Dmitriy Kats
Springboard DS March ‘19

“True Review” a Personalized Recommendation Engine

1. Overview

Many times we find ourselves on Yelp, looking for a new restaurant in the area with 4000 reviews. There is no way we can go through that many reviews to understand how people rate and what they value. Our recommendation system will take a given user’s reviews and compare to others’. Based on content and dining style, we’ll present more relevant restaurants with a better fit. A content / collaborative based recommendation engine that will also average the star ratings and predict what a given user might rate the restaurant. Then sort the recommendations based on predicted star rating of a user (and accuracy of prediction).

2. Tools

Data preprocessing (standard tools)

Tokenization - break down text into sentences / words (smaller chunks)

Remove stop words, punctuation

Stemming - root form of word

Lemmatization - vectorize the words

Bag of words to count occurrences

More advanced analysis - Additional features for content filtering

Vectorize the Categories information for each restaurant

Use Gensim LDA to predict topics from users’ reviews

Add sentiment analysis using TextBlob to improve the model

Positive and negative reviews

QA is going to be very important. We will verify the performance of these NLP tools in order to qualify the applied model. A number of restaurants and reviews will be selected and labeled manually and compared to the models.

3. Prediction Model Product Description

We can start with a simple collaborative recommendation engine:

	Restaurant 1	Restaurant 2	Restaurant 3
User 1	5	5	2
User 2	4	5	2
User 3	5	4	?

From above, given reviews of User 1 and 2 for the same restaurants 1, 2, and 3, we can predict that User 3 will give a review of 2 stars.

If we then look at a content based engine:

	Restaurant 1	Restaurant 2	Restaurant 3
User 1	Topic 1	Topic 2	Topic 3
User 2	Topic 1	Topic 2	Topic 3
User 3	Topic 1	Topic 1	Not recom.

Based on restaurant review topics, we can recommend or not recommend a restaurant based on a given user’s most commonly mentioned review topics. From above, since user 3 typically mentions topic 1 in his / her reviews, Restaurant 3 is not recommended due to Topic 3 being mentioned in majority of the reviews.

Taking this a step further and combining both models, this same concept can be applied to viewing existing reviews from other users delivered in a personalized fashion:

	Chinese		American New	
	Restaurant 1	Restaurant 2	Restaurant 3	Restaurant 4
User 1	5	5	2	1
User 2	4	5	2	3
Yelp Shows	4.5	5	2	2
User 3 Personal View	3.5	4	3	3
User 3	2	?	4	?

This is similar to movielens.org where blue stars represent predicted personal ratings for each movie that are not yet rated by the user. And red stars represent actual user ratings.

4. Data
Yelp dataset will be used.

5. Deliverables
Code and a supporting report will be provided through GitHub.