P-25: Analysing Yelp Reviews using NLP techniques

Shah Haard Prashant (CSCI 6509), Nishant Amoli (CSCI 6509)

1. Problem Statement

Nearly 60% of restaurant don't make it past the first year of opening [1]. One of the major reasons is uncertainty about what they are doing right and where they are going wrong. Reviews are a great source of feedback for any business. However, if there are huge number of reviews, it becomes tedious and time consuming to analyse each review. Hence, this project focuses on determining the positive and negative aspects of a restaurant of a certain region. These topics can be extracted and weighed from Yelp dataset for restaurant reviews [2] available in JSON format. The dataset contains more than 5 million reviews for Canada and USA combined. This analysis would help the restaurants of that region to gain business insights and potentially improve on certain key aspects. Topic modelling techniques like NMF, LDA can help in determining highly weighed topics in the review dataset.

2. Possible Approaches and Relevant work

There has been lot of work done with the Yelp dataset and its analysis. Some of them include sentiment analysis on dataset, predicting ratings of restaurant based on reviews and predicting international restaurant success. However, the area of concern for this project is the reason behind success and failure of various restaurants. This is because positive reviews can greatly impact the restaurant business. A $1 million business restaurant can increase its sale by $180,000 if its Yelp rating increases by just 2 stars [3]. Hence, word of mouth is a great tool when it comes to restaurant business.

Factoring of models having discrete data can be performed in various ways. An older and basic approach is Latent Semantic indexing (LSI) [4]. LSI is an information retrieval technique and makes use of single value decomposition. It helps to determine the frequency of certain topics in a document. In LSI, negative weight words are permitted, and it can become a problem during its implementation on certain applications. However, there is generative probabilistic model called Probabilistic Latent Semantic Indexing (PLSI) [5]. PLSI leads to substantial performance gain as compared to LSI but has problems about overfitting [5].

Another common method for unsupervised learning of Yelp Dataset reviews is Latent Dirichlet Allocation (LDA). However, one major factor to consider while using the LDA is that we must know the number of topics beforehand. Hence, we assume that there should be *k* topics according to which documents (reviews) are generated [6]. As a generalization of this approach, LDA helps to quickly summarize search and summarize a huge collection of documents. It can be used to predict stars of hidden topics from the reviews. Extracting subtopics from Yelp review using LDA gives insights about users that are unlikely to rate high stars during peak times [7].

A more modern approach which has resulted in more successful topic modelling for terms and documents is using non-negative matrix factorization (NMF). The LSI, K-means clustering use similar approach for factor analysis unlike NMF. NMF provides a positive factorization of given positive matrix. Hence, each document is guaranteed to take only non-negative values in all the latent semantic directions [8]. Experimental results have shown that NMF is better as compared to SVD and eigen document clustering in reliable and accurate document clustering [8].

3. Project Plan

In order to complete this project, we have divided the project in 4 phases which we plan to accomplish in due to time. As the reviews are already labelled out of 5-stars, we are not going to perform sentiment analysis to identify positive and negative reviews. Hence, positive and negative group of reviews can be extracted from the ratings. For this, we will keep 1- star and 2-star rating as negative and 4-star, 5-star ratings as positive reviews.

During the first phase, the primary focus would be on restaurant review topic analysis which includes extracting only restaurant reviews from the Yelp dataset, analysing the provinces with most businesses in the dataset and selecting a province using minimum review rule so that further steps can have sufficient data for analysis. Moreover, stop words will be removed and the reviews shall be tokenized using TFIDF and 1-gram approach. This milestone should be completed till 9th March. Hence, we will allocate 6 days for these tasks.

In the second phase, we will analyse effectiveness and implement various NLP methods for topic modelling like LSA, LDA and NMF. This is important phase of project and we will allocate 10 days to perform these tasks. At the end of this milestone, major NLP tasks would have been completed like topic modelling all reviews, adding topic weights and business information to dataframes. This milestone is expected to be completed by 19th March.

Third phase includes Visualization of gathered weighted topic of restaurant reviews. This visualization would be done using Tableau. This will help to analyse a restaurant from restaurant id based on its good and bad aspects. As this is not major concern for the project, we plan to complete this milestone by 22nd March.

During the final phase, we will document our project work with appropriate results and references. The project report will help to gain clear understanding of each task and mention possible future work. This milestone will be completed by 31st March.

Hence, we plan to follow this feasible project plan to achieve our goals and complete the project in time.

4. List of References

[1] H. Parsa, J. Self, D. Njite and T. King, "Why Restaurants Fail", *Cornell Hotel and Restaurant Administration Quarterly*, vol. 46, no. 3, pp. 304-322, 2005. Available: 10.1177/0010880405275598.

[2]"Yelp Dataset", *Yelp.com*, 2020. [Online]. Available: https://www.yelp.com/dataset. [Accessed: 02- Mar- 2020].

[3] R. Fuggetta, "How to Turn Customers Into Advocates," *Restaurant Hospitality,* vol. 96, *(12),* pp. 22, 2012.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G.W. Furnas, and R. A. Harshman. “Indexing by latent semantic analysis.” Journal of the American Society of Information Science, 41(6):391407, 1990.

[5] T. Hofmann, "Probabilistic Latent Semantic Indexing", *ACM SIGIR Forum*, vol. 51, no. 2, pp. 211-218, 2017. Available: 10.1145/3130348.3130370.

[6] D. Blei, L. Carin and D. Dunson, "Probabilistic Topic Models", *IEEE Signal Processing Magazine*, 2010. Available: 10.1109/msp.2010.938079.

[7] J. Huang, S. Rogers and E. Joo, "Improving Restaurants by Extracting Subtopics from Yelp Reviews", *Ideals.illinois.edu*, 2020. [Online]. Available: https://www.ideals.illinois.edu/handle/2142/48832. [Accessed: 02- Mar- 2020].

[8] Wei Xu, Xin Liu, Yihong Gong, “Document Clustering Based on Non-negative matrix Factorization”, In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR ’03), pp. 267-273, 2003.