Creating an adaptable beer recommender: Using Word2Vec and content-based filtering to democratise beer tasting notes

MSc Data Science and AI for Creative Industries
Thesis Project

by Alena Yakushova

UAL, London, 2022

Contents

Introduction

Beer is one of the world’s most popular and most celebrated drinks. The recent craft beer revolution, brought on by a new generation of inquisitive millennials, social media and reduced costs of running small-scale breweries (Brown, 2017), has been accompanied by an increase in machine learning (ML) and artificial intelligence (AI) application in the centuries’ old industry. Particularly, it has resulted in various beer recommender systems being created. Conducted market research however shows that there is still a knowledge gap in consumers' experience of beer consumption, regarding the variety of available options, and that there is a lack of an adaptable, ‘beer novice’ friendly recommendation system. This thesis project aims to create an approach that could fill that gap. This draft academic paper consists of four chapters:

Chapter 1 gives a literature review of related existing research and systems;
Chapter 2 discusses the market research conducted to discover the gaps in consumers’ experience;
Chapter 3 provides methodology details on how the approach was created;
and finally, *Chapter 4* evaluates the results and provides suggestions for future research.

The accompanying notebook with relevant code can be accessed here.

Chapter 1: Literature review

Beer today is not only trendy, but also digital. Since the introduction of social drinking apps in 2007, both industries - beer and technology - have developed a long way individually and in collaboration with one another (Smiaware, 2017). Beer and mobile technology have come together in a fascinating way in 2008 with Steve Sheraton's iBeer app (MelMagazine, 2021). Sheraton first developed a beer drinking video which was sold through iTunes, and when Apple came up with AppStore iBeer was born. Sheraton shared that 'iBeer was a huge success because it allowed people to show their friends what their phone is capable of'. The notion of peer-to-peer sharing is as important and true nowadays, and a great beer recommendation system can be that amusing and fun thing to share.

According to many articles and reviews, Untappd is the most famous beer app and has been for a while with over nine million registered users (HopCulture, 2022; The Beer Connoisseur, 2016; MakeUseOf, 2021). There is a web version and a mobile app for Untappd, with the latter having a unique function which allows its users to share their location. The app is famous for its simplicity and connectivity, users can share what they are drinking, time and location, as well as check out and comment on the choices of their friends. There is also a reward system in place which keeps the users on the app and entertained - they can earn badges when checking in their drink (Untappd, 2022).

Similarly, focus on location is embedded in the BreweryMap app. The app launched in 2021 compiles a searchable map of all breweries in Michigan (Pair, 2022). It was developed as a way to bring people out of their homes after the Covid-19 lockdown. Although the app is still very new with just over 7,500 users, the developer has great plans to expand and bring more people out into the tasting and socialising world. This is thus another example of a beer app which aims to connect and provide users with best options of what is on the market. This close relationship to users is carefully nurtured by the company's founder Kenneth Konarzewski, as they recently decided to add ciders and other alternatives to beer to the app's selection of breweries due to demand for beer alternatives from the app's users. (Pair, 2022). This shows that the popularity of beer alternatives has increased, and it can be predicted that individuals would be more prone to explore different beers, hence a well-developed beer recommendation system will be in demand.

Another noteworthy recommendation app is Next Glass, which uses vast amounts of data to tell consumers what beers or wines they might like. The creators of the company, founded in 2013, ran 15,000 beers through a mass spectrometer that sorted all the chemical compounds in a beer by size, and gave them tens of thousands of data points to produce recommendations. Despite this highly sophisticated approach to recommendations, the app still fails to address the complexity of what a given consumer will actually like: "a "you'll-probably-like-this-too" rating system feels insufficient when it comes to the human palate", notes Shilton (2015). This, however, has not evidently curbed Next Glass' strive to become the ultimate beer recommendation app: in 2015 it acquired the Untappd app and in early 2020 another platform uniting beer enthusiasts all over the world, BeerAdvocate (Furnari, 2020).

Although there seems to be a lot of beer apps on the market, the gap in knowledge is evident. There are no apps which are specifically famous or known for providing good, accurate recommendations of new beer to try for individuals based on their previous choices of beer. Thus, this research aims to deliver a suitable system, where the recommended beers to customers will be precise and on target. If that is successful, the peer-to-peer sharing and reviewing will take place and the app will be recommended and therefore grow in popularity.

When looking into existing beer recommendation apps and approaches, it is important to note the IBM's Mobile innovation lab and design studio experiment that took place at South by Southwest (SXSW) conglomerate of art festivals in Texas in 2016. After gathering vast amounts of data and working through different machine learning algorithms the team finalised five questions to ask SXSW visitors before giving them 'an incredibly accurate' prediction of what beer they would enjoy (Gutierrez, 2016). The visitors were asked about their favourite berry, cheese and dessert, as well as what time of the year they prefer to drink beer and when they prefer to drink beer the most.

Personalisation of beer properties has also been explored in great detail by al-Rifaie & Cavazza (2022). A framework they introduced uses an evolutionary method supporting multi-objective optimisation to enable brewers to map the desired beer properties into ingredients dosage and combination in response to individual consumer preferences. Similar research has been found valuable not only to consumers wondering what beer to order next but to bars and breweries trying to learn the preferences of their demographics and grow their business (Wilson, 2012).

