Detecting wine taste using Recommender Systems

Geronimo Wolf
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
g.wolf-1@student.utwente.nl

ABSTRACT

Wine comes in a wide variety of types and has many different flavours. Of all human senses, taste is often seen as the most personal of all. It is unique and helps determine what food to eat or what drink to consume. Consequently, complex beverages as wine are not perceived in the same manner by everyone. Getting an understanding of wine and of one's pallet can be difficult and very time-consuming. This paper investigates the ability of a recommender system to detect taste and whether it can give wine suggestions correlative to taste preferences. This paper also explores the data that is required to create a wine recommender system and considers how this data is collected. In addition, this research addresses a language interpretation problem. The results of the recommendation techniques used during this research gave useful insights into the dataset required for such a recommender system and allow future possibilities for this approach to be explored.

Keywords

Taste, Wine, Suggestion Algorithm, Recommender Systems, Wine Recommender System

1. INTRODUCTION

One's preference of wine differs from person to person, and developing the ability to learn about the different flavours can be challenging and time-consuming. The evaluation of taste is often a product of social processes, rather than objective characteristics [23]. Even a professional wine taster (sommelier) can misjudge a wine or be biased. A recommender system that suggests wine to users could properly tackle this issue and allow people to enjoy wines they would perhaps otherwise not have been able to enjoy due to their lack of wine knowledge.

Taste is, arguably, the most personal and complex sensation and makes the most private of connections to the material world [9]. Ferguson [9] described it in his 2011 paper as "describing the indescribable, measuring the immeasurable, generalising the singular". Taste also is the most essential element of wine quality and the biggest factor influencing consumer liking [5]. From all five taste modalities, the three largest contributors in wine's flavour are sweetness, sourness, and bitterness [5].

Suggestion algorithms, also known as Recommender Systems, attempt to predict preferable suggestions to a user. These systems are used by all major websites to filter content and make suggestions to the users in a personalised manner [11]. There is a large demand for recommender systems from businesses due to the increase in user engagement. The recommendations customers receive on their phone are faster and more effective than what could be achieved in person by an employee.

Recommender systems have different techniques, often separated into the following main three categories:

- *Content-based filtering*: Recommendations will be determined by the user's preference of items (wine) in the past and suggests similar items [1].
- *Collaborative filtering*: Recommendations will be determined by suggesting items that users with similar preferences have enjoyed [1].
- *Hybrid approaches*: Recommendations are determined by a combination of content-based and collaborative methods [1].

Collaborative filtering systems are extremely useful but often suffer from the cold start problem. There are three cases of cold start [22]. The first case occurs when there are no (or little) users present at the start, which makes it difficult to provide reliable recommendations. Secondly, a new wine that is added to the system will not have been rated yet by users and thus cannot be recommended to users. Lastly, a new user in the system will not have any data yet and therefore it will be impossible to provide personal recommendations. Creating a wine recommendation system with collaborative filtering would require a large user base with ratings to be relevant and give decent suggestions. More user preferences will result in more accurate results. Such data is not publicly available due to its potential commercial applications.

Contrarily, content-based filtering requires a great number of items with distinguishable features. Content-based filtering therefore also does not suffer from the cold start problem. However, if a user does not have any prior personal data, there needs to be an alternative initialisation of the system. To give suggestions, each wine needs useful information and features to distinguish itself from one another. Features of wine include a general description of the wine, location of the vineyard, and numerical attributes such as alcohol percentage. Such information about a wine is often made available to the public.

This paper explores the possibility of a content-based filtering recommender system for wines, that will suggest

35th Twente Student Conference on IT June. 27th, 2021, Enschede, The Netherlands.

wines to users based on their taste preferences. A wine's description and its geographical location will be taken as the main features of the recommender system. The accuracy of such a system (the quality of suggestions) is of importance and makes for the quality of the research. The question of whether taste can be detected will be answered.

2. PROBLEM STATEMENT

The main contribution of this paper is to elaborate on the concept of recommender systems and how these systems can be applied to wine suggestions. The research will also contribute by explaining how to build and use a dataset for such a system. In addition, this paper will investigate the correlation between location and wine's taste.

2.1 Research Question

The problem statement leads to the following goal of this thesis:

To examine if taste can be detected using a recommender system and whether such a system can give wine suggestions correlative to the person's taste preferences.

To achieve this goal, the research will answer the following research questions (RQ):

- **RQ1**: Which features of wine determine the taste and which of these features can be applied in a Wine Recommender System?
- **RQ2**: How is a dataset of wine (reviews) collected for a Wine Recommender System?
  - **RQ2.1**: How should the data be prepared to train a Wine Recommender System?
- **RQ3**: Which recommendation techniques should be used in a Wine Recommender System?
  - **RQ3.1**: What are the optimal variable values for the recommendation techniques described in RQ3?

In addition, the following business-related research questions will be answered:

- **Business RQ1**: Which questions could be asked to a person about their taste preference?
- **Business RQ2**: How are these questions from Business RQ1 derived?
- **Business RQ3**: How can the outcomes of the questions from Business RQ1 be utilised in a Wine Recommender System?

3. RELATED WORK

One of the first prediction methods, Bayes' Theorem, was provided in the 1700s by Thomas Bayes. Newer prediction methods such as collaborative filtering only appeared recently. The first research papers on collaborative filtering appeared in the mid-1990s. The term data mining was also introduced in the 1990s and quickly became popular [19]. Soon it became apparent that data was valuable as its usefulness became clear. Since then, lots more research has been done towards recommender systems due to the commercial demand and the abundance of practical applications [1, 11].

