Project Write-Up

Design:

The aim of this project is to use Natural Language Processing and Unsupervised Learning exploration techniques on all the reviews of the Disney Parks in California, Paris, and Hong Kong to gain insight into the overall sentiment of the parkgoers. Additionally, to further understand the sentiment of positive and negative reviews, basic topic modeling and more exploratory data analysis are performed to help understand the context of these sentiments.

Data:

A dataset of over 42,000 reviews on the Disney Parks from California, Paris, and Hong Kong can be found on Kaggle. Reviews consist mainly of textual data. Each row consists of a unique review along with its unique ID, reviewer, date, Disney Park being reviewed, and a rating score from 1 to 5.

Algorithms:

Unsupervised Learning techniques used to explore the sentiment of the textual reviews included:

- Sentiment analysis with VADER
- Dimensionality reduction and topic modeling with:
    - Latent Semantic Analysis (LSA)
    - Non-negative Matrix Factorization
    - CorEx
- Vectorization with:
    - CountVectorizer
    - TfidfVectorizer

Tools:

Python text processing tools and libraries were used, such as:

- NLTK
- spaCy
- scikit-learn
- Regex

Visualization tools included:

- WordCloud
- Matplotlib
- Seaborn
- Scattertext
- and Canva for the presentation slides