NLP Project Write-up.

Vacation Recommendations

Abstract

The goal of this project is to create a vacation recommender using the sentiments of tourists visiting cities and their attractions at different times of the year. I worked with TripAdvisor reviews of 'Things to Do' to perform sentiment analysis, topic modeling, and linear regression analysis to understand which attractions and cities have a substantial seasonal component. Using the sentiment analysis, I built a recommender that proposes a city and top attractions to visit based on the month of the year.

Design

I scraped the reviews from the 'Things to Do' section of the TripAdvisor website. I focused on Top 10 attractions of Austin, Chicago, and New York, excluding hotels, restaurants, and landmarks.

I developed a recommending system using the sentiments of the reviews aggregated by month for each city. The recommending system proposes a city and top N attractions to visit in the city in each month of the year. I also explored which cities' reviews have a substantial seasonal component, constructed a seasonality score using topic modeling, and analyzed whether the variation in user sentiments can be attributed to changes in seasons.

Data

For each of the Top 10 attractions in these three cities I collected around 600 reviews. As a result, the dataset consists of over 18000 user reviews with dates when the attraction was visited by the user.

Algorithms

- LSA, NMF (with CountVectorizer) for topic modeling.
- Corex for topic modeling using anchor words to identify reviews that discuss weather and seasons. Also Corex was used for constructing and calculating the seasonality score.
- VADER for sentiment analysis
- Linear regression to explore the relationship between users' sentiment variation and the seasonality score.

Tools

- Selenium for scraping TripAdvisor reviews;
- Python, Pandas, and Numpy;
- NLTK for preprocessing;
- LSA and Corex for topic modeling, vectorization using CountVectorizer, VADER for sentiment analysis;
- Linear Regression and statsmodels for regression analysis.

Communication

- 5 minute presentation to client.