CASE STUDY: TEXT ANALYTICS ON HOTEL REVIEWS

In this case study, we will go deeper into using text analytics on reviews posted on websites such as TripAdvisor.

Opening prompt

The hotel industry is rapidly changing, and customer reviews play a crucial role. According to a survey done by TripAdvisor, 87% of travelers consider reading reviews important and 78% customer will read reviews to get a fresh perspective before making a decision. Votes on a review are beneficial for both hotel owners and customers. Reviews with good number of votes are advantageous to hotel owners as they can display or feature them on their website and attract more customers. Likewise, if a review is negative and is likely to get votes they can take remedial steps and pacify the customers.

Initial steps

The first step is to explore the data. Let us assume that the reviews data (there are several publicly available corpora online, but you could also scrape small quantities of data from websites like Yelp and TripAdvisor) has several fields, including hotel class, name, an identifier, address, as well as the text of the review(s). A basic exploration could involve statistical profiling (e.g., is the dataset biased toward five-star hotels?) but a more advanced analysis could involve some of the machine learning techniques you have learned in class, including word-clouds.

Below, we use a real-world reviews dataset from Yelp to show the main regions (states) in the US where the hotels in the dataset under study (this is not necessarily representative of Yelp) are present. We find that there is a rough, but not perfect, correlation with population.

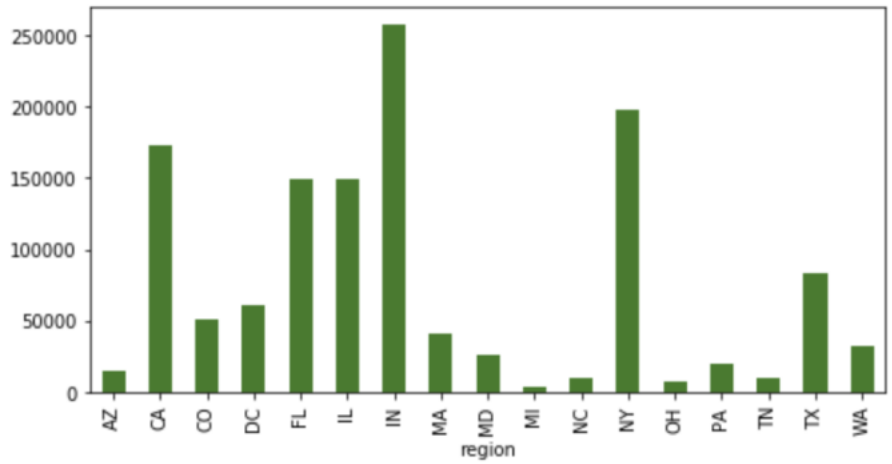

Figure 1: Distribution of hotel breakdowns by region (states in the US) in a real-world Yelp dataset.

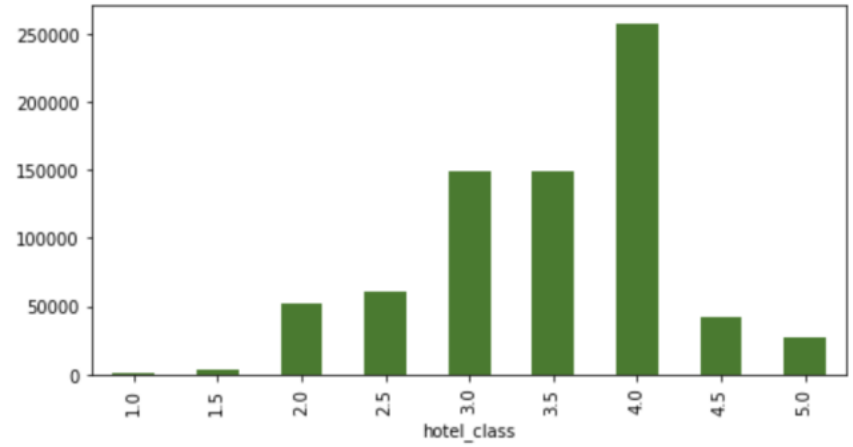

Figure 2: Distribution of hotel classes (in increments of 0.5 stars) in a real-world Yelp dataset.

Review Analytics

The next step is to develop text analysis classifiers to systematically study the review text. One way to do so is to consider a set of 'explanatory' variables such as service, cleanliness, value or location (including whether the hotel is based in a nice part of the city, by tourist attractions, or by the beach). In the case questions, we ask you to define such a set of variables.

How might we train these classifiers? One way is active learning, where we first identify a few examples (through eyeballing, or using clues in the review text), train an initial classifier and then use the examples that are most confusing to the classifier to progressively label more data and re-train the classifier.

Testing the models

Once you have your text classification models, you need to test them. Since you do not have a ground-truth, you could do post-hoc labeling, where you randomly sample some outputs and check whether you agree with the classifier. Another option is to independently label some data and evaluate the classifiers on these data. No matter what you do, make sure to verify that neither your test data nor your classifier are biased in some way that makes the metrics suspect. Checking for statistical significance, especially compared to some simple baseline performance, is also vital, although you may not always have enough data to get a significant result. Regardless, it is a good practice.

Advanced analysis

Another kind of advanced analysis would be to use the variables you selected to *explain* the final overall review that a user may have given the hotel (for example, 4 points on a 5-point scale). You can accomplish this by doing a regression. One thing to note is that your independent variables are themselves predicted and subject to noise. Furthermore, they may not be enough to explain the single review score. Another possibility is that the regression may be inadequate as it is linear. Only through careful experimentation and analysis can these be understood in the context of a specific dataset.

Case Questions

i) We suggested in *Initial Steps* that one reason to do data exploration is to look for certain biases that might be present. For example, the dataset might be over-represented in the number of luxury of 5-star hotels. Suppose you found that 50% of the hotels in the dataset were 3-star, and the other four classes (1-star…5-star, not including 3-star) were equally divided among the other 50%. Would you necessarily conclude the dataset was 'biased'? Why or why not? What other information (such as the city) would you need to determine this? *Hint: In a resort town, would you expect hotels among all the classes to be equally divided?*

ii) How do you explain the distribution in Figure 1? What variables might be most closely correlated to the numbers you are seeing? *Hint: Think of economic variables, but you may also have to get more creative.*

iii) Does the distribution in Figure 2 look normal when you think about a sufficiently large and broad set of hotels (e.g., global or US national)? How would the distribution shift if you were instead plotting it for a luxurious destination like the Cayman Islands or Bahamas? What about a poorer region of the world?

iv) What are some variables that you could train text classifiers for that you believe can be 'mined' from the review text? Name the variable and also state whether (i) it is discrete or continuous, (ii) its domain of values.

v) Next, look up three hotel reviews on Yelp and manually ag those reviews with optimal values for those variables. Did you have more difficulty with some variables compared to others?

vi) Discuss what kinds of classifiers you would use to build models for these variables and with what features. Hint: Would generic text-based features like tf-idf always be enough or should we give more importance to certain variable-specific keywords? What about word embeddings? Would there be an advantage to using word embeddings over tf-idf in this context (think about the active learning suggestion presented earlier in the case)?