In their 2005 paper, Adomavicius and Tuzhilin [1] provided a new, alternative overview of recommender systems. They elaborated on the three main categories and established the differences between content-based, collaborative, and hybrid filtering.

With the great growth in demand for recommending algorithms, it has become more important to evaluate the efficiency and accuracy of these systems. However, the effectiveness and accuracy of many real-world systems are often unreported. In addition, many researchers believe that the quality of such systems cannot properly be judged [6]. As the complexity of a recommender system increases, it becomes more challenging to evaluate the system; the subjectivity of taste makes it especially difficult to evaluate a wine recommender system.

Density-based clustering [13] creates clusters for higher density areas. The density-based-spatial clustering of applications with noise (DBSCAN) [8] is the most popular density-based method. It is a data clustering algorithm proposed in 1996 and received the SIGKDD test-of-time award in 2014 [24].

One of the most common natural language processing algorithms used in recommender systems is TF-IDF. An article from 2016 surveyed that TF-IDF is the most frequently applied text weighting scheme [2]. It is used by 83% of all text-based recommender systems in digital libraries [2].

TF-IDF has been applied to many problems due to its wide range of applicability. In 2018, Juntui and Khoenkaw [12] created an automatic non-personalised book recommender algorithm for book store shelf management that used the TF-IDF algorithm to weight each term.

Charters and Lockshin [3] have shown in their research that even experienced wine drinkers have great difficulty in matching the tastes of the wines with the back label descriptions. During their research, they surveyed which words were helpful (helped the taster identify the taste of the wine) and which were not.

Furthermore, wine critics are creative in their use of language and often use difficult words that can be simplified. In 1984, Ann C. Noble [17] created an Aroma Wheel and Chen and Hambuchen [4] processed this wheel in their paper. It is a multi-level diagram that allows for creative words to be simplified while containing their meaning.

A similar goal is achieved with the unsupervised learning algorithm GloVe [18]. The algorithm attempts to find relationships between words. It maps the words and the distance between the words represent the similarity. It can be used to find lower-dimensional representations [18]. It allows for synonyms to be found and can also simplify complex words.

Recently, using machine learning, researchers have been trying to predict the quality of a wine. In the 2019 paper 'Wine Quality Prediction Using Data Mining', Shruti [25] attempts to predict a wine's quality based on the sample of different wines with their attributes. The data used only included numerical attributes such as levels of alcohol and colour intensity. In 2020, Kumar, Agrawal, and Mandan [14] also attempted to predict wine quality, but used different prediction techniques.

Lots of research has been done towards the influence of location on wine taste and wine quality [15]. The soil largely affects the taste of wine [20, 26]. In addition, the climate is a major factor in wine production and has a large influence on the taste of a wine [27]. Climate conditions vary from year-to-year which causes the ”vintage effect”, variations in yield, quality, and taste [27]. The term used for all these natural environment influencing factors on wine's taste is *terroir*. However, this could all be changed in the future as a ”changing climate and technological advances have threatened the Burgundian notion that the quality of wine depends on regional geography and culture” [28].

Little research has been done towards taste and its application in recommender systems, even less so regarding wine. Wine is a product that varies a lot and is affected by many different factors. A recommender system that recommends wines should take into consideration as many of these factors as possible, most importantly the factors with the greatest impact. Variations of such a system have been created before, but each uses a different set of features. This paper examines which set of features are important to take into consideration.

4. METHODOLOGY

Wine is a product that varies a lot. Among other factors, it differs per vintage year and is affected by the producer, region, and production technique. On each wine bottle, the producer is required to put a back label on the bottle. This is the only way for the producer to communicate with the public. As taste is such a personal sensation, it is still challenging nowadays to make objective judgements and statements about it.

This brings the question, which features of wine determine the taste and which of these features can be applied in a Wine Recommender System?

4.1 Important features

The most important features that are publicly available involve terroir (natural environment factors). For this research, an interview was conducted with Bas Kroese, a certified wine expert. In the interview, Kroese mentioned the following four critical factors of wine that influence the taste of wine besides production techniques:

- The type of grape used for the wine.
- The soil on which the grapes are grown [20, 20].
- The topography of the vineyard.
- The climate in which the grapes are grown [27, 28].

The grapes that were used to create the wine, can be found in the name of the wine; Pinot Noir, Cabernet Sauvignon, Merlot, but also Chardonnay and Sauvignon Blanc. Often, a combination of grapes is used in a wine. The winemaker has to put each of the grape names on the bottle, with their respective percentages. Besides the taste of the wine, grapes also vary in skin thickness, which impacts the tannin in wine. However, the type of grape is not the only determining factor in a wine's taste. The type of soil, altitude, and climate also have a great impact on the taste.

The soil on which the grapes are grown is important as the grapes will absorb water and nutrients from the soil. In addition, the soil should be permeable, allowing water to drain easily. A study from 2009 [26] found that two Cabernet Sauvignon vineyards under the same climate but on different soils produced created different tastes of wine.

The altitude of the vineyard impacts the amount of sun the grapes receive. The orientation and angle also impact the amount of sun received and thus should be taken into consideration.

