Unsupervised/NLP Project

Hotel Review Topic Modeling for Hotel Recommender

By William (Wei-Lun) Mai

Abstract

This project explores the solution of topic modeling for hotel reviews. By utilizing Natural Language Processing (NLP) techniques, we are able to classify hotel reviews into different topics. The model enables the end-users, either hotel management or hotel guests, to filter on the topics they are interested in reading. Furthermore, hotel booking sites might be able to attract more users by providing a more tailor-made recommendation system based on this solution.

Design

First, we would like to classify hotel reviews into different topics. The topics could vary by location and the type of hotel (think snow resort vs. city hotel). To remove this variable and prove the model will work, we conduct the study on 50 hotels in the same area - Times Square. After the success of topic modeling, we further explore building a hotel recommender based on the topic modeling solution with the help of sentiment analysis on the reviews.

Data

We use this Hotels API hosted on RapidAPI to extract hotel reviews from Hotels.com. Specifically, we randomly chose 50 hotels around New York Times Square and retrieved 50 most recent reviews from each hotel. JSON responses from API were later flattened into a pandas DataFrame. Each observation is one review from one hotel. We later split the reviews into individual sentences for further analysis.

Algorithms

First of all, we used Snowball Stemmer to pre-process the review sentences. Second, CountVectorizer and TfidfVectorizer were used to construct document-term matrices. We subsequently performed two different topic modeling techniques, Latent Semantic Analysis (LSA) and Non-Negative Matrix Factorization (NMF), on our two document-term matrices. Four combinations were evaluated, and the best one was chosen - the combination of CountVectorizer and NMF. We also used VADER to score the sentiment for each sentence. Last but not least, we took the average of the sentiment score for all sentences in the same topic for each hotel. The average score for each topic and hotel was later used to implement our hotel recommendation system.

Tools

- Numpy and Pandas for manipulating data
- Requests and json to work with API
- NLTK for stemmer, tokenization
- VADER for sentiment analysis
- Sklearn for vectorizer and matrix decomposition

Communications

Presentation pdf, write-up, and Excel files can be found on this GitHub page:
https://github.com/zyzzyva1423/NLP_Project