There are also a number of educational platforms and apps about beer, BJCP Styles being one of the most prominent services in the sector (HopCulture, 2022). The app from Beer Judge Certification Program (BJCP) works as an encyclopaedia for beer. The app is a lot more informative and formal than other beer apps, and in contrast to the examples above lacks social and communicative aspects. However, the app is perfect for a very specific target audience. The success of this particular app signifies that there is technical and in-depth interest in beer in our society, so it is important to consider informational and educational content about beer for the beer recommender system developing.

Generally speaking, recommender systems commonly use two approaches to generate recommendations: content-based filtering and collaborative filtering, or a combination of both (Kane, 2018). While content-based filtering uses the characteristics or properties of an item to suggest similar recommendations, collaborative filtering looks at similarities between users. Both have its advantages and challenges, yet this project focuses solely on content-based filtering. This approach has an advantage that only information about the items, beers in this case, is needed and no extensive data about the user of the recommender. It also bypasses the new user 'cold start' problem, which occurs when the system cannot generate recommendations for users about which it has not yet gathered sufficient information (Bobadilla, 2012).

While Natural Language Processing (NLP) as a subfield of linguistics, computer science and artificial intelligence dates back to the mid-20th century (Nye, 2016), the recent advances which play a major part in the industry today became widespread in the 2010s (Otter, 2018). For instance, the NLP technique Word2vec was created and published in 2013 by a team of researchers led by Tomas Mikolov at Google. The trained model represents words as vectors, as the name suggests, which then allows to indicate the level of semantic similarity between the words or sentences with the help of a simple mathematical function such as cosine similarity (Singhal, 2001). Precisely for this purpose Word2vec has been used in this project, as discussed in Chapter 3.