Lastly, the climate in which the grapes are grown has a great impact on the taste [27, 15, 28]. Ultimately, it decides whether an area is even suitable for wine growing. The majority of vineyards in the world are located between the 30° and 50° degree of latitude on both hemispheres [10]. Grapes grown in a cold climate tend to produce lighter-bodied wines, crisp acidity and often include fruity flavours. Contrarily, grapes grown in a hot climate tend to produce fuller-bodied wines, with higher levels of alcohol, soft acidity, and typically brings bold flavours. There can be great flavour differences in the same variety of wine, grown in a different part of the world. Some great wine growing areas have a microclimate which is of greater importance than the climate itself. Examples of this are the Moselle river in Germany and the Alsace region in France. The Moselle river is used by the winemaker to reflect sunlight onto the grapes, acting as a mirror. This will result in the vineyard being slightly warmer than the surrounding areas; Germany's usual climate is too cold to grow wine grapes.

Since the wine's taste is dependent on the location it can be assumed two wines from the same region will have a similar taste (if the wines are of the same variety). Therefore, a person will be likely to enjoy wines from a similar region. There are lots of geographical attributes that such a recommender system can take into account: climate, amount of sun, angle of the sun, temperature, soil, altitude of the vineyard, microclimate (if present). For the size of this project, latitude and longitude will be to represent the terroir.

Furthermore, since taste is so personal and hard to measure, it can only be described with as much detail as possible. Wine critics review wines and post their description of the wine online. These reviews can be utilised by a wine recommender system to suggest wines that contain similar important characterising words. These words were extracted from the descriptions by preparing the data to compare descriptions with each other as objectively as possible.

For this research, it was important that the features were publicly available. Recommendation techniques were applied to the geographical locations (in latitude and longitude) and the detailed descriptions of the wines. The location has a great impact on the taste of wine and a description is the closest measurement of taste currently available. Combining these techniques allows a recommender system to suggest wines in a similar region or wines with a similar taste description.

4.2 Data Collection

To gather the data required for such a system, wine collections websites have been scraped. These wine collection websites contain information about the wine, including a detailed description of the wine taste, described by a professional wine taster. For this research, the website Wine Enthusiast[1] was used as the source of data.

After the HTML of a review had been requested, it was cleaned of its HTML tags. The Python library *Beautiful Soup*[2] aided in this process. It sits atop an HTML parser, providing possibilities of iterations, searching, and modifying the parse tree.

The pseudocode of the code used to collect the data for this research can be found in Algorithm 1. As most websites block IP addresses when receiving large amounts of unusual requests, collecting this data can be quite difficult. The implementation of Algorithm 1 should therefore ensure that not too many requests are made in a short period of time, this causes the data collection to be time-consuming. Due to the size of the project, a dataset of

[1] https://www.winemag.com/

[2] https://pypi.org/project/beautifulsoup4/

wine reviews found online on Kaggle was used instead [3].

Algorithm 1 Pseudocode for web scraping websites

```
Require: $Pages \geq 0$
data ← {}                                          ▷ Accumulated review data
for all $n \in 1, \ldots, Pages$ do
    function $ScrapePage(n)$
        $Reviews$ ← Reviews on page
        for all $review \in \{Reviews\}$ do
            function $ParseReview(review)$
                $review\_data$ ← $\{description, country, province, \ldots\}$
                $data.append(review\_data)$
```

4.3 Data Preparation

The data collected cannot be used without first being prepared. For this research, the geographical location and the description of the wine were used. Therefore, these two attributes should be prepared properly.

First, the location of the wine was not given in coordinates by the website. Therefore, the geocoding information of the wines was obtained. This was achieved through the Python library *geopy*[4] which uses the geocoding service of OpenStreetMap Nominatim. The returned information includes the latitude and longitude of the given location.

Second, each wine description was prepared properly. This process can be described in the following three steps:

1. Text normalisation
2. Find common phrases
3. Simplify creative words

The first step in this process is to normalise the text. Stopwords and punctuation were removed. In addition, a stemmer was used to reduce words to their stem. To illustrate, a stemmer would turn ”acidity” into ”acid”. This is important since both of these words have the same meaning, but phrased differently. Some wine critics could describe a wine as having ”a bolt of acidity” instead of saying it is ”acid”.

Secondly, combinations of words were found that are be contracted into one word for the recommender system to take into consideration as an entirety. The Python library *gensim*[5] offers a package Phrases that was used during this research to achieve this. To illustrate, Phrases detects that the words ’light’ and ’bodi’ are often found together in wine descriptions, and therefore consolidates them to ’light_bodi’. (The term ’bodi’ represents ’body’ but is stemmed by the stemmer, as this occurs throughout all descriptions, this will not have any effect on the quality of the recommender system).

Lastly, wine critics tend to use creative language when describing a wine, so these creative words were simplified while holding the same meaning. Ann C. Noble [17] created an Aroma Wheel (See Figure 1) and Hambuchen [4] processed her findings. As the database created by Chen and Hambuchen was no longer available online, a project on GitHub[6] still had their map and turned it into *cvs* file that was used instead.

