CS 410 Project Proposal

Project Title: Sentiment Analysis on Disneyland Reviews

Team Members

- Jingxin Deng (jdeng19): jdeng19@illinois.edu
- Yujia Qiu(yujiaq5): yujiaq5@illinois.edu
- Luxia Yin(luxiay2): luxiay2@illinois.edu

Project Coordinator

- Jingxin Deng (jdeng19): jdeng19@illinois.edu

Project Keywords

- #Text Mining
- #Sentiment Analysis
- #Feature Engineering

Project Description

- **Project Objective**
    - This project aims to conduct sentiment analysis on Disneyland Reviews data to develop a model that accurately predicts ratings based on customer review content. Analyzing sentiment in this data provides valuable insights, such as identifying commonly used words in positive and negative comments, uncovering factors that visitors frequently mention, and exploring specific concerns related to different Disneyland theme park locations. This project is valuable because it helps Disneyland understand what matters most to their visitors, allowing them to address any issues and improve the overall experience. The insights gained could also support better marketing strategies and customer service improvements, making Disneyland an even more enjoyable place for visitors.

- **High-Level Approaches:**
    - Data Preprocessing
        - Text Cleaning:
            - Lowercasing
            - Removing URLs, handles, punctuation
            - Removing Nulls & Duplicates
        - Tokenization
        - Stemming
        - Stop Words Removal
        - Label Encoding
        - Feature Extraction: like TF-IDF or Word2Vec.

    - Data Splitting
        - Split Data: Divide the data into training and test sets (e.g., 80/20 split) to ensure the model is trained and evaluated on separate datasets.
    - Model Training and Evaluation
        - We plan to use various regression models to predict the rating from review text and evaluate their relative performances. The models used include:
            - Logistic Regression
            - Decision Tree Classifier
            - XGBoost Classifier

- **Evaluation and Demonstration of Effectiveness**
    - Evaluation Metrics:
        - Precision, Recall, F1 Score
        - Accuracy
        - Confusing matrix
    - Visualizations:
        - Word Cloud of Review Text
        - Classification Report Plot
        - Feature Importance Plot

Tool, System and Dataset

- Kaggle Dataset: Disneyland Reviews https://www.kaggle.com/datasets/arushchillar/disneyland-reviews/data
- Python (and python libraries)
- Google Colab

Workload and Timeline

- Identify project research: Data Collection **4 HOURS**
- Data Exploration and Visualization **6 HOURS**
- Data Cleaning & Text Preprocessing **6 HOURS**
- Feature Engineering **8 HOURS**
- Project Progress Report **2 HOURS**
- Model Exploration **7 HOURS**
- Model Training and Model Fine-tuning on different models **12 HOURS**
- Model Evaluation - Testing and Validation **7 HOURS**
- Algorithms Comparison **6 HOURS**
- Overall Documentation and Presentation **10 HOURS**