Chapter 2: **Market research**

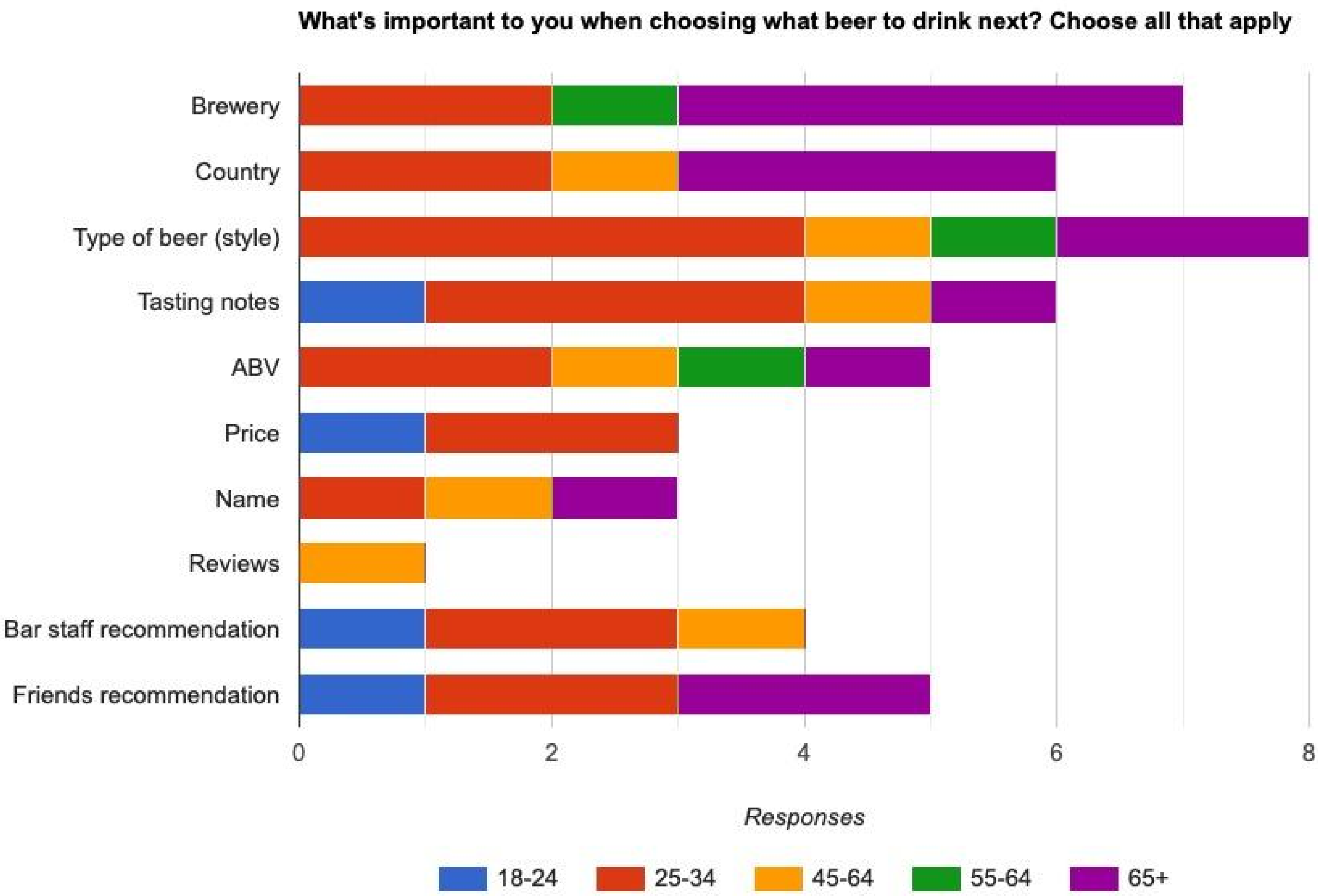
2.1 Offline survey

To gain a better understanding of the gaps in consumer experience of beer recommendations I first conducted several interviews at the Great British Beer Festival (GBBF), which took place on 2nd-6th August in Olympia, London. That event presented a perfect opportunity to reach a variety of beer drinkers, including those who traditionally prefer cask ales and those who like discovering craft beer. After conducting qualitative interviews with a randomly selected sample of 10 people I was able to record all the factors they considered when choosing a beer. This led to making a list of 10 most mentioned factors which would be subsequently used in an online survey.

In this random sample, I was able to collect responses from representatives of a variety of age groups. All 10 interviewees were asked to identify their age group, but no other personal data was recorded. Informed consent was obtained from every person surveyed, and they were given my contact details in case of any queries regarding the usage of their responses. Covering as many age groups as possible in the limited time allocated for this part of the investigation and getting qualitative data from different individuals was in pursuit of a greater external validity of this offline survey, and so that the findings would be more appropriate and resourceful for later quantitative data collection. Analysing the gathered offline data revealed that the only group not represented was the 35-44 segment. That, however, was not deemed enough of a limitation to make the research invalid, but rather a good consideration for further investigation.

The key findings of my offline survey are presented in a bar chart in Figure 1, displaying the number of people who mentioned a particular factor as important when choosing a beer, with colour-coded age breakdown. Generally, people approached at the festival did not struggle when choosing what beer to drink next, given the prior knowledge of different components of a good beer choice. Outside of this event, as the online survey would show, that would be a significant issue for some respondents. Given the specific nature of the event where thousands of people gather to try beers from all the different breweries and countries, it was expected that one of the most popular answers was the tasting notes, provided in the festival's programme and on every bar. In the real world, however, tasting notes are not always readily available, and that factor becomes less

Fig.1: *Factors guests of GBBF considered important when choosing a beer*



important, as wider market research would show later. The least mentioned factor was reviews, and only by the 45-64 age group, showing that many participants from the sample did not consider reviews important when choosing a beer. It is highly likely that this factor was not mentioned as many times as others because, once again, of the event's context where it would take considerable time and additional effort to locate online reviews or gather offline reviews of one of the hundreds of beers present. Reviews would be a more important factor when purchasing a beer online, for example.

Notably, a third of respondents at the festival also highlighted the importance of trying something unique, such as rarely available on tap Imperial Stout by Sussex-based Harvey's Brewery (9% ABV (Alcohol by volume)) or interesting beers from European countries, never seen in the UK before. This trend was highly specific to the nature of the event where the sample was selected, and is not included in the final representation of the results. Still, Figure 1 reveals several interesting trends worth considering outside of the event's context:

- Style or type of beer was mentioned by most people when talking about important factors behind their beer choice. The only age group that did not consider it an important factor was the 18-24 year olds. For this age group tasting notes, price and person-to-person recommendations were more important.
- Brewery that made the beer was the second most mentioned factor in beer selection for the whole sample, with the 65+ age group considering this factor the most important, followed by the country the beer was made in. This breakdown suggests that older generation are less prone to trying new beers outside the established pool of beverages they trust.
- Tasting notes and ABV got a moderate amount of mentions, but since four out of five age groups mentioned these factors, they can be considered as important to any generation or age.
- Bar staff and friends' recommendation factors were also noted by a fair amount of people. Particularly, the 25-34 and 65+ age group considered friends' recommendations very important.
- Price and name of a beer were also discussed by the sample, but less than other factors mentioned above. Notably, price is an important factor for 18-24 and 25-34 age groups of the sample. This is an expected finding, as other age groups in most cases would have a more stable income and lifestyle and may not consider price an important factor.

When analysing the findings from this qualitative research, it is unavoidable to acknowledge the setting where the data collection took place as a limitation, since it takes the research out of the real world context. However, it can be argued that the external validity of the festival's survey was balanced out by choosing to interview participants from the different age groups. Most importantly though, the aim of this data collection was to explore what factors influence consumers' beer choice as a starting point for a larger research in the real world. Indeed I used these findings in further quantitative research, when asking a bigger sample of participants to answer this question and several more in an online survey. To the same question about what is important to an individual when choosing a beer, I was able to provide more informed and researched options of answers for the participant to choose from.

2.2 Online survey

To expand the scope of my market research I conducted an online survey of beer drinking habits, which was taken by 108 people, with the help of SurveyMonkey service. Participants were recruited through the means of social media, and their responses were anonymous, revealing only their age group and no other personal data. The survey began by assessing lifestyle patterns of the participants, starting with broader questions and then zooming into more specific preferences.

