Final Project Proposal
Hotel Review Sentiment Analysis

Team Mr. Goose

Runyu Cao (runyuc2@illinois.edu) (Captain)
Weijie Wang (weijiew2@illinois.edu)
Qijing Zhu(qijingz2@illinois.edu)

Background

Hotel reviews are one of the most important factors influencing a person's booking selection and thus have a high impact on marketing strategies. Our team has chosen to utilize sentiment analysis, a technique of extracting emotions based on the selected hotels' textual reviews and producing feedback and solutions to benefit travelers by comparing different options. A challenge we will be discussing is whether a sentiment analysis model learned on hotel reviews can process different kinds of text across domains.

Implementation Planning

Dataset

We use the dataset from Hotel Reviews Data in Europe and separate them to be used as the training, validation, and testing data. We will perform a series of data cleaning on the dataset before training the model, such as stop word removal, word stemming. In addition, we could convert the review text to n-gram corpus as appropriate for the model of choice.

Algorithm

Stemming, BM25, Logistic Regression, K-Nearest Neighbor, XGBoost

Tech Stack

We will use Python 3.8 as our main programming language. Jupyter notebook will be used to simplify the process of presenting the data and trying out different models. Scikit learn will be used for its processing functions and classification models. Nltk will be used for its pre-defined English stopword list, stemming tool, and tokenizer. Metapy will also be used.

Evaluation

To evaluate the tool and demonstrate, we will show 10 reviews to users, and users are expected to mark these reviews from 1-5 based on their sentiment. After collecting users' scores (true label) and predicted scores, we will measure the average accuracy and average difference.

Timeline

Task	**Time estimate**	**Note**
Dataset Selection	3 hours	Background study
Proposal	3 hours	
Dataset Cleaning	10 hours	Feature engineering
Coding	25 hours	Debugging, Evaluation
Team Meeting	6 hours	Weekly 0.5-1 hour sync up * 6 weeks
Progress Report	2 hours	
Final Report	5 hours	Drafting, Proof-reading
Presentation	5 hours	Recording, Video editing
Total	≈ 59 hours	