Data 620 - Final Project Proposal

Bikram Barua & Leticia Salazar

Proposal: Analysis of TripAdvisor reviews

Motivation:

* Everyone love travelling and hotels play a crucial role in traveling. To plan your trip better, you want to find the best choice given your preferences.
* With this dataset, consisting of 20k reviews crawled from Tripadvisor, you can explore to find a great place to rest and use this to plan your travel journey.

Data Source:

The data will be obtained from Kaggle <https://www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews?resource=download>

1. Identify the links between key travel related phrases and their related ranking and sentiment.
2. Evaluate the relationship between the text contents and the sentiment rating.

Project Goals

Statistical Analysis:

1. Word frequency count
2. Percentage rating
3. Word cloud by rating
4. Review length
5. Top 20 most common words

Sentiment Analysis:

Using sentiment analysis, we will attempt to classify reviews as negative, neutral or positive. We will use natural language processing to examine the following questions:

1. Do useful review attributes affect rating?
2. Does review text affect the review attribute?
3. Can review text determine if a hotel receives a high or low rating (stars)?

Work plan

* Proposal
* Data load
* Data pre-processing
* Data cleaning
* Statistical analysis
* Data Visualization
* Building a Machine Learning Model / classification
* Sentiment Analysis
* Conclusion
* Review
* Presentation

Concerns

…