LING539 FINAL PROJECT PROPOSAL: WINE REVIEW ANALYSIS

Name: Hanh Nguyen

1. Overview

In this project, I plan to cluster similar wines based on the provided description such as tasting reviews and their winery, region where grapes were harvest, price and so on, by using an unsupervised learning algorithm. In a secondary analysis, I want to build a predictive model to identify wines.

2. Motivation

I'm curious to see the capability of the machine learning algorithm to cluster wine groups based on similar features of the wines.

3. Anticipated challenge

This is a multi-label-multi-class classification problem. The model will use data from both numeric features as well as the NLP-processed text features. The most challenge is to transform the text to numeric in order to feed in the training machine learning process, such as using TF-IDF to find document similarity, word vectorization along with lot of regex and matching pattern

4. Timeline

Task	Timeline
Exploring and visualizing data	1w
Preprocessing text in reviews: tokenizing, vectorizing … the text to transform text to numeric data for machine learning algorithm	2w
Data splitting to train and test set	1w
Unsupervised learning using K-means clustering	
Choose model for multi-class-multi-label classification	1w
Build a pipeline	
Report and presenation	1w