Comparing LDA and NMF Techniques for Topic Modeling Positive Yelp Reviews

By: Vikas Gudhe

Introduction

Finding the perfect restaurant or store is hard – whether you're looking for something new, something similar, and especially when traveling someplace you haven't been before. While there are Google reviews, Yelp, and social media, none of them can tell you where to go without extensive searching and evaluation. But despite the needed research, according to surveys conducted by Statista, 51% of US travelers said they would "spend less than one week conducting research" before they go on a trip and 40% of users who preferred using apps said, "an important feature was how quick it was to book". While there are many elements to planning a trip, restaurants play a key role and Yelp provides us with a large dataset to start evaluating and building the beginning of a travel recommendation engine. I decided to compare a term frequency–inverse document frequency statistic (fit to a Non-negative Matrix Factorization) and a bag of words fit to latent dirichlet allocation model for Topic Modeling to try and find which would be more effective for identifying unique clusters of words that could help build the parameters for a recommendation engine.

Preparing the Data

1. First, download the yelp dataset from their website. Then use the Python package json and string to parse and clean an arbitrary amount of business's data (this was the first 10,000 reviews).

2. Given this data, store it in a dictionary with the business ID as the key and the review text as the value.

3. Use a python cleaning script to clean all the review texts of punctuation, and ensure that everything is utf8 and ascii readable for the model.

4. Utilize a python script to tokenize the reviews and generate text tokens based on the reviews, and clean the tokens of stopwords that will negatively impact the analysis, then lemmatize the tokens and join the lemmatized words together.

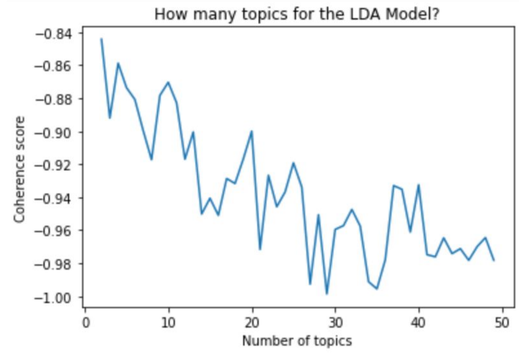
5. Using the gensim python module, generate the LDA model for analysis based on the number of topics and plot the topics as a bar graph. Then, do the same for the NMF model.

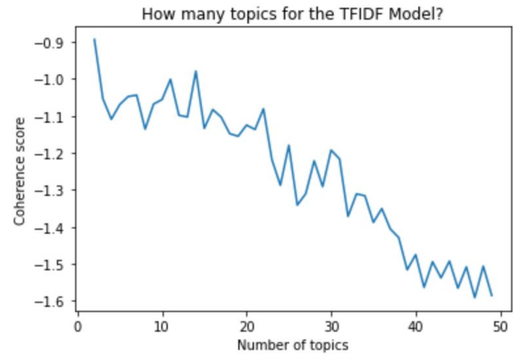
6. Plot the models on graph with respect to changing the number of topics that go into generating the model and see how this impacts the coherence score measure in a plotted line graph.

7. There were 4,299 unique businesses' reviews that contributed to this analysis after this process was complete.

Coherence Analysis

Once the pre-processing was done, the first form of analysis I did was with a coherence model, coherence of a topic is quantified by measuring the semantic similarity between the words that have scored the highest. This works similarly to most natural language processing techniques and results in the words being more human-readable. The topic coherence model selects the most frequent words in a topic and then aggregates and computes all pairwise scores (UMass) for each of the words.

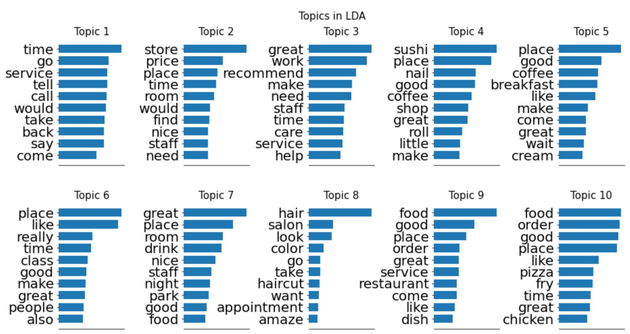


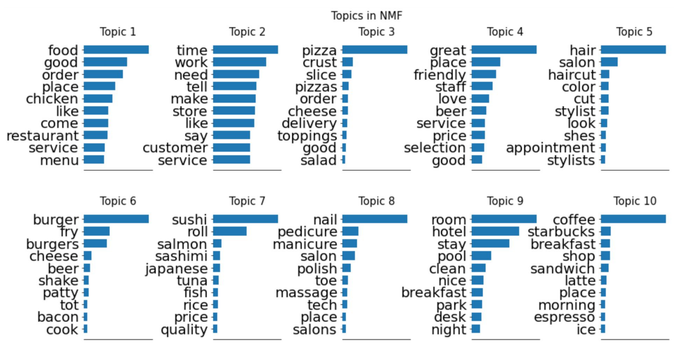


Since reviews tend to be short, even with aggregation as some restaurants may not have a large quantity of reviews, a corpus with short documents (reviews for us) will likely be more difficult to use with a coherent model than a corpus with longer texts. Choosing the number of topics is subjective but utilizing where the average score plateaus is common. Since both the TFIDF model and LDA model show sharp declines around 10, that was the number of topics used. In line with previous research, the TFDIF model had a significantly lower coherence score across all topics as it is less resource intensive to compute and faster but generally not regarded as useful.

Topic Modeling Analysis & Conclusion

After the coherence score was completed, the corresponding bag of words model and TFIDF model were fitted to LDA and NMF models, respectively. Interestingly, while the coherence score was higher across the board for the LDA model, the words identified when fitted to the NMF one seem more related to the given topic. Comparing Topic 5 from LDA and Topic 10 from NMF, the words seem related to coffee shops and thus words like "morning", "breakfast" and "shop" are more relevant for a recommendation engine than "great, "like", "wait", "cream" etc. This was similar when comparing topics like topic 7 from NMF and Topic 4 from LDA as well. The TFIDF model fitted to NMF seems to be finding words that appear less frequently but more correlated across the topic as there is a sharp decline in frequency of the words within topics as opposed to the LDA topic models.





Future Work

With more time, there is more work to be done, such as creating a word cloud for easier comprehension, building a domain-specific dictionary with words derived from user-testing and more. But this provides a starting point for future work in more exhaustive topic modeling.

Acknowledgements: Special thanks to Professor Katherine Elkins and Professor Jon Chun for assistance and the knowledge to conduct this exploratory data analysis.

References

Some code modified from:
https://www.kaggle.com/devmaxime/finding-topics-in-yelp-s-bad-reviews/

https://www.sciencedirect.com/science/article/pii/S0306437920300703

http://tfidf.com/

https://www.statista.com/markets/420/travel-tourism-hospitality/