**ST: Parallel Analysis Final Project Proposal**

Presenter:

Kuan-Chih Lee

Sai Sneha Yachamaneni

Instructor: Prof. Stratis Ioannidis

**Topic:**

Restaurant Recommendation with Yelp dataset

**Dataset Description:**

Source: Yelp dataset <https://www.yelp.com/dataset/download>

Background: Public provide their reviews about the service provided by

businesses from their personal experience. Also, Yelp collects this precious

experience and open it to researchers.

Description: As part of Yelp data challenge, Yelp has provided data which consists

of various JSON files.

Size: Primary file is review.json which consists of 6 million records (4.5GB)

Purpose: This task is to build a restaurant recommendation system to new

customers. By performing sentiment analysis on reviews, we can help Yelp by

providing suggestions to customers with better experience exploring pocket list of

restaurants.

**To do list:**

1. Loading data with Spark and rename columns

Sample:

{"review\_id":"VKGvSe0SmYcFc5PqjbxLDQ","user\_id":"eG6HneK9zLcuZpVuKcsCGQ","business\_id":"0bjFYstj8USMzEV4ZQldjA","stars":5,"date":"2017-01-19","text":"Cool little place, nice atmosphere and staff. Great coffee too, go check em out!","useful":0,"funny":0,"cool":0}

1. Data cleaning and extraction
2. Feature engineering and selection
3. Deal with data into MLLIB form
4. Data preprocessing/feature selection
5. Modeling and model selection