CASE STUDY: WHAT MAKES A RESTAURANT POPULAR?

The restaurant industry is extremely competitive, and it is important for someone running or opening a restaurant to understand what makes a restaurant popular and profitable in their context (location, price point…). We use text analytics on actual reviews from Yelp to try and figure this out.

Opening prompt

As a crowd-sourced business review and social networking website, Yelp has hundreds of millions of reviews. However, it only provides a holistic view by giving review ratings. In this project, we separated overall ratings by different categories to gain an insight into which categories will influence the stars most. For example, a 4 stars review: "Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks." This example includes different topics such as environment, service, food and price. By generating the topics and sentiment intensity score from the user's review, we can use linear regression to generate coefficients of each topics and understand the relationship between stars and topics.

High-level description of solution

We focused on the reviews from Las Vegas restaurants. Firstly, we cleaned the data and tokenized reviews by sentences. Secondly, we use TF-IDF generate word features and non-negative matrix factorization (NMF) to reduce the dimensions of the vectors. Then, we grouped the word features under different topics. We labeled sentences with extracted topics and then generated sentiment intensity score, using those scores as predictors and each user's star as response to analyze the relationship between stars and topics through linear regression. By analyzing the coefficients of linear regression, we can find out which topic influence the stars most.

Dataset

The first step is to select a subset of the dataset from the Yelp Academic Dataset containing 6,685,900 reviews and 192,609 businesses. In this case study, we only focus on restaurants in Las Vegas, Nevada. We first applied our model to one restaurant, Hash House A Go Go (5,847 reviews), then after validating our model we applied it to the dataset with the restaurants grouped by their postal code. We did our analysis on five of the postal codes in Las Vegas surrounding the famous Las Vegas Strip (89103, 89109, 89118, 89119, 89169).

Summary of Key Findings

By extracting topics, we found out that when people giving reviews they care about the overall experience, price, food, service and if the restaurant is worth a try. After applying linear regression with mean square error of 1.034 and AIC of 12958, it shows that among all the topics service influences the number of stars the most, with the second influence being the food. We also noticed that the restaurants near the Vegas Strip (89109) showed a significant difference from others. Compare with service, food contributes more to the overall stars in postal code 89109.

Case Questions

i) Describe some basic and advanced data cleaning and preprocessing steps that you believe are necessary for effectively working with restaurant reviews. How do these steps differ from what you would do if you were instead processing comments on social media?

ii) Could topic modeling be used for discovering the different factors (like food and service) that influence the reviews? Describe in depth how you would discover such factors and what would be ways to validate them.

iii) What is the difference between MSE (mean squared error) and Akaike Information Criterion (AIC)? Give the formulae and discuss why we may want to use one or both of these for our problem setting.

iv) What is the role of sentiment analysis here? Can we attempt a solution to the problem without it?

v) Rather than use tf-idf followed by NMF, discuss a more principled approach to directly get low-dimensional vectors. Are there advantages to using this other approach compared to the default approach we used (tf-idf with NMF)?

vi) Suppose you had to scale this model to the entire country. Would you train a single model? A different model for each city or state? Discuss what your strategy would be, and how your validation approach would change.