Popular Items from Yelp Reviews

Ashwin Sura Ravi | Dileep Kumar Gunda | Suryakiran Suresh Gurumoorthy Iyer

Problem Description

We propose an end to end method to identify popular food items of a restaurant from the sentiment of reviews, for example: did a user like an item and how many liked it! Yelp dataset contains millions of reviews for thousands of restaurants but this doesn't include any information about the menu or items of a restaurant and each review might talk about multiple items. This makes the problem interesting and difficult to solve.

Motivation

Finding popular dishes in a restaurant has been a long-standing problem. There exist numerous applications that rate a restaurant but give little information about what items the users liked. It's a common dilemma everyone faces at new places and even at old places with never tried items on the menu. With the rapid growth of the internet and mobile phones, Yelp has accumulated millions of reviews for thousands of restaurants which makes this problem of recommending popular dishes of a restaurant feasible. There is an information overload in today's world and extracting useful information makes it easier for everyone.

Datasets

We are planning to use Yelp dataset. This contains 6.6 million reviews from 192 thousands of restaurants across at least 10 metropolitan areas. This doesn't include menu items for each restaurant. So, we are going to mine the menu from the reviews and additionally use the New York public library's food dictionary containing menu items since the 1850s to filter out items. For the scope of this project, we will consider a subset of this data like confined the scope to one metropolitan area. Each review is associated with a rating, and this can be used to categorize review as either positive or negative.

Approaches

We approach this problem in two phases. The first phase is to extract food items like 'Alfredo Chicken' from the reviews. The Yelp dataset doesn't include menus for restaurants, which makes it tricky. One can adhere to manual extraction for higher precision, but considering 100,000 reviews makes it a very unfavorable option. Another way is extracting through Named Entity Recognition (NER) techniques. NER is associated with true negatives and false positives like it might capture too much like 'Alfredo Chicken Large' or a partial name like 'Chicken'. These can be rectified after NER extraction with less effort by expert feedback, as the extracted list would be smaller than the original corpus.

The second phase is to rank this extracted list of items for a particular restaurant. One of the approaches is to calculate TF-IDF of these items in all reviews of a restaurant and rank them. This approach doesn't distinguish most liked items from most disliked items, cause it just deals with the count. An alternative approach is considering the sentiment of the review towards a particular item, which identifies the most liked items from most disliked and weights them appropriately. Each review can talk about multiple items like 'the chicken is good but veggies are bad'. The central problem to solve in this phase is to break the review into segments that talk about a particular item.

Evaluation

Once the NER predictions are available, the truth values of an item to review would be gathered by crowdsourcing. This is done since the ground truth is not readily available. The evaluators would be shown restaurant reviews and asked to pick popular dishes in a particular restaurant from among the NER picked items for that restaurant. A subset of this is selected as Train and while rest is considered as the Test set and used compared with the test results from our model.