

Extracting Rating Dimensions with Text Reviews

Xinzhe Yang, Xuwen Shen

October 22, 2017

1 Research Context

The same rating from different users usually stands for different meanings. Individual users may assign different weights to such aspects when determining their overall score. For example, a spendthrift hotel reviewer might assign a low weight to "price" but a high weight to "service", thus explaining why their overall rating differs from a miserly reviewer who is only interested in price. There exists hidden information in reviews that leads to the final rating. By extracting the information, we can get where a restaurant shines and what it needs to improve. If we have k topics in reviews then we extract k (the same number as the number of topics in Yelp reviews) dimensions in rating for each topic respectively and then use the k dimensional rating to compute the recommendation score for an individual.

The importance of customer reviews has been proven to be important in terms of giving insights to businesses and also providing customers personalized services. LDA has been widely used to analyze the latent meanings in reviews and study customer behaviors. A previous study investigating the topics in hotel reviews uses LDA to extract dimensionality of customer satisfaction and it simply compares the topics extracted by LDA to those in previous studies as evaluation.

2 Research Objective

The ultimate goal for the project is to create a recommendation system which recommends restaurants to a specific user given the user's preference and the restaurants' rating with respect to the user's preference. To achieve the goal, the most important part is to extract dimensions for both overall rating for restaurants and user preference using information from reviews. Information we care about includes what a specific user cares about, food or view, price or service and what factors lead to the overall rating for a specific business.

Goals:

1. Use LDA to set a baseline for exploratory methods
2. Find a way to preprocess the review data so that it fits the autoencoder model
3. Build a neural network to learn complete sentence from reviews with tensorflow.
4. Evaluate the two methods and explain why one outperforms the other
5. Apply the learned topics to a recommendation system

3 Research Road Map

3.1 Experiment Design

To learn the summarization from the reviews, topic modeling methods are considered at first. We will start with LDA, a popular topic modeling method to get m keywords in n topics. However, there may exist problems in the results. On the one hand, topics we get from LDA may fail to describe the correlations among the words in each topic. For example, the topic can include the word "food" and a name for a specific food such as "steak", which suggests strong co-occurrence or hierarchical relation among words in one topic. However, similar words are not necessary for topic modeling since they point to the same topic. On the other, the topics from LDA are composed by keywords instead of a complete sentence. We may fail to get the precise meanings from keywords

without link between them.

To address the first problem, we can use the dictionary to define the types of words and then substitute similar words with one word. Alternatively, we can use hierarchical topic modeling to get different layer of topics. Intuitively, the method searches for a set of latent factors that best explain the correlations in the data as measured by multivariate mutual information. The method is unsupervised, requires no model assumptions. To address the other problem, sequence to sequence can be applied to compress the original text into its summary, importantly, with complete sentences.

3.2 Experiment Validation

Set LDA as a baseline and then compare all the results with topics we get from LDA. If the two main problems are addressed, for example, the number of similar words decreased or links of words make sense, the research results can be verified. We evaluate the word similarity by computing the correlation among words in each topic and links of words by human evaluation.

3.3 Research Timeline

Oct 21 - Oct 28: Analyze outputs from LDA and delete words with strongly positive correlation or hierarchical relation in the topic using dictionary. Evaluate the results by comparing the results from LDA.

Oct 28 - Nov 4: Use hierarchical topic modeling to get different layer of topics.

Nov 4 - Nov 11: Do research on text summarization using neural network, especially previous work on sequence to sequence.

Nov 12 - Nov 19: Apply sequence to sequence to compress the original text into its summary. Evaluate the results by comparing the results from LDA.

Nov 20 - Nov 27: Compare the results from methods above and evaluate if they address the problem in LDA.

Nov 28 - Dec 1: Build a recommendation system which recommends restaurants to a specific user given the user's preference and the restaurants' rating related to the user's preference.

3.4 Resources

We use the datasets of yelp reviews, business and user to extract topics from reviews for popular restaurants (restaurants with most reviews) active users (users with most reviews).

References

- [1] *Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation.*
- [2] *Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach..*
- [3] *Multi-label text classification with a mixture model trained by EM.*
- [4] *Hierarchical Topic Models and the Nested Chinese Restaurant Process.*