Capstone #2 Project Proposal

The problem that will be solved in this project is predicting a positive or negative sentiment given a yelp review and restaurant attributes. The yelp review will need to be converted to a vector so that it can be inputted into a machine learning model.

The client is any review site that allows for users to input a text and star rating for an item. A client would find conversion of text to a positive or negative review as more intuitive for the user because he/she does not need to go through the process of giving a star rating that has five options and does not give viewers a clear understanding whether the subjective star rating represents a good business. For example, some reviewers may consider a 3 star rating as positive while others may consider it negative. Sentiment analysis is important because many times a review contains sarcasm, ambiguity, or a play on words that can be misconstrued as either positive or negative but positive or negative label can give more information to a user. A statement that simply states if a user liked that restaurant or did not based on his/her text review is more realistic in a conversation where a person will more likely describe a restaurant based on it whether they liked it or not, not by star ratings. Yelp will be able to forgo the star rating system and just let users input a text review and then yelp's overall score of a restaurant would be the percentage of users that liked the restaurant.

Problem solution:

1.) Define what is a positive or negative review based on distribution of star ratings.
2.) Remove punctuation and numbers from all reviews so that only words are left.

3.) Lowercase and then Tokenize sentences because word2vec accepts lists of sentences.

4.) Don’t remove stop words that occur frequently in the reviews, word2vec uses these words for broader context. Punkt tokenizer.

5.) Convert to vectors using unsupervised word2vec.

6.) Two options for transforming for supervised model. Each word is a vector is 300 dimensional space, so we can average the word vectors in a review so that each review has the same vector length as review length is variable. Other is to create k clusters that is ⅕ of review vocabulary size with an average number of words per cluster as 5. Then do a bag of centroids that produces a numpy array with the length representing the number of centroids and the elements as the number of words in the the cluster.

7.) Used supervised ml: logistic regression, random forest, deep learning model, and gradient boosted tree model to predict sentiment.

The data can be found here:

https://www.kaggle.com/yelp-dataset/yelp-dataset#yelp_academic_dataset_business.json

The data will be downloaded in .json format and uploaded into a pandas dataframe for tabulated data. A subset of the data will be used to decrease prediction time. The biggest city in the dataset, Las Vegas, will be chosen as well as only restaurants in Las Vegas as the business looked at in order to decrease the sample size of the number of reviews. The features used for prediction will mainly be the restaurant review. Overall restaurant attributes, such as number of reviews, type of restaurant, and operating practices will be added as predictors to the model if they are shown to correlate with restaurant sentiment. More specifically some of these restaurant attributes include review count, whether the restaurant is open, restaurant price range, and type of restaurant.

The deliverables for this project include two jupyter notebooks, a paper, and a slide deck. The first notebook will contain the EDA of the project and the second notebook will contain the machine learning section of the project.

Instructions

- What is the problem you want to solve?
- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?
- What data are you using? How will you acquire the data?
- Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
- What are your deliverables? Typically, this includes code, a paper, or a slide deck.