LING539 FINAL PROJECT PROPOSAL: WINE REVIEW ANALYSIS

Name: Hanh Nguyen

1. Overview

In this project, I will explore the Wine Review dataset on Kaggle (scrapped from WineEnthusisast: https://www.winemag.com/?s=\&drink_type=wine) from the following link:

https://www.kaggle.com/zynicide/wine-reviews/data\#winemag-data\first150k.csv

More precisely, I will build a model to predict the quality of wine based on the text review in this dataset. This would be a sentiment analysis problem and hence it can be re-casted as a classification task.

2. Motivation

I would like to try to use NLP technique and standard machine learning algorithms (logistic regression) to apply into a specific problem. Sentiment analysis is a standard problem in NLP and it is worthy exploring this in the dataset Wine Review I mentioned above. It would be useful to understand customer’s reviews in depth and predict valuable information based on these. Wine quality is clearly one of the main concerns for customers to decide to purchase, so this is a good motivation to pursue this classification task using text description of the reviews for this wine dataset.

3. Anticipated challenge

The model will use data from both numeric features as well as the NLP-processed text features. The most challenge is to preprocess and transform the text to numeric types in order to feed in the training machine learning process, such as using CountVectorizer or tf-idf. The size of the dataset is not small as it has more than 150,000 reviews (observation) and while fitting and transforming with vectorizer it would have a large number of features, so preprocessing the text is critical to boost the performance of the predictive model. This step is probably the longest step as it requires me to explore and extract valuable information by removing unnecessary things such as stopwords and non-alphanumeric characters. Another challenge is that this is a multi-class classification problem. Therefore, I need to examine the data carefully to decide the weights of each class rather than the default choice of making weights equally. After that, choosing a model to predict wine quality is also another difficult task. Logistic regression for multi-class problems is my choice and tuning the hyperparameters is also important. Also picking a suitable optimization solver is important too. The result could serve as a baseline benchmark for more sophisticated methods such as deep learning (Word Embedding and Recurrent Neural Network layers). For these approaches, it is anticipated that they would require a computer with GPU to run and I do not have it, so this is another challenging.

4. Evaluating my system

To evaluate my system, I run the model on another independent part of my data set, i.e., the development set. I will check the accuracy metric as a first evaluation. As this is a multiclass classification problem, I will use confusion matrix to evaluate my model in depth. I would compare the unigram and bigram to select the one with better accuracy and confusion matrix coefficients.

Timeline

Task	Timeline
Exploring and visualizing data	March 10, 2020
Preprocessing text in reviews: tokenizing, removing stopwords, punctuation, lowering case letters, removing non-alphanumeric characters, etc	April 10, 2020
Data splitting to train and test set	April 20, 2020
Preprocessing text in reviews: tokenizing, vectorizing … the text to transform text to numeric data for machine learning algorithm	
Picking a model for multi-class classification task. Tuning hyperparameters.	April 30, 2020
Wrapping up codes, writing report/slides and doing a video presentation	May 06, 2020