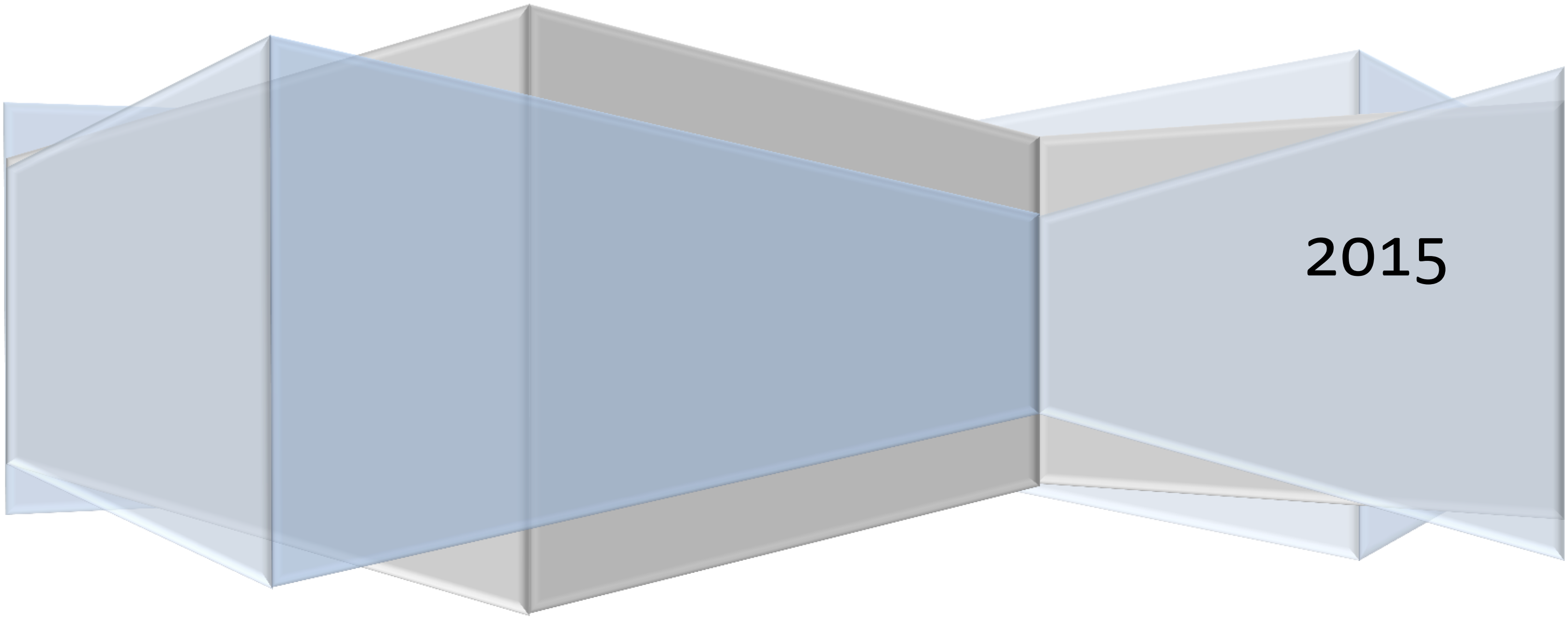
CSE 5334 Course Project

Yelp data – Project Progress

Puneeth Umesh Bharadwaj

Shreekanth Kuppur Shreedhar



2015

Contents

Proposal:

We see that most of the reviewers visit and provide reviews to the restaurants that are in their vicinity as they have knowledge of restaurants in and around there are or hometown. We provide user with suggestions in other locations (cities, neighborhoods) based on his/her reviews, ratings, check-in and cuisines. This would help a user when he is out of town or in other location and wants to know the best restaurant of his favorite cuisine in that locality.

Project Progress:

Based on the project proposal, we first tried to analyze the business, checking and user data their data format and ease of use of data. We found that data is in Json format, i.e. in key value pairs, where each attribute name was the key and value was the attribute value. This was repeated for all the data set. Though it is easy to work with this format of data in python where we can create dictionaries with key value pairs as given in files, we found that we needed the data to be in csv format for us to work on the tools we were planning to use.

Challenges Faced:

As mentioned before the challenge we faced was to convert the data in json format to csv. We did find few online tools which would convert to the required file format, but they had a restriction on file size to be less than 1MB as the conversion was done in the browser, hence it didn't bear us any fruit. Upon a little bit of searching we found that yelp had sample code to do the conversion on their Git hub Repository. We used the code to convert the files to csv format.

Approaches tried:

We tried to work on tools Weka version 3.6 and Rapid Miner Studio, and we found that Rapid miner though having more features required more time to learn and took more steps as it required a workflow to be created which took more time to understand. On the other hand Weka was easier and straight forward to learn and use in mining tasks known to us. Hence we decided to go with Weka.

We tried to cluster the data and see the results on which we could work further. Clustering the user reviews based on the user ID to find his food interests is what was expected. We also ran clustering on Checkin data which should ideally provide all the checkins in one business. This we could compare with the user data and find his checkins was the plan of action.

Results:

The results of clustering in Weka was not satisfactory, the approach did not return the result as expected. And we would not be able to use the result obtained further. As Weka being a new tool we are not aware of its full functionality as we were not sure about the selection of criteria based on which clustering to be done.

Next Steps:

We will study the tool further and see if we can change the criteria of clustering. We would work towards finding the required result so that we could predict restaurants of user's interest in different area based on his previous reviews. Also try and implement a classification method, preferably Naive Bayes. . This will be an initial approach to find out the best possible solution for our presented problem.

References:

http://www.yelp.com/dataset_challenge

http://engineeringblog.yelp.com/2014/08/the-yelp-dataset-challenge-goes-international-new-data-new-cities-open-to-students-worldwide.html

https://github.com/Yelp/dataset-examples

http://engineeringblog.yelp.com/2014/11/yelp-dataset-challenge-round-3-winners-and-dataset-tools-for-round-4.html

http://www.cs.waikato.ac.nz/ml/weka/