BT 5153 Group Project Proposal

wine recommendation system

Cun Xiaofei(A0186051J)
Hong Li(A0186004M)
Liang Shijie(A0186001U)
Liang Xinran(A0186708R)

1. Purpose of the Project

The main purpose of this project is to develop an automatic wine **recommendation system** for wine based on reviews of professional wine tasters. The input of the recommendation system is the aroma, taste, price interval and other possible features getting from text mining on reviews, and the output is the basic information of recommended wines and an one sentence description of each wine.

Singapore's wine market is predicted to reach 1 million US dollars by 2021 as the result of the country's growing economy and increasing consumer spending. But at the same time of wine's increasing sales volume, it is not as easy to select by consumers as other FMCG because of its complicated variety constitution. Different wineries, different grape breeds and different years of making are jointly affecting wine's taste. How to help consumer select wine product which matches the consumer's personal taste is a significant problem for online wine sellers who seek to enhance customer experience.

To solve this problem, we try to generating an automatic wine recommendation system. Words describing aroma and taste of wine are captured from reviews to create labels for each wine. The automatic wine recommendation system are built from these labels as well as a price interval set by users, and the one sentence description shows in the output helps users better understand the recommended wines. This recommendation system can be applied to online wine retailing platforms such as WineExpress.com and yesmywine.com.

2. Dataset description

The wine dataset (*http://www.kaggle.com/zynicide/wine-reviews*) is chosen from Kaggle . The data of this dataset was scraped from a wine reviewing and ranking website *Wine Enthusiast* during the week of June 15th, 2017. The dataset contains 129,970 rows and the following 10 columns:

- **Points:** the number of points Wine Enthusiast rated the wine on a scale of 1-100

- **Title:** the title of the wine review(which is regarded as the name of wine)
- **Variety:** the type of grapes used to make the wine (i.e. Pinot Noir)
- **Description:** a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc. This is the main column for text mining.
- **Country:** the country that the wine is from
- **Province:** the province or state that the wine is from
- **Region 1:** the wine growing area in a province or state (i.e. 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
- **Winery:** the winery that made the wine
- **Designation**: the vineyard within the winery where the grapes that made the wine are from
- **Price:** the cost for a bottle of the wine
- **Taster Name:** name of the person who tasted and reviewed the wine
- **Taster Twitter Handle:** Twitter handle for the person who tasted and reviewed the wine

One highlight of this dataset is the aroma and taste description of wine which can captured from tasters' reviews. This information seldom appear in other sources, but it is of great importance for consumers to correctly select wine by their preference of flavor. Moreovers, this dataset recorded various information of wine in detailed, which can be applied to multi-purposes. Some further explorations of the dataset are listed in the last part of the proposal.

3. Exploration Steps

3.1 Data preprocessing

Firstly do data cleaning, and in order to fully understand our data set, we use **Tableau** to plot the data.

3.2 Getting features from aroma and taste description words

According to our hypothesis, wine's aroma and taste are of great importance in wine recommendation. As a result, using keywords that describe aroma and taste as features to build our recommendation model is reasonable. Below are steps to get our features.

3.2.1 Extract words describing aroma and taste

Firstly, we union all the reviews and extract all words that describe taste and aroma. In order to get the most accurate words set, we use two method: **LDA** and **word2vec** to get two different words sets.

3.2.2 Get the features

After getting the words sets of aroma and taste, calculate the **word importance** of each word based on the principle of **TF-IDF** and use the value as features.

3.3 Getting the recommendation target

Since the data set contains more than 10 thousand wines based on the *title* column, it is difficult to regard every wine as a unique type. After consulting some wine classify papers, we decide to category all the wines according to **their year of making, grape breed and the country the wine** is from to get our recommendation model target.

3.4 Select features

We separate the whole data set into train data set and test data set. Then we build our model using some algorithms, such as **XGBoost** and **Naive Bayes,** and conduct feature selection based on **feature importance**. Finally, we re-build the model using selected important features and test its accuracy.

3.5 Adding more features

In this step, we try to find some extra important and meaningful features through other text mining methods such as **Topic Modeling** and **Sentiment Analysis** to perfect our recommendation model furthermore. Then we use the same train and test data set above to test the new model's accuracy.

3.6 Generating one sentence description

In order to giving our customer better using experience, we also try to provide one more **abstract description** about the wine we recommend. **Document Summarization Method** is used to generate an one sentence description for each wine tetle.

4. Text Mining Methods Related

4.1 LDA

The intuition behind the LDA topic model is that words belonging to a topic appear together in documents. It can infer topics based on word counts and the bag-of-words representation of documents. In this case, we can extract topics and similar words from whole reviews to form features.

4.2 Word2Vec

Word2Vec is one of the popular methods in language modeling and feature learning techniques in natural language processing (NLP). This method is

used to create word embeddings in machine learning whenever we need vector representation of data. In this case we use Word2Vec to cluster similar words. Since we have already decided to use ‘aroma’ and ‘taste’ as two kinds of feature groups, we can get similar words together by capturing the distance between individual words with the help of Word2Vec.

4.3 Feature Importance and Feature Selection

Tons of features can be burdensome and lead to high variance. After using ensembling models to predict the wine category, we will use its feature importance function to figure out which features have significant impact on deciding the category of wines in order to make feature engineering.

4.4 Document Summarization

The main idea of summarization is to find a subset of data which contains the "information" of the entire set. In this case, we try to select a subset of existing words, phrases, or sentences in the original text to form the summary. After that, we will build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express.

4.5 Other Methods

Methods such as **Dictionary Based, Tokenization**,**Topic Modelling and Sentiment Analysis** will also be used in this case.

5. Further Applications

5.1 Verify the correlation with wine enthusiasts ranking and customer reviews

Using crawler in Python to get customers’ wine purchasing and reviews information from Amazon or other online wine selling websites. Information consists of 4 parts:

- **Category:** the category of wine
- **User ID**: customers’ ID
- **Reviews**: customers’ reviews after consumption
- **Ranking**: The number of points customers rated the wine

For a certain category of wine, verify whether the wine taste and aroma from customers’ reviews are the same as what wine enthusiasts evaluated. And also, the ranking given by winetasters is supposed to be similar with customers’ evaluation.

5.2 Verify the importance of grape variety and vinery for wine quality

The wine type which got high rate from tasters is supposed to be produced by some certain range of wine growing area, type of grapes and venery. In other

words, the source of grapes, variety and vinery are correlated with ranking of wine.

5.3 Figure out which feature contributes to wine prize

Conduct a simple linear regression for grape variety, vinery, vintage year and final wine prize to see which feature contributes most to wine prize. We have made a hypothesis that the fame of vinery is able to give a high premium for the prize.

5.4 Test the price elasticity of customers

Evaluate whether the price is the most important feature that customers consider when they purchase wine. And also, whether the price elasticity goes down when the actual wine price exceeds a certain value, because those rich man are likely to choose a high quality wine without considering its price premium.