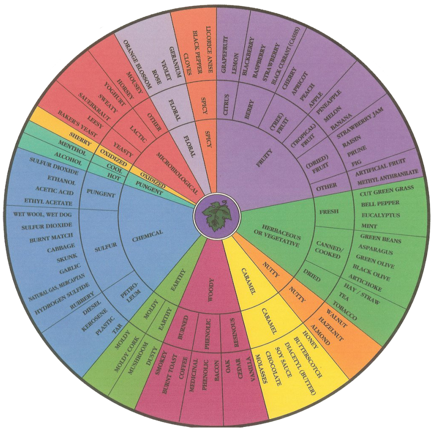

Figure 1. Ann C. Noble’s Wine Aroma Wheel that provides terms to describe wine aromas in different tiers [17].

4.4 Recommendation Techniques

The recommendation techniques used in this research solely used content-based filtering as no user preference data could be found online. The Python machine learning library *scikit-learn*[7] was used to implement the different recommendation techniques.

The recommendation technique that was used to analyse the geographical location is DBSCAN (Density-based spatial clustering of applications with noise) [7]. As it can detect noise, this algorithm was chosen for this research. Since the retrieval of latitude and longitude could have errors, detecting noise was of importance.

The DBSCAN algorithm groups together points that are close to each other. In addition, it is also capable of detecting noise (points that do not belong to any group). The groups returned are referred to as clusters. Scikit-learn contains an implementation of DBSCAN. Unlike k-means clustering, DBSCAN will determine the number of clusters itself. DBSCAN takes two variables, *Epsilon* and *minPoints*. Epsilon is the distance that will be used to locate the points within the neighbourhood; the radius of a hypersphere. MinPoints on the other hand is the minimum number of points required in that sphere to have it qualify as a cluster.

To evaluate the clusters created by the DBSCAN algorithm, the Silhouette score was calculated. The Silhouette score is a measure to compare the tightness and separation [21]. The goal of the evaluation method is to see which objects lie well within their cluster and which are somewhere in between clusters.

The technique that was used during this research to compare descriptions of wines is the Term Frequency/Inverse Document Frequency (TF-IDF) measure [1]. This technique is ideal for a content-based recommender system; any newly added wines to the database can instantly be recommended to users if it contains a description. Other natural language processing techniques include Word2vec [16] and GloVe [18].

GloVe is good at finding relationships between words and thus can be used to find synonyms [18]. It attempts to find lower-dimensional representations, similar to what is done during the data preparation in Section 4.3, simplifying creative words.

On the other hand, Word2vec (proposed and supported by

[3] https://www.kaggle.com/zynicide/wine-reviews

[4] https://pypi.org/project/geopy/

[5] https://radimrehurek.com/gensim/

[6] https://github.com/RoaldSchuring/wine_recommender

[7] https://scikit-learn.org/stable/index.html

Google) [16] has the ability to predict a word depending on the context or a context depending on the word. It is great for getting a better understanding of the content of a piece of text, as the resulting vectors represent the context of each word. However, for this research, it is more efficient to compare results when the output is a score.

Therefore, the TF-IDF technique is ideal for such a system. It is a way to determine the taste of a wine by the words it contains. It uses the prepared detailed description of a wine written by a professional wine taster. This description can very effectively portrait a wine's flavour. It measures the importance or originality of a word by comparing the number of times a word appears in a description with the number of times a word appears across all descriptions.

Term frequency (TF) is a measure for the frequency of a word within a single text string (observation). In the case of a wine recommender system, each description will be of a different size. Hence, it is useful to normalise the document based on its size. The TF value for a word would express a word within a wine description as a percentage of all words. Term frequency, *tf(t,d)*, is the frequency of term t and calculated with Equation 1.

$$tf(t,d) = \frac{|t \in d|}{|d|} \tag{1}$$

Where:

- t: *term* or *word*
- d: *document* or *wine description*

Inverse document frequency (IDF) is a measure of how often it occurs across all the observations. It assists by acting as a measure of how much the information the word provides. If a word appears in many documents, it cannot be distinguished as a relevant or irrelevant keyword. However, if it is rare across all documents, the word could be of greater value. In the case of a wine recommender system, the IDF value of a word would express the number of times a word appears across all the wine descriptions in the dataset. Inverse document frequency, *idf(t,D)*, calculated with Equation 2.

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

Where:

- N: number of descriptions in the dataset $N = |D|$
- $|\{d \in D : t \in d\}|$: number of documents where term t appears

TF-IDF is calculated as the product of TF and IDF.

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D) \tag{3}$$

The Python library *scikit-learn* offers a TFIDF implementation named TfidfVectorizer

4.5 Questions to Ask the User

A recommender system cannot be initialised without any data. If a user is new to the recommender system, no wines have been rated yet which makes it impossible for the system to suggest a wine. To introduce the user to the recommender system, a set of key questions can be asked. This brings the question, which questions can be asked to a person about their taste preference?

These questions should be simplified for the general user and will therefore concern relatable taste sensations. This also ensures that users without prior knowledge of wine can make use of the system.

To derive the questions that the system will ask the user, it is important to know which words are valuable. Charters and Lockshin showed in their research that some words were helpful than others [3]. The questions asked to the user should result in useful words to ensure accurate recommendations. The TF-IDF algorithm calculates the weights of each word to achieve a similar goal. As the results of the questions asked to the user should be helpful but not too specific, these questions should be derived with a good understanding of wine.

Wines can be put on a spectrum for their general taste. If a person enjoys wine's on one end of the spectrum, the same person is more likely to enjoy wines on the same end. The position of a wine on this spectrum can be determined through keywords. Words such as ”sweet” and ”light” appear on one side while ”dry” and ”bold” appear on the other (See Figure 2). The key questions should result in such keywords and should help the system determine where on such a scale a person's preference lays.

