Final Project Proposal

Team: Forhad Akbar, Adam Douglas, and Soumya Ghosh

Introduction

With the Covid19 pandemic beginning to recede people are looking to travel again. The country has been locked-down for a while, and people are looking to book vacations to a variety of destinations.

One of the key choices for travelers looking to book a big trip, is to choose the proper accommodations. The website www.tripadvisor.com allows travelers to review and rate hotels based on their personal experiences. Those ratings are then summarized and presented to would-be customers to make decisions about where they want to stay while on vacation.

Although the ratings are aggregated in some form, do they present a reasonable metric for others to make choices? In other words, would sentiment analysis techniques applied to written reviews match the 1-5 numeric ratings?

Data Source

Data is provided by the UCI Machine Learning Repository and is part of the OpinRank study[1]. The dataset contains written text reviews for hotels in 10 different cities worldwide as well as aggregate scores for each hotel as of the date the data was compiled.

For this project, we will focus on a single city: Las Vegas, Nevada. There are 230 distinct hotels represented in the data, though we will likely remove any which do not have at least a reasonable number of reviews in the sample.

Data Concerns

The largest concern is that some hotels represented in the data might not have enough written reviews to generate a realistic sentiment score, however even if that is the case for a few hotels, there are 230 of them represented in the city data, so a meaningful sample size should still be attainable.

Work Plan

The work plan, in general terms, looks like this:

Task / Step	Primary Team Member
Data collection and loading	Adam
Data parsing and cleaning	Adam
Initial Analysis and Modeling	Forhad
Model Validation and Write-Up	Soumya
Presentation	TBD