Firstly, the participants were asked about the frequency of their beer consumption, including a non-alcoholic option, since the focus of this survey was beer as a beverage, and not just as an alcoholic one. Participants' views varied, spreading the results evenly between all answer options, shown in Figure 2. Interestingly, almost a quarter of participants (23.15%) answered that they do not drink at all, whereas the least number of participants, 12.96%, reportedly consume beer every day.

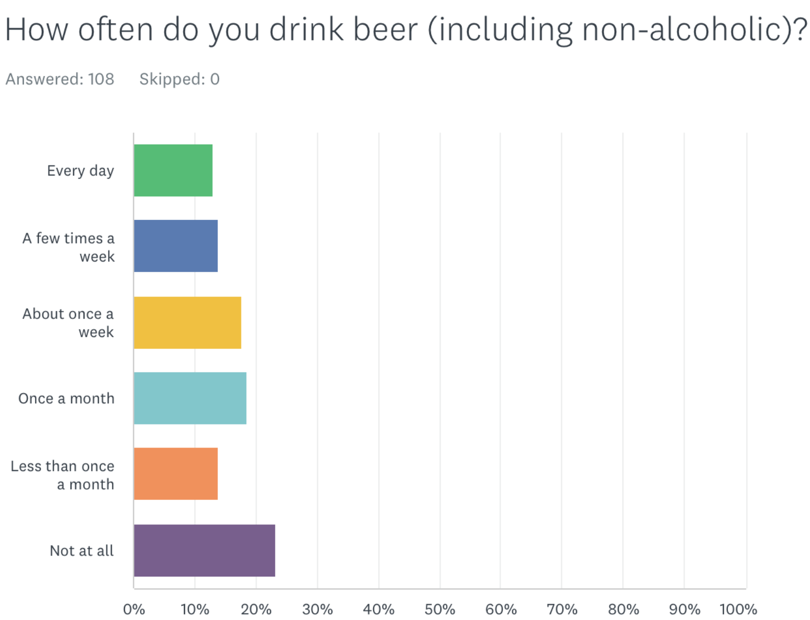
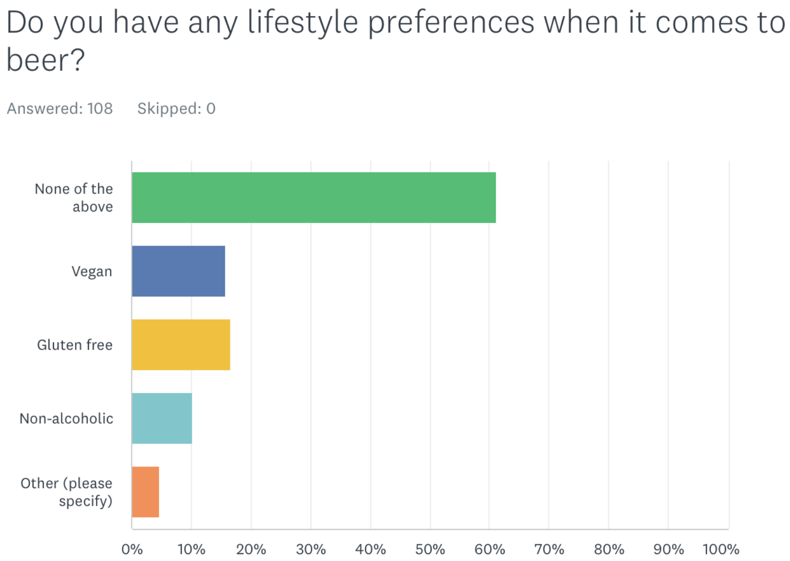


Fig. 2: *Frequency of beer consumption of online sample*

To identify different lifestyle patterns of my sample further, I then asked participants if they had any intolerances or dietary preferences (Figure 3). 61.11% stated that they did not have any specific lifestyle preferences regarding their beer consumption, while 16.67% of the sample noted the gluten free diet and 15.74% – vegan. 10.19% of the sample selected a non-alcoholic option, supporting my decision to include it in the previous question. A small, but significant in this context group of 4.53% of the sample provided their own answers, which allowed for receiving more in-depth data from those participants. In this pool of answers, options like 'light' and 'low calorie' appeared more often than others, thus supporting



the notion of beer's ability to complement

Fig.3: *Lifestyle preferences of online sample*

a person's lifestyle choices and diet. Overall, these findings indicate that beer can be suitable for many lifestyle preferences, especially since the variety of non-traditional beers has massively increased in recent years.

After a brief exploration of their lifestyles, participants were asked about their favourite type of beer, with a possibility to choose up to three options. To little surprise, lager was the most common answer, with 36.11% participants choosing that option (Figure 4). The second most popular answer was the 'other' option, allowing participants to provide their own answer, not mentioned in the options given. The answers provided were of great value to the research.

