Case Study: How Keywords in California Restaurant Reviews Influence Ratings

DS 4002 - CS 3 - Mohini Gupta

Purpose

Do you like traveling? What factors influence which restaurants you go to when you're in a new city? If you're in a state with many tourist attractions, and therefore a lot of restaurants wanting to attract those tourists, chances are you're going to use overall restaurant star ratings to guide your decisions. But have you ever wondered what features of a restaurant influence the rating that it gets and why people rate restaurants differently? This case study will introduce you to the idea of analyzing text data, and for this case study, you will learn about which keywords are linked to certain restaurant ratings and be able to understand how overall sentiment changes depending on the rating.

Context

You are a young data scientist that loves travelling, and you are a major foodie. You are about to go on a 7-day road trip in California, and you want to hit up the best restaurants in California while you are traveling. But, part of you is unsure if you can trust the opinions of people online. You want to understand what influences people to rate restaurants differently, and you want to verify that overall sentiments for restaurants increase as the rating increases.

Your task

You will be using Yelp's open dataset for this case study. After downloading the dataset, you will filter for reviews in California. Then, you will remove stopwords from the reviews. Next you will identify words that appear most frequently in both low and high rated reviews. Lastly, you will find out what the overall sentiment scores are per rating.

All materials for this project can be found in the GitHub repo:
https://github.com/rde6mn/DS4002-CS3