ExecSummary

Hangyu Kang, Xiangyu Wang, Ruyan Zhou

November 2020

1 Introduction

There are more than 40 hotels in Madison area, we had wondered which hotels are evaluated as good hotels and what made them obtain good reputation. Hence, for the project, we had planned to analyze yelp reviews about the hotels in Madison area. Our questions for the analysis are the same as the following:

- What affects to customers' evaluation towards hotels especially in Madison area?
- What is the difference between high rated hotels and low rated hotels in Madison area?
- How AC hotel (Our virtual client) can be improved?

2 Data Cleaning and Pre-Processing

We have a given 4 datasets from the Yelp. There are the files that shows details of the businesses, the reviews about them, detail information about the users who leave the reviews and useful tips for the business owners.
First we merge review set and business set together according to their business ID where each row contains one business. Second we filter all rows to find which business contains 'Hotel' in categories column and filter again to eliminate irrelevant business. Third we cut off those users whose ratio is less than 2 which is $\frac{user.useful}{user.review\_count}$. Finally we break each review into single word to calculate frequency for future use. Then we get df.csv and df_tip.csv.
After pre-processing part we select hotels whose reviews are more than 5. And we use 3rd quantile as a threshold to determine which words are frequently used.

3 Analysis

3.1 Analytical Insights

3.2 Data-driven Action Plan

3.3 Important Evidence