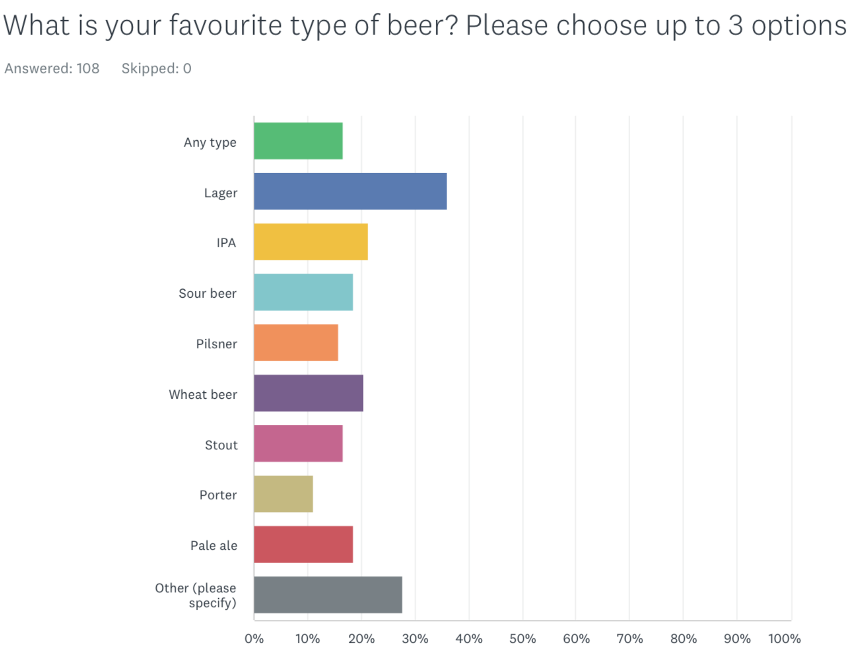


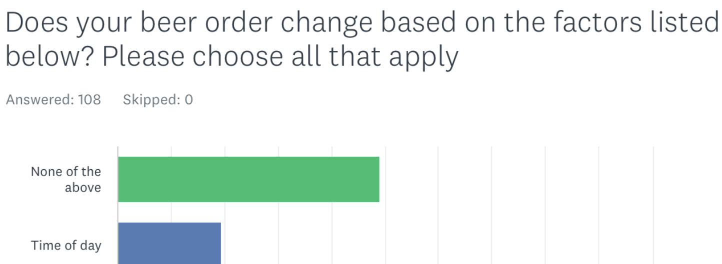
Fig.4: *Favourite types of beer of online sample*

Most participants who chose the 'other' option, specified that they simply do not have a favourite beer or that they do not drink. The two answers which stood out were 'Not too familiar with the difference above. I know I dislike stout' and 'Whatever Corona and other light beer is'. These qualitative answers indicate the lack of knowledge about available options and styles of beer. 'Corona' and 'Stout' are most likely used here as a symbolic representation, which suggests that many people simply put beer into light and dark categories. Another recurring specification was 'Heineken', which alongside Corona, is a well-known lager, thus supporting the finding that lager is the most liked type of beer by the sample. Unsurprising that the names of these two beers are used in lieuof a type of beer, as both are made by the largest companies in the industry: Corona Extra is a brand owned by AB InBev conglomerate with 31.3% global market share based on volume sales, while Heineken producer has second largest market share of 12.4% (Statista, 2021)

Following lager as participants' favourite style of beer was IPA with 21.3% having chosen this option. A fifth of the sample (20.37%) said wheat beer was their favourite type, while an equal number of people (18.52%) chose pale ale and sour beer. Stout was chosen by 16.67%, and the same number of people said they liked any type of beer. The two least popular beer types within the sample were Pilsner (15.74%) and Porter (11.11%).

To assess the participants' awareness of not only different beer styles but also different ways it can be served, I posed a question whether they preferred cask or keg. 42.59% of participants stated that they did not know the difference, further supporting the idea about the lack of knowledge of the beer industry. 21.3% of the sample chose keg as their preferred beer delivery method, and the same number of people said they liked both. Cask received only 14.81% of the votes.

Drawing inspiration from IBM's Mobile innovation lab and design studio experiment (Gutierrez, 2022) where participants were asked what time of the year they prefer to drink beer, I asked my sample whether their beer order changed depending on time of the day, day of the week, season and weather. The participants were also given the option 'None of the above', which was chosen by almost half of them (Figure 5).. These findings can signify multiple things, an important one to highlight would be the fact that due to the lack of knowledge about how beers could vary and suit different times and occasions, the participants perhaps never experienced an ability to make beer choice based on any outside factors. However, in the future, when conducting