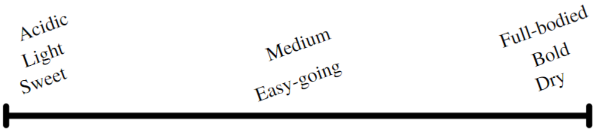

Figure 2. Red and white wine scale.

The questions should, ideally, result in a position on the scale in Figure 2. A word that helps determine the position is seen as 'helpful'. Charters and Lockshin [3] have researched many wine terms that were used a lot and analysed their positive or negative utility. These words can be found in Table 1. A word is considered 'helpful' with determining a wine's taste if it has positive utility (helps the taster identify the taste of the wine), otherwise, it is deemed 'unhelpful'.

However, while 'chocolate' is classified as helpful, there are many types of chocolate that each pair with a different end of the spectrum shown in Figure 2. White chocolate tends to be mellow and buttery and therefore is a good determinant for sweeter wines. Milk chocolate pairs well with a medium-bodied wine. Contrarily, dark chocolate is rather bitter and therefore fits well with fuller-bodied, intensely flavoured wines with bold fruits. By asking the user what type of chocolate they prefer, a better estimation of their taste preference can be made.

The questions that are asked to the user should be created surrounding these helpful words. The result of a question asked should in all instances be a helpful word to increase the accuracy of the recommender system. The system could ask the user ”Do you prefer melons or peaches?”. This question could result in one of two outcomes; ”melons” or ”peach”; both of these words are deemed helpful.

Questions can also be created with more than two outcomes. For instance, by asking ”How do you prefer your tea?”. This could result in many different outcomes. For example, a sweetened cold tea should result in terms that describe a sweet wine that tastes better cold. On the other

Table 1. Words with a high frequency of mention categorised by a widely agreed sense of positive or negative utility [3]

Positive	Negative
Acidity	Balance
Aromas	Basket
Blackberry	Concentration
Butterscotch	Days
Cherry	Fermentation
Chocolate	For
Citrus	Full-flavoured
Crisp	Great
Fruit	Lingering
Melons	Oak
Nutty	On
Oak	Pressing
Peach	Seven
Rich	Soft
Soft	Tannins
Velvety	Varietal
Violet	Well-integrated

hand, a strong warm tea without sugar should result in a different type of wine, full-bodied with many rich flavours. Such questions are ideal as they offer more options for the user.

The results of these questions can be utilised by a recommender system. To allow this, an alternative TF-IDF function was created that allows for words to be entered, instead of wines. The function attempts to find wines that score high with the given words.

5. EXPERIMENT

For the size of this research, a dataset found online on Kaggle was used. In addition, this dataset had been reduced to 50.000 samples to reduce the run-time of the algorithms during this research.

5.1 Data Preparation

To obtain the geocoding information of each wine, combinations were made of the limited information about the location that was available in the dataset (i.e. the name of the vineyard in combination with the country). The information that was given in the dataset was: *country*, *province*, *region_1*, *region_2*, *vineyard*, *winery*. If the combination did not return any results, a request was made for another possible sensible combination. If no combinations were successful, the sample was removed from the dataset instead. In addition, since the searches sometimes resulted in inaccurate results, these were filtered out as good as possible and also removed from the dataset. This time-consuming process resulted in 44.801 samples for which geocoding information was available.

Afterwards, the descriptions of each wine were prepared for the recommender system to apply the TF-IDF technique more efficiently and accurately. The three data preparation steps described in Section 4.3 (text normalisation, find common phrases, simplify creative words) were used and applied to all wine descriptions in the dataset. To give an example, the following description is that of a white wine (Jacquère) from France:

> *Crisp and fruity, this bright wine is full of lively acidity, tangy citrus and apple fruit and refreshing acidity. It is ready to drink.*

By applying each of the data preparation methods, the following descriptive terms were obtained from the description of this wine:

> *crisp fruit bright wine full live acid tangy citrus apple fruit refreshing acid readi_drink*

Throughout this project, the descriptive terms returned by these preparation methods will be referred to as the wine's 'tokenised description'.

5.2 DBSCAN

For this project, the minimum amount of samples required to form a cluster was set to 50, due to possible errors in geocoding information retrieval. Decreasing this amount will result in more clusters being created as groups of outliers will become their own cluster. Any outliers are classified as noise and should not be taken into consideration, they could ultimately be the result of mistaken geocoding information requests. To decide the optimal value of *epsilon*, several DBSCAN algorithms were performed on the wines in the dataset with different values of epsilon. The Silhouette score of each run was then calculated and the results can be found in Figure 3.

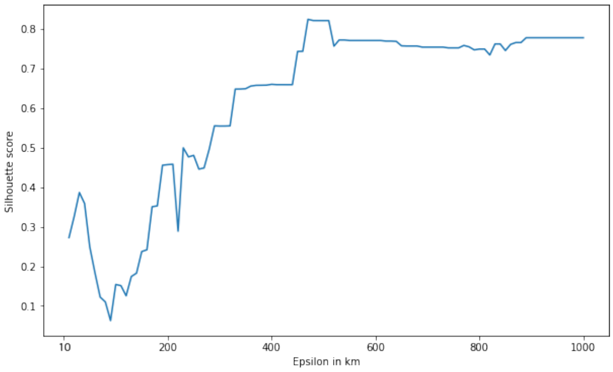

Figure 3. Silhouette score of DBSCAN algorithm with different values for epsilon represented in km.

