أكاديمية سدايا
SDAIA Academy

Trip Advisor Hotel Reviews

NLP Project

By:

Alanoud Alhussain

Amal Altamran

Amirah Alotaibi

Introduction:

The goal of this project is to build a topic modelling for Tripadvisor hotel review

And we know hotels play a crucial role in traveling and with the increased access to information new pathways of selecting the best ones emerged. Travelers across the globe use the Tripadvisor site and app to discover where to stay and where to eat based on guidance from those who have been there before.

Dataset:

The dataset is taken from Kaggle, which consists of 20k reviews from TripAdvisor. However, in this project, we will explore the hotel reviews and the rating base on customers' hotel experience.

Review	Review Text
Rating	Review Rating (stars)

Tools:

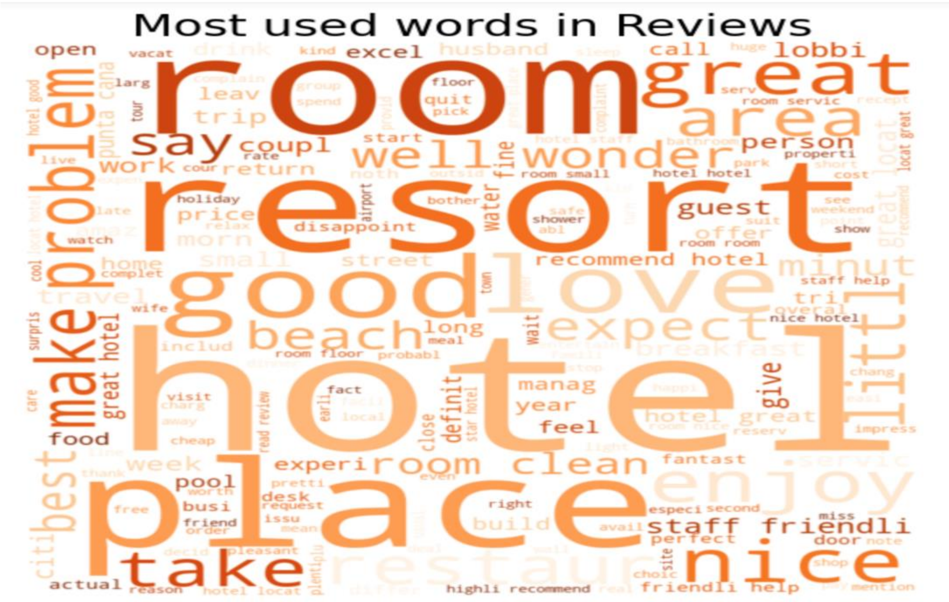
- Scikit-learn to perform unsupervised learning.
- NumPy and Pandas to perform data manipulation.
- NLTK to perform text manipulation.
- Word cloud to represent the frequency or the importance of each word.
- Seaborn and matplotlib for visualization.

Algorithm:

Text Preprocessing

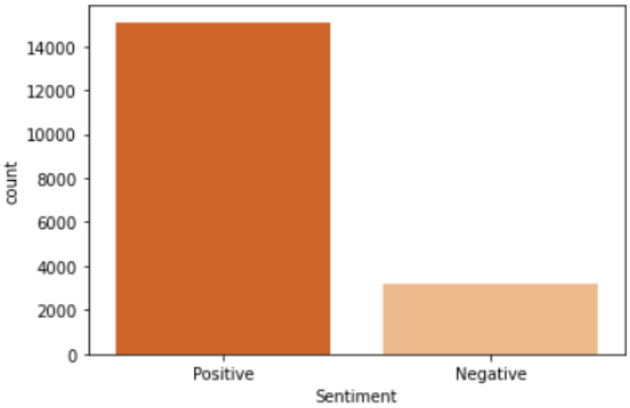
1. Removed digits.
2. removing stop words, punctuations and Extra space.
3. remove HTML tag/markups and duplicate letters in words.
4. Converted all characters to lowercase
5. Applied lemmatization and Stemming with NLTK
6. Removing useless words like [also, tell, meanwhile, however, arriv, felt, time].

And after text Preprocessing this is our Most used words in Reviews



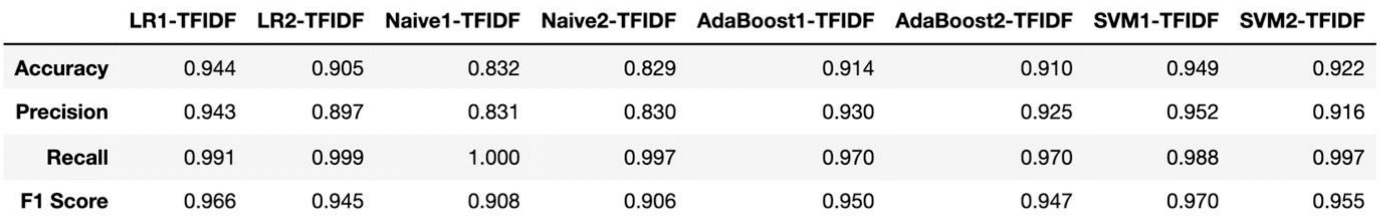
Sentiment Analysis

After we do the text preprocessing, we do Segregating and Encoding Positive and Negative labels



Classifying

We Split the data into training and test sets, and then create TF-IDF versions and create a function to calculate the error metrics, and in the end compile all of the error metrics into a dataframe for comparison



	LR1-TFIDF	LR2-TFIDF	Naive1-TFIDF	Naive2-TFIDF	AdaBoost1-TFIDF	AdaBoost2-TFIDF	SVM1-TFIDF	SVM2-TFIDF
Accuracy	0.944	0.905	0.832	0.829	0.914	0.910	0.949	0.922
Precision	0.943	0.897	0.831	0.830	0.930	0.925	0.952	0.916
Recall	0.991	0.999	1.000	0.997	0.970	0.970	0.988	0.997
F1 Score	0.966	0.945	0.908	0.906	0.950	0.947	0.970	0.955

By using TF-IDF we able to see that the recall and precision of the first SVM model still outperforms the other models.

Overall, the first SVM model with F1 score 0.97 (using unigrams) seems to best classify positive and negative Trip Advisor reviews.

Models

We apply 4 models to obtain the best topic model that describe our data which is

- NMF
- LDA
- CorEx
- LSA

By using Count Vectorizer and TF-IDF Vectorizer

And the model performing the best topic modeling is **LDA for TF-IDF Vectorizer with 3 topics.**

Future work

Add another dataset contain the hotels name to do recommendation system.