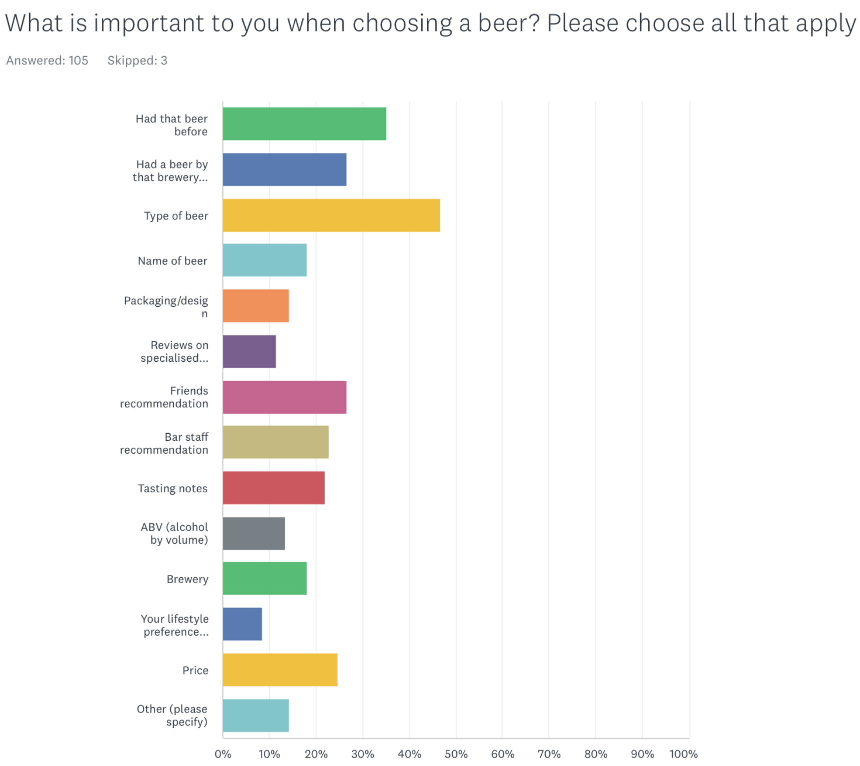


the same investigation, I would suggest adding a 'Please specify why' in a comment section when a participant chooses 'none of the above'. If that is done, more qualitative personal data will be collected which will produce findings which can more specifically explain

why the majority

Fig.5: *Changes of online sample's beer order on time/weather*

of the sample stated that their beer order does not change based on the factors provided. 29.63% of the participants said that their beer choice will change due to season, closely followed by 22.22% of those whose beer order changes due to weather.

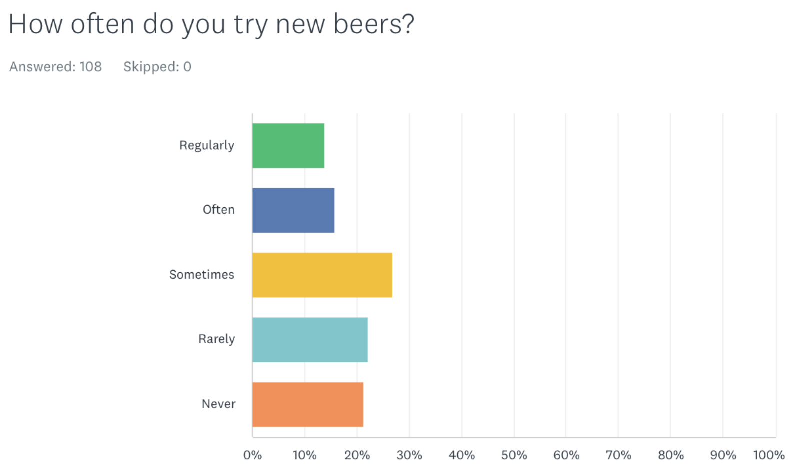


The participants were then asked what is important to them when choosing a beer and given multiple options to choose from with the ability to select as many as apply. Most participants (46.67%) stated that type of beer is important to them when choosing the beverage (Figure 6).. This was closely followed by 35.24% of

Fig.6: *Factors online sample considered important when choosing a beer*

participants making their choice based on criteria 'had that beer before', and with 26.67% making a choice based on 'had a beer by that brewery before'. These findings clearly show that a big majority of the sample is used to sticking to what they know and rarely try any new beers. The least amount of people (8.57%) shared that they make a choice based on their lifestyle preference (vegan/gluten free/non-alcoholic/other), supporting the decision to include a question about it earlier. Interestingly though, only 9 people said this factor is important to them, whereas in the earlier question 51 responses were for having a lifestyle preference. This suggests that while people do follow a vegan or gluten free diet, or avoid alcohol, some of them feel that it should not be constraining their beer choice.

The rest of the findings showed an even spread between the rest of the options (from 11.43% to 24.76%), meaning that friends' recommendations, price, bar staff' recommendations, tasting notes, name of beer, brewery, packaging/design, ABV and reviews on specialised websites are all also important when it comes to beer choice. The findings from this specific question also show the validity of the sample, as answers varied and included all options provided, therefore providing greater external validity. In order to provide a flexible range of answers for all participants, there was also 'Other (please specify)' option. This allowed me to collect more qualitative data and gain a better understanding of the participants in the sample. 14.29% of the sample chose this response, and when asked to specify most answers followed the notion that they 'don't know' what is important to them when making a beer choice. This again, highlights the need for more knowledge and understanding of the beer industry and variety of options within, thus proving the importance of this study.