From Figure 3, it can be determined that around 500km is the optimal value of *epsilon*, as this number received the highest Silhouette score. This will result in a total of 12 clusters, and a visualisation of the clusters with an epsilon value of 500km can be seen in Figure 4. Each cluster has its own colour and the points that are not in a cluster (noise) are coloured black.

5.3 TF-IDF

When a word occurs often, it does not imply that the word is of great importance. Instead, the opposite is often true. Words that have a high occurrence are often of little value when trying to discriminate between wines. This problem is tackled by the TF-IDF technique. For this project, TfidfVectorizer from the Python library *scikit-learn* was is used and trained with all tokenised descriptions. This way, the TfidfVectorizer knows of all wines and all possible descriptive terms.

To analyse the effect of the TF-IDF algorithm, only the tokenised descriptions of all Chardonnays were used to calculate the TF-IDF score for each term. The resulting matrix of 3939 (rows) by 19531 (columns) represent the 3939 Chardonnay wines and the 19531 descriptive terms representing the vocabulary of the TfidfVectorizer. As the TF-IDF score of a term is dependant on the description

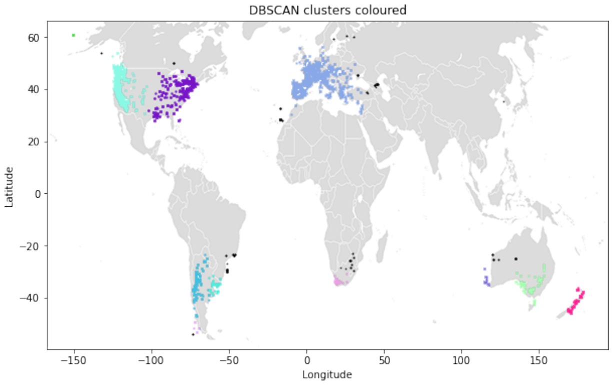

Figure 4. The clusters returned by the DBSCAN algorithm with epsilon as 500km. Each colour displays a cluster and noise is portrayed by the colour black.

in which it was found, the average of all TF-IDF scores across all descriptions was taken and plotted against the frequency of that term across all tokenized wine descriptions. The resulting graph can be found in Figure 5.

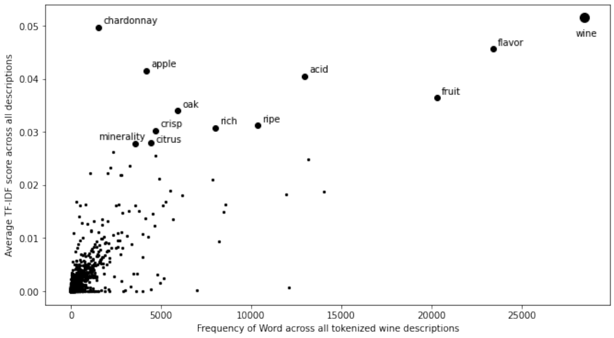

Figure 5. Every descriptive term found across all tokenized descriptions where only tokenised descriptions of Chardonnay wines are measured.

The graph in Figure 5 can be separated in four sections; top-left, top-right, bottom-right, and bottom-left.

- **Top-left**: These words are of great importance. As they do not occur in many wine descriptions, but still have a high average TF-IDF score, when it does occur in a description it tends to have a high TF-IDF score.
- **Top-right**: These words are in general also important, but do not help distinct the different wines from each other very much. As these words occur in so many descriptions, on average the word will not have a large impact on determining the taste.
- **Bottom-right**: These words are of little or no value, occur often, and do not help distinguish the descriptions from each other. Stopwords are often found in this area. Since stopwords are removed during the data preparation, there are not many words in this area in Figure 5.
- **Bottom-left**: These words do not occur often, but not much can be said about their value. In a few descriptions, a term can have great value but in general, be very rare. In this case, such words will result in very specific results.

The descriptive terms that resulted from the Jacquère from France (shown in Section 5.1) after applying the preparation methods were then used to compare with the descriptive terms of other wines. The number of occurrences of the distinct descriptive terms from this wine is shown in Figure 6. While some words occur a lot more often than others, not much can be said about the importance or originality of each of these terms yet.

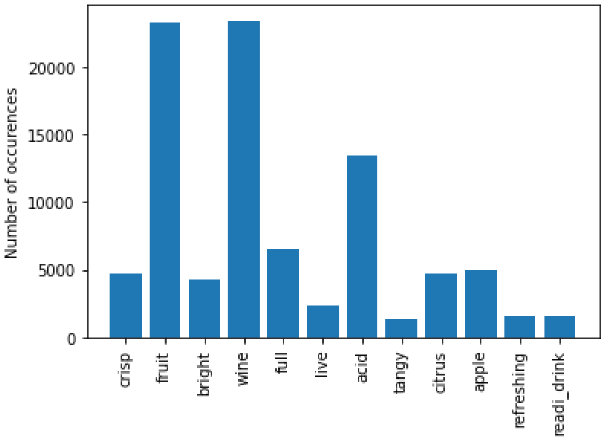

Figure 6. The number of occurrences of the descriptive terms returned from "Domaine de l'Idylle 2015 Cruet Jacquère (Savoie)" in that of other wines.

