**Extracting Sub-categories using Textual Analysis and Classification**

**Team Information:**

Pavan Chowdari Ghanta(1001118625)

[Pavan.ghanta@mavs.uta.edu](mailto:Pavan.ghanta@mavs.uta.edu)

**Objective and Overview:**

Yelp dataset contains ratings and reviews about various businesses and services. Yelp data (reviews and ratings) helps the users to choose from the list of available the best service or business. The reviews contains important details about the businesses. The users searching for a business of specific type have to go through all the reviews to evaluate the quality of the business. Sometimes, the search for services also depends on the time and situation. For ex., the users who are looking for a specific type of restaurant (calm environment, good ambience, wifi available) may have to read all the reviews to find the required information. Users may search for restaurants that are open till late night that can home deliver food. Some users may further need to know the amount of time they need to deliver the food. This kind of information is not directly available but can be extracted from the reviews. The businesses or services needs to be sub-categorized and are presented to the users so that they can find the appropriate information quickly. The sub-categories includes businesses with less check out time, parking availabilities, restaurants with quality service, coffee shops with peaceful environment and wifi availability, restaurants suitable for game day e.t.c., The businesses can be sub-categorized using already available category information and performing texual analysis of the reviews. This sub-categorization is useful in making the data more transparent to the users.

**Data mining tasks:**

Words are extracted from reviews text and analyzed. Data classification is performed by constructing a decision tree and various businesses or services are further sub-classified based on different attributes.

**Deliverables:**

Deliverable will be a program that takes the dataset and analyses the data and classifies the businesses based on the attributes. When users are searching for a kind of business, they are presented with various sub-categories. User can search for anything and can further filter it based on the sub-category. Directly selecting a sub-category lists all the highly rated businesses or services in that sub-category.

**Challenges:**

The data set is large in size. Since the data is about various businesses and services, many attributes need to be considered. Constructing a decision tree will be a complex task.

**Efficacy:**

Analyzing multiple reviews of a business and estimate the sub-categories manually. Calculate the sub-categories of the business using the program. The efficacy of the program can be evaluated by comparing the sub-categories calculated manually and returned by the program.

**Note:**

**References:**

**http://www.ics.uci.edu/~vpsaini/**