In order to find out how experimental the participants are with their choice of beer, I asked how often they try new beers. They were presented with options to choose from: regularly, often, sometimes, rarely and never (Figure 7). Fig.7: *Frequency of trying new beers by online sample*

Most participants, 26.85% try new beers sometimes, closely followed by participants who rarely and never try new beers. These findings further highlight the lack of knowledge about diversity of flavours in the beer industry. There were, however participants who try new beers regularly (13.89%) or often (15.74%), thus proving the validity and importance of this research and that the interest to try new beers is there.

Predicting such a large proportion of responses from people who do not tend to try new beers, I wanted to gather more specific qualitative data from them, so participants who have stated that they rarely/never try new beers were asked to provide the reasons for this. I was able to gather several interesting responses which added great value to the investigation. One of the responses specifically said: 'I don't drink alcohol very often, only in social situations', there were other responses similar in notion, which highlights that some of the participants in the sample simply lack the knowledge that there are a lot of non-alcoholic beer options out there. Therefore this research is important, as the beer recommendation app can be used not only by alcohol drinkers, but can also be a useful tool for those wanting to know and explore non-alcoholic options. Interestingly, one participant shared that they rarely or never try new beer 'because usually non-alcoholic beer choices are limited', this clearly highlights that non-alcoholic options of beer need to gain popularity and be distributed and advertised more in our society. Another common response was 'because I like to stick to what I know', showing that some participants in the sample simply do not have the need to try new beers due to their personality or behavioural patterns. One participant answered that they do not try new beers often because 'I have an addictive personality', again highlighting that some participants do not try new beers as a life choice and not due to the lack of knowledge or tools. Another common and definitely worth noting response was 'cost', 'don't want to waste money' or 'I don't want to buy beers I don't like and won't finish'. These findings show that the participants have experienced trying new beers and have been left unsatisfied, thus highlighting the need for a better beer recommendation system which will leave users satisfied with their choice. In support of this view another response I would like to highlight is: 'I like light beers and most of the time I've tried others, I haven't enjoyed them'. This response, like many others, highlights the lack of good recommendation systems or tools in the beer tasting world.

The participants were then asked whether they use any beer recommendation apps, to which the findings are very important. The vast majority (93.52%) of the sample have stated that they do not use any. This shows the significance of the research this study poses, and underlines that a better recommendation system needs to be developed in order for beer recommendation apps to gain popularity and be used more. Amongst the small number of those who have stated that they have used an app with such a purpose before, the app 'Untappd' was the most commonly mentioned.

Finally, the survey asked participants to identify their age group, as the only personal data about them. Putting that question at the end of the survey was done with a purpose not to distract participants with such a trivial question, to which all would know the answer, at the start and keep their attention focused on answering more thought-provoking questions about beer preferences. It was still crucial to collect personal information about the participants, in order to have the ability to evaluate and place the findings in the context of society.So it further adds validity to the research, and helps to indicate why the participants have answered the way they have and what influences they have. Most participants (28.70%) identified themselves as part of the 35-44 years old age group, closely followed by 25-34 year olds (25%). Less than a fifth of the sample (18.52%) were reportedly aged between 45 and 54 years, and 13% of the sample were between 18 and 24 years old. The two least represented age groups were those aged over 65 years (8.3%) and 55-64 year olds, who made up 6.48% of the sample. Such distribution of ages suggests that the research conducted is valid and can be taken as representation of the general population.

The market research conducted both offline and online shows two important insights into the current beer consumption landscape:

- There is a considerable lack of motivation to try new beers which can largely be explained by a lack of knowledge of available options and how best to match them to a given consumer's taste.
- There is a space for an easy to use and 'beer novice friendly' recommendation system.

These insights have hence become the motivation for my technological investigation. Its goals can be summarised as to create an approach to recommending beer, which is a) easy to use, b) adaptable to any dataset, including non-traditional beers, c) able to provide varied recommendations to uninformed consumers, thus tackling the lack of knowledge outlined above.

Chapter 3: Methodology

3.1 Data collection

To attend to the goals outlined in the previous chapter I chose to utilise the content-based approach to building a recommender, as mentioned in Chapter 1. The starting point was building on my mini project completed for the 'Personalisation and Machine learning' module. I collected my own dataset from Ferment magazines published by Beer52 Craft Beer Club. This dataset (referred to as 'df_52' in the accompanying Jupyter notebook) includes descriptions of 101 beers from 75 breweries in 24 countries across 51 styles. There are both numerical and categorical data that provide a good platform to attempt building a content-based recommender system. Each beer in the dataset has information on its name, brewery, country where it is made, style, ABV, temperature it is best enjoyed at alongside the taste rankings across eight categories and tasting notes. The dataset was complemented with review scores and number of reviews available for each beer on the Beer52 website.

To expand the scope of my training data I combined my dataset with a 'Beer Profile and Ratings Data Set' ('df_BA' in the notebook), taken from a reviewing website BeerAdvocate.com and available at Kaggle. This dataset was in turn compiled from two other datasets from the same source: 'Beer Tasting Profiles Dataset' and '1.5 Million Beer Reviews'. In the combined dataset there are 11 columns representing the tasting profile features of every beer. They have been defined by word counts found in up to 25 reviews of each beer. The key assumption here is that people writing reviews are more than likely describing what they do experience rather than what they do not. Building on that assumption, I scaled the scores in each dataset to the same range of 0 to 1, and focusing only on the features present in both datasets, combined them into one dataframe. As a result, each beer has six scored aspects of its tasting profile: how hoppy, malty, sweet, fruity, sour and bitter it is.

