Final Project: NLP analysis for Wine Taster's review.

Team: Group 3 Fireflies

Roles:

Data cleanup and exploration: Chunhui Zhu,

Data Modelling: YueChun Wong, Chunmei Zhu

Overview

We are interested in the wine review dataset to perform NLP analysis. We also try to figure out any insights such as wine preference or bias among the review.

https://github.com/czhu505/Data620-Fireflies/tree/master/Final%20project

The data was scraped from WineEnthusiast on November 22nd, 2017. The code for the scraper can be found https://github.com/zackthoutt/wine-deep-learning.

The author collected the title of each review, the tasters name, and the taster's Twitter handle.

Data

We have the 11 attributes in 37420 samples in the data set.
https://www.kaggle.com/zynicide/wine-reviews

Column names:

Country - The country that the wine is from description

Description - example “Raw black-cherry aromas are direct and simple but good. This has a juicy feel that thickens over time, with oak character and extract becoming more apparent. A flavor profile driven by dark-berry fruits and smoldering oak finishes meaty but hot.”

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 Rutherford inside the Napa Valley), but this value can sometimes be blank

Taster_name: 16 unique tasters

Taster_twitter_handle:

Variety: The type of grapes used to make the wine (ie Pinot Noir) :

Winery: The winery that made the wine

Project Plan:

Data overview

- # of wine
- # of tasters
- # of reviews

1. Sentimental analysis 1: Determine what is the sentimental score for each wine review by NLTK's package (e.g. SentimentAnalyzer/NaiveBayesClassifier)

2. Sentimental analysis 2: Build up a sentimental keyword list for wine based on good/bad rating (e.g. light is negative in wine description while it is positive in NLTK's package)

3. Taster favors for wine: based on taster's keywords and review score, identify taster's preference of wine / bais.

4. Taster's commons and links in network graph

5. The top influencer on social network for wine (# of followers vs # of review)

6. Wine and grapes type relationships network graph

Concerns:

There are many wine specific vocabularies that we may require us to look up. Non-english words could be troublesome in textming and clean up. Also, pronouncing the wine type in presentation could be challenging but educating and interesting to us.