To determine the importance of each term, the TF-IDF score for each term across all descriptions was calculated. The TfidfVectorizer was again trained with all tokenised descriptions so its vocabulary includes every possible descriptive term. The resulting 44802 by 19313 matrix that was returned by the TfidfVectorizer correspond to the 44802 rows of wine descriptions of the dataset and the 19313 columns of different terms.

An extract of the returned matrix can be found in Table 2. The first row of this extract shows the entered wine (the Jacquère from France). The other rows shown are the rows of the top three recommendations in Table 3 for the entered wine (with similar TF-IDF scores). The recommendations are determined by how similar the TF-IDF scores of each term is compared to the entered wine. The columns show the words tokenised description of the entered wine, words that do not occur in this tokenised description are left out. The other rows contain words that do not occur in this table but these words are not shown.

The similarities between the rows of Table 2 can be observed. Rows with similar TF-IDF scores for each term can be seen as accurate recommendations. To find these rows, Cosine Similarity is used. Cosine Similarity is a metric used to measure how similar documents are, regardless of their size, through the TF-IDF weights given. Any number of wines can be given and the program will attempt to find cosine similarities for each of the entered wines.

The Jacquère from France (from Section 5.1) was entered in the TF-IDF system and the top five recommendations with the highest cosine similarities were retrieved and can be seen in Table 3. The top five recommended wines had extremely high TF-IDF similarity scores and also seemed to be grown in a similar climate. In addition, the top five recommendations were all white wines.

A similar method was created for this project which allows a set of words to be given, instead of wine(s). To test this

Table 2. Extract of the 44802 by 19313 matrix returned by the program.

Wine	Live	Readi_drink	Tangy	Bright	Refreshing	Crisp	Acid	Wine	Fruit	Citrus	Apple
Jacquère	0.347	0.345	0.357	0.270	0.344	0.258	0.356	0.133	0.298	0.264	0.271
Wine 1	0.314	0.313	0.323	0	0	0.234	0.323	0.120	0.270	0.239	0
Wine 2	0.301	0.300	0.301	0	0.299	0.224	0.155	0.115	0.129	0.229	0
Wine 3	0	0.327	0	0.511	0.326	0.245	0.169	0.252	0.141	0.250	0

Table 3. TF-IDF Technique recommendations for "Domaine de l'Idylle 2015 Cruet Jacquère (Savoie)"

Wine	Country	Variety	TFIDF similarity
Wine 1	France	Albert Bichot	0.668
Wine 2	Portugal	Fernão Pires	0.646
Wine 3	Portugal	Portuguese White	0.628
Wine 4	France	Chardonnay	0.613
Wine 5	France	Chardonnay	0.562

algorithm, the descriptive terms from the Jacquère from France (shown in Section 5.1) were entered. The top five results included the top four of Table 3 in addition to the Jacquère from France itself. This wine had a TF-IDF score of 1 against itself, as the entered words were identical to the tokenised description of the Jacquère from France. This alternative method can be used in combination with the results of the questions described in Section 4.5.

6. RESULTS

This paper has analysed the effect of applying a density-based clustering technique on wine locations and attempted to receive recommendations for wines through the TF-IDF algorithm. The resulting tables and plots gave useful insights on future work for a wine recommender system.

The clusters created by the DBSCAN with *epsilon* set to 500km, were not ideal for a wine recommendation system. Europe received its own cluster and in Europe alone, there are many different flavours of wine. A different approach should be taken to analyse the terroir.

Figure 5 shows the sum of TF-IDF scores for each term concerning all Chardonnay wines. It shows which words are important when identifying a Chardonnay wine. Some of the resulting highest rated words can also be found in Table 1. Charters and Lockshin [3] found in their research that some words have a higher utility, and the results found in Figure 5 prove this. The following words had a high TF-IDF rating across all observations for Chardonnays, and are categorised as 'helpful' by Charters and Lockshin: acid, citrus, crisp, fruit, oak, and rich.

The results found in Table 3 were very accurate and had high cosine similarities for the terms' TF-IDF scores. However, after searching for these wines on Wine Enthusiast, it was found that each of these wines was reviewed by the same wine critic, Roger Voss. This is because each wine critic uses their own personal vocabulary to describe the taste of a wine. While some terms occur more often, and these terms are analysed and weighed through the TF-IDF technique, the wine critics still have their own personal manner of describing the taste which is often distinct from that of others. The vocabulary they use is not widely spread and there is not a commonly accepted list of descriptive words. While creative words were simplified, this does not take into consideration the writing style of the critic. After conducting more experiments, the resulting recommendations were mostly limited to wines that were described by the same wine critic. This is undesired behaviour as it indirectly limits the user's possibilities.

However, this problem is self-correcting when multiple wines, reviewed by different wine critics, are given to the system. The system takes the descriptive terms of each of the given wines. At this point, the vocabulary used, writing style, and sentence structure is mixed with that of other critics so it can no longer be traced back to a singular person.

From this, it can be concluded that the TF-IDF technique is not a reliable recommendation method when given a small number of wines. The method should only be used when a sufficient amount of user data (previously enjoyed wines) is available. Alternatively, the method can be used in combination with the questions from Section 4.5.

