CSCI 5622 - Machine Learning

Feature Recommendation-based Hotel Popularity Improvement

Nithin Veer Reddy - nithin.kankanti@colorado.edu
Sankaranarayanan - sana6469@colorado.edu
Lokin Sai - lokin.makkena@colorado.edu

1 Acquiring Data

The data that we have acquired until this stage consists primarily of hotel reviews and ratings. It also contains other secondary information like latitude, longitude, trip type. All of this information is extracted in json format which is readable and easy for implementation purpose. Going through this data format, few people have given sub ratings on individual components, like service and food. This can be used as supervised data and can be easily extracted.

But most of the reviews don't have individual ratings on components listed. So there is a need to extract these components from the review. Also we need to extract the information that this component encodes. For ex. if the review reads 'Service is good.', we need to extract information related to the service being good.

2 Proposed Method

To extract important components, we try out the below method.

Extracting components using Part-of-Speech Tagging

This method starts out with extracting the text part of the 'review' and passing it to VaderSentiment. VaderSentiment gives us a polarity score dictionary. This gives us a proportion of the 'positivity' and 'negativity' in the review which can be used to arrive at the sentiment of the review. This sentiment has the potential to be used as the label of the review.

From here on, we use StanfordCoreNLP for extracting the components from

the reviews. We set up the StanfordCoreNLP Server and access it through API. We break up the reviews into sentences. As we are breaking them based on the period, we also need to perform a few basic regex operations in order to ensure the format is correct. The reason we are perfoming the regex operation is because the sentences start with a space after a period. This is why we need to eliminate the space and extract just the text of the sentence, which is done using regex operations.

Now we proceed to Part-of-Speech Tagging using standard functionality in StanfordCoreNLP. Our aim here is to identify adjectives that are associated with nouns. We believe that components mostly fall under the category of nouns. The information that the component encodes can be extracted by the adjective that is associated with the noun. If there is any adjective associated with a noun, then we would take that noun as a component.

So for now, the components being identified, we need to give the components a 'value', more specifically a rating. This is done by passing the sentences into VaderSentiment to get the overall sentiment of the sentence. After this, we use a weighted measure of the 'positivity' and 'negativity' of the sentence, based on which the rating for the components are decided.

After all this, we have managed to extract the components from the review and alon with the rating of the hotel combined, we have managed to create supervised data.

3 Moving Forward

Moving forward on the project, the team has decided to go forward using feature selection techniques to get important features from the data. Feature selection methods can broadly be categorized into the filter, wrapper and embedded methods. We have planned to implement feature importance methods from each of these categories in moving forward. For instance, Pearson Correlation, Recursive Feature Elimination, SelectKBest, and PCA. Later, we will use important features to train a model and predict the label. More information on which models to be picked for training will be discussed in detail and will be stated in the final report.

4 References

[1] *Julio-Omar Palacio Nino Fernando Berzal Evaluation Metrics for Unsupervised Learning Algorithms*

[2] *VADER Sentiment Analysis*

[3] *TripAdvisorAPI https://developer-tripadvisor.com/content-api/documentation/*

[4] *Stanford CoreNLP https://stanfordnlp.github.io/CoreNLP*