Project Proposal

Author: Yilai Yan(yy97), Weijian Zeng(wz48), Wanrong Cai(wc52)

What is the problem you are solving

In the information searching part, we are going to collect restaurant names and apply tokenization, normalization, and other IR preprocessing steps on it. After that, for example, if a user types an ‘e’ in the search bar, the website gives all restaurant names with an e.

In the machine learning part, we are going to filter data out of stopwords and count the frequency of each word. Then add a positive or negative tag on it. We will predict the next year trending.

What data will you use

We will use the Yelp dataset

What work do you plan to do the project

The main steps include crawling yelp data, preprocessing the data, indexing the token, and finishing the UI.

Which algorithm, techniques, model you plan to use,develop

Algorithm:term-document incidence matrix, inverted index, permuterm index, spelling correction
Techniques and model: Solr, Django, Lucene, sentiment analysis model

Who will evaluate your method, how will you test it, how will you measure success

Methods:
We will build a machine learning model, using the previous five years' users' reviews to predict the next year’s reviews from users. For example, after preprocessing the data, we can calculate the number of good reviews and bad reviews. We can use that data to predict how good reviews are and bad reviews will be in the future.

Test:
We will split the data into training data and testing data. Use training data to build the model and use the testing data to calculate the accuracy. In order to get the accuracy of our more, we will use precision and recall and some other accuracy methods.

Success:
Training the model until the accuracy reaches 65%

What do you plan to submit, accomplish by the end of the semester

We are creating a website that allows users to search Houston restaurant names and the website shows the trend of customer comments in five years in the form of cloud words and line charts and the website will give the prediction of next year toward restaurant reputation.
In the cloud world, users will see positive words in bright color and negative words in gray color.
In the line chart, users will see years on the x-axis and positive comment levels on the y-axis.
Besides, there will also show some tags of the restaurants. All the tags come from the review of the previous. And the tags that we show, are the most common tags from users.