**YELP DATASET CHALLENGE USING OPTIMIZED FRAMEWORK FOR MAP REDUCE QUERY LANGUAGE:**

**CONTRIBUTORS:**

Rohan Patil (1001052795) Ahmed Ulde (1001090889)

**OBJECTIVE AND OVERVIEW:**

Our objective is to find meaningful information from yelp dataset. That information must be useful for new startup companies and business analysis of existing companies. Also from reviews and user profiles we should be able to find out customer preferences and choice.

Our project will focus on finding such data with combination of tools and techniques. We will use json dataset along with apache MRQL or SPARK and write queries to extract info. Also we have CSV file for dataset to use tools like weka and scikit learn to apply some machine learning algorithms on dataset and create some visual graphs and plots to answer general queries.

**DATA MINING TASKS**

**TASK 1:** Writing scripts to extract unique categories from business.csv and grouping of data based on category, reviews grouped by user id, group of business based on their working hours or location and so on to get some clear data.

**TASK 2:** Visualize business attributes vs locations and other relevant attributes. This will help to visualize the data and make decisions which area is potential for which business.

**TASK 3:** Identify good yelper profiles from user.json and filter the tips.json to keep good tips remove bad ones as they don’t have likes or any trust worthy info. Basically this task would focus on extracting valuable reviews and tips based on user profiles.

Written a script (att.py present in dataset-CSV folder ) to open and read file for finding the attributes user and review files. Below are the findings.

Below are the attributes of user.json file

['yelping\_since', 'compliments.plain', 'review\_count', 'friends', 'compliments.cute', 'compliments.writer', 'fans', 'compliments.note', 'type', 'compliments.hot', 'compliments.cool', 'compliments.profile', 'average\_stars', 'compliments.more', 'elite', 'name', 'user\_id', 'votes.cool', 'compliments.list', 'votes.funny', 'compliments.photos', 'compliments.funny', 'votes.useful']

Below are the attributes of review.json file

['user\_id', 'review\_id', 'text', 'votes.cool', 'business\_id', 'votes.funny', 'stars', 'date', 'type', 'votes.useful']

**TASK 4:** Build a model to support queries based on business attributes. Sample would be like how likely is a business to be successful if you don’t accept cards but only cash.

**TOOLS AND TECHNIQUES:**

ScikitLearn, Weka, Orange, Rapid Miner, Anaconda, NLTK: These tools will be used initially for small queries, visualizations ad understanding the dataset.

Main approach would be using an optimized framework for map reduce query language (mostly MRQL) or Apache Spark(If better performance required) on hadoop cluster.

**CHALLENGES AND OUR APPROACH:**

Dataset is too large covering a lot of information in different domains. Challenge is choose which quries we would like to solve. Very difficult challenge would be to have one interface and user should be able to query any info from the dataset which is difficult to achieve and may be beyond the scope or time of course. So we will focus a few queries initially and increase the scope of the project as the time permits. Hence we have divided our projects into small achievable tasks. Once we achieve the mentioned tasks we may work more towards enhancement of project.

**DELIVERABLES:**

Deliverables would be source codes and queries and a document specifying which code or query was used for what part of the project. Website will include a background of tools and techniques used along with the content of document written.

**Work Division**

Ahmed: work on MRQL, version controlling

Rohan: work with data mining tools, website

We will work on each other’s tasks too so that both know about everything we are doing.

**REFERENCES:**

<http://www.stat.ncsu.edu/people/zhou/courses/st810/hw/yelpProject.pdf>

<http://www.ics.uci.edu/~vpsaini/>

<http://java.dzone.com/articles/yelp-graph-checkin-based>

<http://wiki.apache.org/mrql/>