Stat-495 Revised Proposal

Emily Ye & Shukry Zablah

2019-11-15

Team Name: YZ Analytics

Title: Wines Around the World.

Purpose:

A lot of people want to know about wine but don't know where to start. We are not wine experts, but we can leverage the experts and what they have to say to form an interactive, state of the art explorative visualization for the laity. In class we take visualization to its bare minimum. This project will try to embrace new visualization technologies in the context of learning more about wine, and present a couple of techniques that branch off from the curriculum such as working with extracting features from text.

Data:

We will be working with Kaggle's wine review data, found here: https://www.kaggle.com/zynicide/wine-reviews (Thoutt, 2017). There are close to 130,000 wine reviews.

The original data set poster also included his code for scraping the data off of the WineEnthusiast website, so we will also explore to see if his code still works and scrape the data ourselves so we have more up-to-date wine reviews and more data. There are currently about 270,000 reviews on WineEnthusiast (WineEnthusiast, 2019).

Variables:

Variable descriptions from the data set:

- country - The country that the wine is from
- description - A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- designation - The vineyard within the winery where the grapes that made the wine are from
- points - The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- price - The cost for a bottle of the wine
- province - The province or state that the wine is from
- region_1 - The wine growing area in a province or state (ie Napa)
- region_2 - Sometimes there are more specific regions specified within a wine growing area (i.e. Rutherford inside the Napa Valley), but this value can sometimes be blank
- variety - The type of grapes used to make the wine (ie Pinot Noir)
- winery - The winery that made the wine
- title - The title of the wine review, which often contains the vintage if you're interested in extracting that feature
- year - pulled from the vintage in the title
- taster_name - name of the taster
- taster_twitter_handle - Twitter handle of the taster
- latitude - to be geolocated
- longitude - to be geolocated
- altitude - to be derived

We intend to use some sort of text analysis and text mining to create additional quantitative variables from the description, in particular using Bag of Words (BOW) representation to calculate the Term Frequency-Inverse Document Frequency (TF-IDF) for the unique words of the description (Nabi, 2018). We can achieve this calculation through the `text2vec` package ("Predicting," 2018).

Model:

The model is not the focus of our visual exploration. However, we wish to seamlessly incorporate a predictive model for new wines not in our dataset that predicts multiple categories at once, especially price and quality. We are thinking of clustering, maybe something similar to k means.

End Product:

We are envisioning a web application with a lot of maps and actions on clicks. When clicks happen on certain countries with wineries we have data on, we want to display some summary information and graphs that are centered around teaching the broad strokes of wine type and quality. Additionally, we also want to provide search capabilities of a catalog. All of this under the umbrella of visualization.

References

Nabi, Javaid. (2018). "Machine Learning - Text Processing." Towards Data Science. https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958 Thoutt, Zack. (2017). "Wine Reviews." Kaggle. https://www.kaggle.com/zynicide/wine-reviews 2018. "Predicting Wine Ratings Using LightGBM + Text2Vec." Kaggle. https://www.kaggle.com/nnnnick/predicting-wine-ratings-using-lightgbm-text2vec 2019. WineEnthusiast. https://www.winemag.com/?s=&drink_type=wine&page=0