**CIND820 Big Data Analytics Project**

**Abstract**

Yingqi Hu

As one of the most popular platforms which aim to connect people with local businesses, Yelp highly relies on it review and rating system. Yelp has its own recommendation software to automatically valuate all reviews daily. Those are considered as maybe useful and reliable from active users will be presented as Yelp recommended reviews. Review plays an important part in making recommendation and promoting business.

Yelp’s dataset provides business with both text data (reviews) and non-test data (other attributes like location). As a business management team, how should they allocate the budget/time in analysing the data to get a better prediction on rating? Should they focus on only text data analysis, only non-text data analysis or both?

This project is trying to help business answer these questions through text classification and sentiment analysis.

We will try to identify the most common features that drive the rating and the most common features that lower the rating. This would help business to improve their services and make business decisions. Furthermore, we will try to find a better model for business to predict rating just based on reviews. This could help business to predict the impact on the rating from their reviews.

The data used in this project is the Yelp Open Dataset (<http://www.yelp.com/dataset>). The main techniques include sentiment analysis (n-grams) and classification (Naïve Bayes, Support Vector Machines and Multinational Regression). Python and related packages (e.g. NLTK) will be used for solving the problem.