SDAIA T5 bootcamp

NLP and Topic Modeling Project

Disneyland Review Analysis

Final Report

Basma Ghazi Abdullah Alduaiji

Mai Abdullah Fahad Aljuaid

11/11/2021

Contents

Table of tables

Introduction

Disneyland is a famous amusement park with eight themed amusement rides, shows and costumed characters. Moreover, the visitors of Disneyland have over 52 million yearly. We will develop a topic modelling and take the result as input for the classification model can predict the rating based on the review text.

Design

The dataset in this project is taken from the Kaggle website. This data set presents the complete status of the Disneyland reviews and the stars rating.

Algorithm

The steps to analyse the data set are loading data from Kaggle. In addition, exploring data by using all the functions like info and describe. Then cleaning data by removing null values and duplicates as well as, removing any irrelevant data. Then, cleaning the text data using NLTK. After that, plot the graphs using seaborn, matplotlib, plotly, folium and word cloud modules from python. Finally, building models, clustering and topic modelling.

Used data set

The dataset that will be used in this project was found in Kaggle. In addition, dataset columns are shown in table 1.

Table 1 Disneyland Reviews Dataset

Review_ID	Unique number of each review
Rating	Number of rating point
Year_Month	The date of publishing review
Reviewer_Location	The reviewers' countries
Review_Text	Text to represent Disneyland experiments
Branch	Disneyland branch name

In addition, The Individual sample from the Disneyland reviews data set are 20K rows. Moreover, the expected characteristic used to predict the rating is the reviews text. Furthermore, the predicted target is to predict if the review is good or bad.

Used tools

The tools provided as modules in Python such as Pandas, Matplotlib, Seaborn, NumPy, folium, plotly, word cloud and sklearn. The python libraries support multiple data analysis and cleaning methods to ensure the data is clean and ready to be visualized.

Conclusion

NMF is good in displaying topics.

The results show that people love Disneyland despite the long waiting hours in rides.