CS410 Project Proposal

Group Name: Team Experts

Group members:

Zhiyan Jiang(zjiang2), Xin Peng(xinp2), Shan Shan(ss163).

Captain: Zhiyan Jiang(zjiang2)

Topic: Perform sentiment analysis on Tripadvisor Hotel Review.

Description: Tripadvisor is a website where users can find travel-related information and leave reviews for hotels, attractions, etc. It is an important source for travelers to share their travel experience and plan for future trips. The goal of this project is to practice sentiment analysis on the reviews, and train models for proper classification and clustering.

Approach: The review analysis has multiple sentiments, which are negative, neutral and positive. We will follow the process of tokenization and bag of words to process the review texts. We will use Python to perform the analysis. Some libraries we may use are pandas, numpy, re.

Dataset: “Trip Advisor Hotel Reviews”.

Expected outcome: The outcome may include the general terms people use to express like or dislike about the hotels. We are going to analyze the words in the comment dataset and use F1 score to evaluate our model(s). We will also perform topic mining and provide top features 1-star and 5-star rating hotel reviews have.

Programming language: Python

Workload:

Each group member will spend at least 20 hours, and the total time will be at least 60 hours (N=3). The main tasks include:

- Weekly group meeting, two hours per week in 7 weeks. Estimate 14 hrs.
- Locate Tripadvisor Hotel Review training and test datasets. Estimate 1 hr.
- Perform data cleaning if required. Estimate 10 hrs.
- Train multiple potential multi-level classifiers, tune hyperparameters, and identify the classifier(s) with best performance. Estimate 20 hrs.
- Using Clustering techniques to identify most popular features among different ratings. Estimate 12 hrs.
- Present the results in the form of a report, Python source code, and a live demo. Estimate 12 hrs.