7. CONCLUSION AND FUTURE WORK

The goal of this paper was to analyse if taste can be detected using a recommender system and if a bottle of wine could be suggested correlative to the person's taste preferences. This research has found that taste cannot truly be detected yet, as it is such a personal sensation. In addition, the recommender system is not aware of what taste is as it cannot be described through numbers. However, since the sensation of taste can be described through words, these words can be utilised to gain knowledge about the taste which allows for accurate and efficient suggestions to be made depending on a person's taste preference.

The geographical location recommendation technique used during this research (DBSCAN) appeared to not be a good choice for such a recommendation system. Instead of only coordinates, multiple factors of the terroir should be taken into account. This problem should be seen as a supervised learning problem and could be solved with a complex multi-label classification system in future research.

This research has discovered by utilising the TF-IDF technique that there is a great correlation between the geographical location of wine and its taste, as many papers suggest. Applying the TF-IDF technique on wines in a specific region, the resulting recommended wines were often grown in the same region or in a similar climate. This further strengthens the argument that geographical location has a great effect on the wine's taste.

However, since each wine critic uses their own personal vocabulary to describe the taste, a single wine's suggestions are often limited to wines described by the same wine critic. Therefore, this recommendation technique should not be used on a singular wine with a dataset that contains multiple wine critics. A dataset that uses a specific set of descriptive words could solve this issue. However, such a dataset is currently not available online.

Finally, future work could further investigate the effect of the geographical location; then see if there is any gain in combining the TF-IDF technique with the geographical location of wine. Furthermore, research could be done towards a wine recommender system that filters through collaborative filtering, by generating user preferences data, as such a system will not be affected by the limited vocabulary of the wine critics and possible biased descriptions.

8. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[2] J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems : a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, 2016.

[3] S. Charters, L. Lockshin, and T. Unwin. Consumer responses to wine bottle back labels. *Journal of Wine Research*, 10(3):183–195, 1999.

[4] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen. Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel. *2014 IEEE International Conference on Data Mining Workshop*, 2014.

[5] V. Cheynier and P. Sarni-Manchado. Wine taste and mouthfeel. *Managing Wine Quality*, pages 29–72, 2010.

[6] Z. Dehghani Champiri, A. Asemi, and S. Siti Salwah Binti. Meta-analysis of evaluation methods and metrics used in context-aware scholarly recommender systems. *Knowledge and Information Systems*, 61(2):1147–1178, 2019.

[7] D. Deng. Dbscan clustering algorithm based on density. *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, pages 949–953, 2020.

[8] M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[9] P. Ferguson. The Senses of Taste. *The American Historical Review*, 116(2):371–384, 04 2011.

[10] T. Goldammer. *Grape Grower's Handbook: A Guide To Viticulture for Wine Production.* Apex Publishers, 2018.

[11] D. Jannach and M. Jugovac. Measuring the Business Value of Recommender Systems. *ACM Transactions on Management Information Systems*, 10(4):1–23, 2019.

[12] S. Juntui and P. Khoenkaw. Automatic non-personalized book recommender algorithm for bookstore shelf management. In *2018 International Conference on Digital Arts, Media and Technology (ICDAMT)*, pages 49–53, 2018.

[13] H. Kriegel, P. Kröger, J. Sander, and A. Zimek. Density-based clustering. *WIREs Data Mining and Knowledge Discovery*, 1(3):231–240, 2011.

[14] S. Kumar, K. Agrawal, and N. Mandan. Red Wine Quality Prediction Using Machine Learning Techniques. *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020.

[15] P. Lock, S. Mounter, E. Fleming, and J. Moss. Wineries and wine quality: The influence of location and archetype in the hunter valley region in australia. *Wine Economics and Policy*, 8(2):180–190, 2019.

[16] L. Ma and Y. Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897, 2015.

[17] A. C. Noble. Ann c. noble's wine aroma wheel.

[18] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.

[19] M. Ramzan and M. Ahmad. Evolution of data mining: An overview. *2014 Conference on IT in Business, Industry and Government (CSIBIG)*, 2014.

[20] G. Retallack and S. Burns. The effects of soil on the taste of wine. *GSA Today*, 26(5):4–9, 2016.

[21] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[22] J. B. Sancho, F. O. Requena, A. H. Esteban, and J. B. Bermúdez. A collaborative filtering approach to mitigate the new user cold start problem. *Knowledge-Based Systems*, 26:225–238, February 2012.

[23] H. Scander, N. Neuman, R. Tellström, and A. Yngve. Acquiring competence: Sommeliers on 'good' food and beverage combinations. *International Journal of Gastronomy and Food Science*, 20:100199, 2020.

[24] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: Why and how you should (still) use dbscan. *ACM Trans. Database Syst.*, 42(3), July 2017.

[25] P. Shruthi. Wine Quality Prediction Using Data Mining. *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, 2019.

[26] J. Ubalde, X. Sort, A. Zayas, and R. Poch. Effects of Soil and Climatic Conditions on Grape Ripening and Wine Quality of Cabernet Sauvignon. *Journal of Wine Research*, 21(1):1–17, 2010.

[27] C. van Leeuwen and P. Darriet. The Impact of Climate Change on Viticulture and Wine Quality. *Journal of Wine Economics*, 11(1):150–167, 2016.

[28] M. White, P. Whalen, and G. Jones. Land and wine. *Nature Geoscience*, 2(2):82–84, 2009.