

# Project Proposal

**Team: Hungry Panda**

*(Captain) Tingyu Zeng - tzeng11*

*Sicheng Meng - meng29*

*Yiwei Kuang - yiwei9*

*Jiachun Tang - jiachun5*

*Fangsheng Yang - fyang28*

## Background

When Yelp users rate a restaurant, they tend to give the rating based not only on taste and quality of the food, but also other factors such as service, prices, comfort, atmosphere, parking, and location, etc.

As a group of Chinese students, our favorite dim sum restaurants usually don't have a high rating on Yelp, partly because traditional dim sum places focus on taste of food, but many of them don't focus on service, known as "cantonese hospitality". You have to read each review if you want to know how good the food is.

We don't want to let bad service overshadow good food. Through text retrieval and mining techniques, we think it would be a great idea to analyze reviews that focus on food only.

## Proposal

In this project, we are going to create a new rating system - an algorithm that will only be based on how good the food is! This Chrome extension will conduct text analysis on Yelp user reviews, and offer a one-click function to filter out the noise ratings not related to food quality. A new rating will be generated to help people who focus on food taste only and ignore other factors such as services, environments or any other factors.

Our tool will include following three main features:

- Ratings based on food only
- Dish recommendations - dos and don'ts
- Review highlights

## Data and Techniques

By using what we have learned in CS410 course, the algorithm will focus on specific words in each review to calculate the final score. We will use Yelp Fusion API and web crawling techniques to fetch restaurant review data on Yelp. The techniques we want to use include but not limited to the following:

- Web crawling (20 h): relevant comments will be pulled from the targeted restaurant page on Yelp and used as the data input.
- Text retrieval (40 h): methods we learned in CS410 course, such as VSM and language model, will be used to analyze and vectorize the data input as “documents”.
- Text mining (20 h): documents will be clustered and categorized. Methods like PLSA will be used to remove unrelated topics. Based on the cleaned data, new scores can be generated and dish recommendations will be pushed to the users.

On the front-end (10h), a Chrome extension will be implemented to provide an interface for users to interact with the proposed features. Javascript will be used.

On the back-end (10 h), requests will be processed with frameworks and libraries like Flask and Meta.py, and the result will be sent back to users.

## **Expected Outcome**

We're planning to spend around 100 total hours on this project - that is 20 hours from each group member.

Since this is a student project, we are not expecting a business level outcome, but we will spend time on both user interface and back-end algorithms, and try our best to implement the three main features mentioned in the proposal session, and offer end users an easy-to-use experience - with just one click, noise reviews that are not related to food will be filtered out. New ratings, dish recommendations and review highlights will be generated to help users only focus on food taste regardless of services, environments or any other factors.