**YELP DATASET CHALLENGE:**

**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 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:**

Some machine learning or classifier algorithms along with visualization techniques would be used initially for some parts like prediction of success for new business.

Another approach would be using an optimized framework for map reduce query language on hadoop cluster.

**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>