Wine Recommendations

By: Rachel Goodridge

Abstract

The purpose of this model is to create a wine recommendation system for wine enthusiasts of all experience levels that allows users to explore new options and effortlessly make decisions about their next wine purchase. The dataset includes 49,546 unique wine tastings and descriptions written by sommeliers. Using text preprocessing, dimensionality reduction, and topic modeling, I created a content-based filtering system that can be used to find the closest match to any given wine in the dataset. The search can also be narrowed, based on user input, such that the algorithm only returns wines with the specific features requested by the consumer.

Design

It can be difficult to decide upon the next wine to drink and enjoy, regardless of wine tasting experience. This model can take some of the guesswork out of trying a new wine by creating user-specific recommendations based on past interests. The idea is to effortlessly provide options and assist with or eliminate the need for decision-making as well as introduce the user to more possibilities.

Data

The dataset can be found on Kaggle and each row represents a unique wine tasting with a description by a sommelier and various wine features. There are 49,546 rows and 10 columns including country, description (a text document of around 42 words on average), designation, points (a score out of 100), price, province, region, taster name, variety, and winery.

Algorithms

First, I cleaned the corpus using regular expressions to remove numbers, letters, and punctuation. Then, using the SnowballStemmer from the Natural Language Toolkit package in Python, I trimmed/modified all words to form their root and removed English stop words as well as some specified wine-related terms. Next, I tried both the CountVectorizer and the Term Frequency Inverse Document Frequency Vectorizer from scikit-learn to create a document-term matrix with both unigrams and bigrams. The topic modeling algorithms I attempted include LSA and NMF from scikit-learn, LDA from gensim, and corextopic. I chose to use CorEx and anchor specific words to create the topics for this model. Lastly, I performed content-based filtering by calculating pairwise distances, collaborative filtering using LightFM, and created a hybrid recommender also using LightFM.

Tools

- Pandas and numpy for data manipulation
- Matplotlib and seaborn for visualizations
- Re, string, and nltk for text preprocessing
- CountVectorizer and TfidfVectorizer from sklearn for dimensionality reduction
- TruncatedSVD and NMF from sklearn for topic modeling

- LdaModel from gensim for topic modeling
- Corextopic for topic modeling
- LightFM for collaborative filtering and hybrid recommender

Communication

The topics that were revealed with CorEx include citrus, berry, woody, vinification, varietal, body, spicy, and miscellaneous. Then using pairwise distance, the algorithm can produce some decent wine recommendations given that the user has chosen a wine they like from the dataset. These wines (the one the user has previously enjoyed and the one being recommended to the user) generally fall into similar topic categories and even share many of the same words in their textual descriptions (see below for an example recommendation with similar words highlighted). Furthermore, the model can allow for additional user input if the consumer wishes to specify a country, point range, price range, province, variety, or winery. Then the search will be narrowed and the user will only receive wine recommendations that match those feature requests. On the other hand, it turns out that collaborative filtering (and thus, the hybrid recommender) was not a good fit for this dataset. Although there is a wine rating system, the wines that the sommeliers tasted were nearly completely unique and there was little-to-no overlap between the tasters' preferences. Thus, the algorithm struggled to find similar users and the average error for each prediction was very high. In conclusion, the nature of the data made it unfit for collaborative filtering, but content-based filtering proved to be a success.

Wine the User Likes

Description:

Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby.

Wine Recommendation

Description:

Concentrated blackberry and black currant aromas are ripe and come with a balsamic note. It feels juicy, full, lively and fresh, with oaky dark-plum flavors accented by rooty, herbal notes. Dry tannins mark the finish.