**Extracting Sub-categories using Textual Analysis of Reviews**

**Team Information:**

Pavan Chowdari Ghanta(1001118625)

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

**Objective:**

Perform textual analysis of Yelp reviews and extract sub-topics. Using sub-topics, categorize the business. As part of this project, I will be a sub-categorizing restaurant. Sample of sub-categories expected to generate are Good for group hang out, Coffee shops to work, Good for game night, etc. The sub-topics are presented to the user in the user interface. When a category is selected, relevant restaurants are displayed in the order of their ratings.

**Tasks:**

I will be performing Topic modeling on the reviews to extract meaningful topics All the reviews related to a restaurant are combined and treated as a single document. Reviews belonging to all the restaurants will be a corpus of documents. Data pre-processing is performed on text in each document. The entire text is converted to lower case. The punctuations and special characters in the document are removed. The Stop words from the text are removed. The frequency of each word in a document is calculated. Based on the word frequency, a sub-category is generated.

**Deliverables:**

* A program, which takes the dataset as input and performs textual analysis of the reviews of the restaurants. The outputs of the analysis are list of categories with each category mapped to at least one restaurant.
* Simple website, where user can select a category and all the restaurants related to the category are displayed.
* The restaurants are displayed in the order of their ratings.

**Challenges of Project:**

* The dataset is huge and no. of reviews are high, which is good for improving the accuracy of the analysis. But requires lot of computing to run the program. Data processing is taking lot of time to run.
* The no. Of categories extracted may be high in number. Need to narrow down them to only meaningful categories, which can be of useful and relevant to user requirements.

**Methods:**

* Methods for reading and pre-processing the data are completed.
* Currently, writing method for calculating the frequency of terms in the document and in the corpus.

**Evaluation Plan:**

Run the program on the entire data set and document the sub-categories generated for the restaurants. Pick any N restaurants randomly and estimate the sub-categories of the restaurants. Compare the sub-categories generated by program and estimated manually

**Change of plan:**

During the project proposal, thought of implementing a decision tree for sub-categorization. But decision tree classifier doesn’t fit in this scenario. Decided to perform topic modeling on the corpus of reviews.

**Difficulties encountered so far:**

* The implementation of the project is going smooth so far.
* Currently analyzing the model Latent Dirichlet allocation (LDA) to get more insights into topic modeling.

**Tasks to be accomplished:**

* Extracting sub-categories based on frequency of the words in the review.
* Taking a sub-category and determining the restaurants those are related to the sub-category and group them together in the order of their ratings.
* Create a website to display the sub-categories outputted by the program.

**Expected Challenges**

* Narrowing down the sub-categories to few meaningful sub-categories is a challenging task.
* Given the timeline, creating an appealing and less error-prone website to display the results of the program