites. Yelp has their own social network. They also have lot of business information making it worth a study. Primary idea of our project was to use dataset from Yelp Dataset challenge and try out various analysis to find out any possible good results worth using in real life. We went through following extensive procedures

* Extracting required data from Yelp dataset
* Grouping the users into clusters and visualization of these clusters via a web interface
* Prediction of User rating

Challenge:

1. Data cleaning and extraction: Earlier in the initial stages of project we tried to make small sets of data for ease of testing our program as Yelp dataset is huge. As a first step we cleaned data files but it resulted into removal of preexisting markers in dataset so that it could be used with json parser but we failed. Finally we read the data line by line and used dictionary in python to extract relevant data.
2. We created web interface omega.uta.edu/~aau08889/D3/post.php where you can upload the extracted dataset and mention the number of clusters and see the visualizations created using D3. But due to large file size upload functionality didn’t work and we had to use only visualizations of 3 clusters. <http://omega.uta.edu/~aau0889/D3/test.html> this link is available to use but loads slow.
3. Possible graphs from yelp dataset are the social graph and co-rating graph. It is also possible to create a entire network of users who rate the business and in return they get rating for their review on a business. This was observed in a previous yelp challenge winner paper who suggested a mathematical model for the same. After trying out various things with this dataset we had decided to use this similar network and use HITS algorithm on MRQL platform to rank business based on the yelp social network. But this task turned out to be too ambitious and couldn’t be completed.
4. We have implemented a naïve bayes classifier to predict the user rating but we have discovered that user rating involves some form of intuition so it varies from the classifier.

Method:

1. Data Extraction:

As we faced challenge while data cleaning , we followed line by line extraction of required data from dataset. User related data.

1. Clustering using Weka: The Weka tool provides a collection of machine learning and data mining tools. Freely available built upon Java which allows it to run on platforms that support Java. It’s maintained and supported primarily by researchers at the University of Waikato.

The dataset was modified to create nominal columns from some of the numeric columns in order to facilitate usage in Weka for Tree analysis and simple cluster analysis. The table is then converted and saved into the Weka Attribute-Relation File Format (ARFF)

For our cluster analysis, we chose the Simple K Means, just for simplicity. Using the cluster diagram we can visually analyze the clusters for relationships within the dataset.

WEKA QUERIES:

java -cp weka.jar weka.core.converters.CSVLoader uploads/user.csv > uploaded.arff

java -cp weka.jar weka.filters.unsupervised.attribute.AddCluster -i uploaded.arff -o output.arff -W "weka.clusterers.SimpleKMeans -N 3"

java -cp weka.jar weka.core.converters.CSVSaver -i output.arff -o output.csv

1. D3 Visualizations: The output.csv from above is used to generate a scatter plot. The javascript code extracts the header of the file and on the web interface u can select the X and Y axis and code will generate the plot for you.
2. Weka visualizations: Also this visualization and clustering was done using coding method with weka.jar.
3. We Have extracted user.csv file and made a naïve based classifier to predict the user rating given a tuple. The dataset is read and summarized using pandas and then the test set is evaluated for prediction of user rating.

Technologies used:

PHP, javascript, D3, weka, python, java, Apache MRQL

Future work:

Continue research with this social graph and research some possible analysis on this social graph and if we get any good finding we will try and submit to yelp by June 30.

References:

[1] http://www.yelp.com/html/pdf/YelpDatasetChallengeWinner\_NetworkEfficiency.pdf