A Recommendation System for Museum Selection

Anne Chen

09/20/16

Project overview

Why Built this App

- There is no existing solution to locate museums based on one's preference
- Finding a right museum is time consuming

Data Collection and Visualization

- Scraped ~1600 data from TripAdvisor

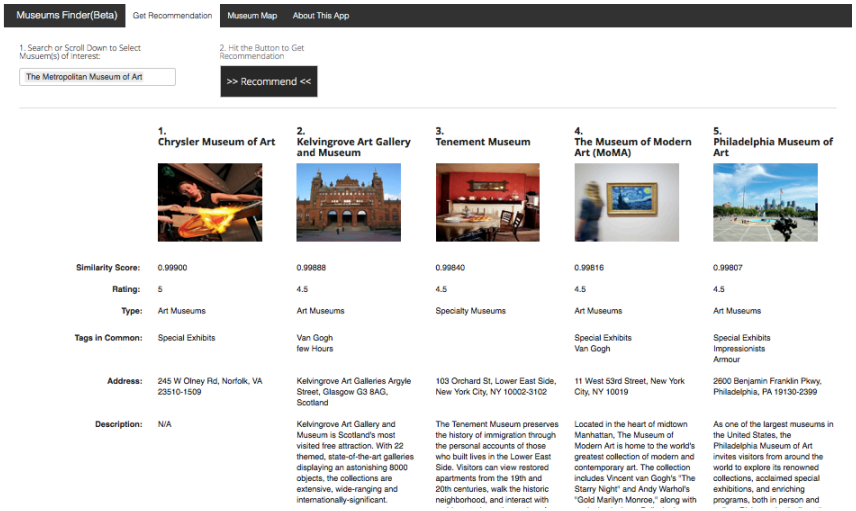
Recommendation System Construction

- Engineered new features including sentiment analysis scores
- Imputed missing values with KNN
- Developed content–based recommendation system
- Implemented system into Shiny App

The app is built for people to locate museums of interest

1. Explore museums based on users' preferences

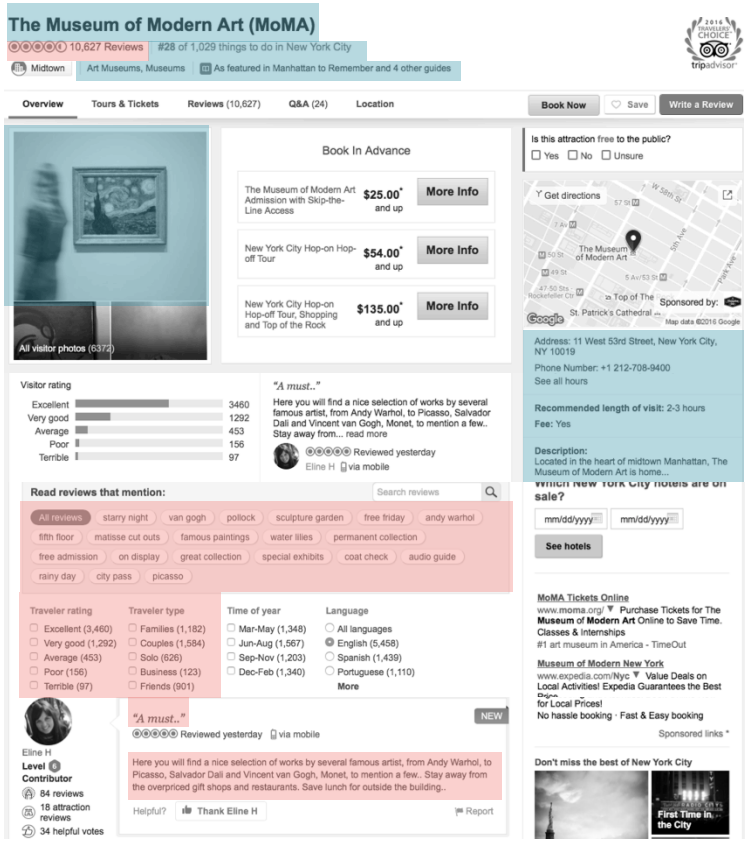
2. Allow users to filter museums that meet their needs



First, I scraped data of ~1600 museums from TripAdvisor

Museum Info

- Museum name
- Address*
- Category
- Relative Popularity**
- Times being featured
- Recommended length of visit
- Fee
- Description
- Displayed rating (0-5)
- Image



Traveler & Review Info

- Number of reviews
- Review tags
- Traveler rating
- Traveler types
- Quotes
- Review contents

* Addresses were passed to Google Geocoding API to convert them to latitude and longitude
** Relative Popularity is later computed as X/Y where rank X of Y things to do in the same city

Next, I designed new features with museum categories, review tags, and textural data

83 Distinct Categories

- Art Museums
- Children's Museums
- History Museums
- Natural History Museums
- Science Museums

100 Most Frequent Review Tags

- Gift shop
- All ages
- Special exhibits
- Few hours
- Well worth a visit
- Permanent collection
- Free admission
- Audio guide
- Beautiful building

Textural Data

10 Quotes

10 Reviews

Museum Description

Sentiment Analysis

Polarity & Subjectivity

Max

Average

Min

Variance

Finally, I constructed a recommendation system based on cosine similarity

Select Museum(s)

Compute and sort cosine similarity across ~1600 museums using 219 features*

Exclude scores of museums selected

Display 5 museums having highest similarity scores

*"ReviewCount" and "TotalThingsToDo" were removed

Demo Time! 

Questions?

Other Slides

If I had more time…

- Predict if a museum will be featured or not
- Cluster museums based on review polarity or subjectivity
- Apply association rules to see if there are rules for combinations of review tags

I imputed missing values using KNN and dummified categorical variables

- Imputed missing values with KNN using k = 40
- Created dummy variables for “Country”, “Fee”, and “Recommended Length of Visit”

Rank, review polarity, and quote subjectivity are important features in predicting rating

- Random Forest (Classification)
- Some important features:
  - Rank & Relative Popularity
  - Review Polarity
  - Quote Subjectivity
  - Review Count