To check that such assumption did not deter the quality of data for the purposes of the project, I looked at the distribution of data in each column in each of the three datasets, as shown in Appendix 1. While there are notable differences, for example, in mean averages, these can be explained by the source of the 'df_52' dataset. Collected from eight issues of Ferment magazine the 106 beers presented do not claim to be a fair representation of the whole landscape of available tastes, and as seen by the 'hoppy' column of the dataset description with mean=0.6 and 50 percentile=0.75, it is not normally distributed. So it is fair to assume that in the 'df_52' dataset there are more hoppy beers than in the general population, due to the custom selection of the Ferment magazine authors. Also, it is important to note that while 'df_52' tasting profiles have been made by presumably experts in the field, the 'df_BA' scores for the same aspects are derived from the consumers' perception of a given taste, and those consumers are not necessarily experts. However, it would be fair to assume that members of the public leaving reviews on the specialised beer-reviewing website do have an above zero level of expertise in that regard.

3.2 Stage one: Recommendations for desired tasting profile

With the dataset ready my first task was to create a framework that would provide a user with top-N recommendations based on their input of a desired tasting profile. The idea behind my approach was to be able to make recommendations with as little as possible information about the user. In the first stage all the user needs to do is to input their desired tasting profile, so how hoppy or malty they would like their beer to be. Based on that information the model suggests ten beers most similar. In addition, recommendations can be sorted by review scores.

To check the validity of the results, I asked five people to get their recommendations and provide feedback. At least one of the beers recommended to different people was available to acquire, so I asked my participants to try that beer and rate the recommendation on the scale from 1 to 10. The results are shown in Figure 8. It is worth mentioning that two out of five people surveyed felt confused by the required input of the beer's tasting profile, as they did not know what 'hoppy' or 'malty' tasted like. With that in mind, creating a recommender that would need even less input from a user was needed.

Fig. 8: *Feedback from Stage one recommendations*

Participant	Beer tried	Score	Comments
1	Prima Donna by Uiltje	9	'Loved it, would not have tried it if not suggested'
2	Lucky by Trouble Brewing	6	'Good beer, but a bit boring'
3	Session IPA by Popples	10	'Exactly like the beer I usually drink'
4	Extra Stout by Jopen	7	'I rarely drink stouts but this suggestion was not so bad'
5	Birra Moretti by Heineken	8	'Great lager'

3.3 Stage two: Recommendations from tasting notes

Following the logic of similarity based filtering another approach I could see to recommending beers was to ask a user what beer they had already enjoyed and would like to try something more like it. However, with the Stage one model it would mean that the user would have to have had a beer from the dataset, which only has 1955 titles. This in turn means that more beers need to be added to the dataset with the appropriate tasting profile. Unfortunately, most beers available in the physical world and online do not have scores on how hoppy, malty, sweet, sour, fruity and bitter they are. What most beers have is their name, brewery, style, ABV and tasting notes. Hence, the task is to use that information to predict the tasting profile scores and make recommendations as in Stage one model.

To use style as a factor in making recommendations additional preprocessing was needed, since our dataset had 162 unique styles. Upon examination, it could be seen that a lot of them were the same styles written differently. After removing duplicates, there were 157 unique styles left, which was still too high. Most of this came from the distinction within a style, such as: 'Porter - American', 'Porter - Baltic', 'Porter - English', 'Porter - Imperial', 'Porter - Robust' and 'Porter - Smoked'. To unite entries like this into one broader category a new column called 'Type' was made, which after numerous steps of processing had only 43 unique entries. It is important to note that while in my market research style and type of beer were used interchangeably, when working with the dataset there was an important distinction, in that 'type' was a broader category encompassing different styles of the same nature and made to get a better visualisation of the data.

In the processed dataset ten most common types of beer were lager, with 14.2% of all entries belonging to that category. Second most popular was stout (9.5%), followed by IPA (7,7%). In the top five most popular types there was also porter (5.83%) and wheat beer (5.78%). The next five most common types were strong ale, pale ale, bock, brown ale and red ale, with less than 100 entries each, and ranging 3-4% of all entries. Given the ten most popular types, I looked at the relationship between each taste ranking and ABV (Figure 9). It can clearly be seen that there are clusters of particular types of beer, for example, stouts tend to be less sour and fruity regardless of the alcohol content, while IPAs tend to have slightly higher than average ABV and be more hoppy and more fruity.

Fig. 9: *Distribution of taste rankings across ABV spectrum, colour-coded by type*

Figure 10 in turn shows the relationship between taste rankings and the overall review scores, also with a possibility to draw conclusions about ten most popular types. It is interesting to note that lagers tend to have lower review scores and lower than average scores on every aspect of the tasting profile.

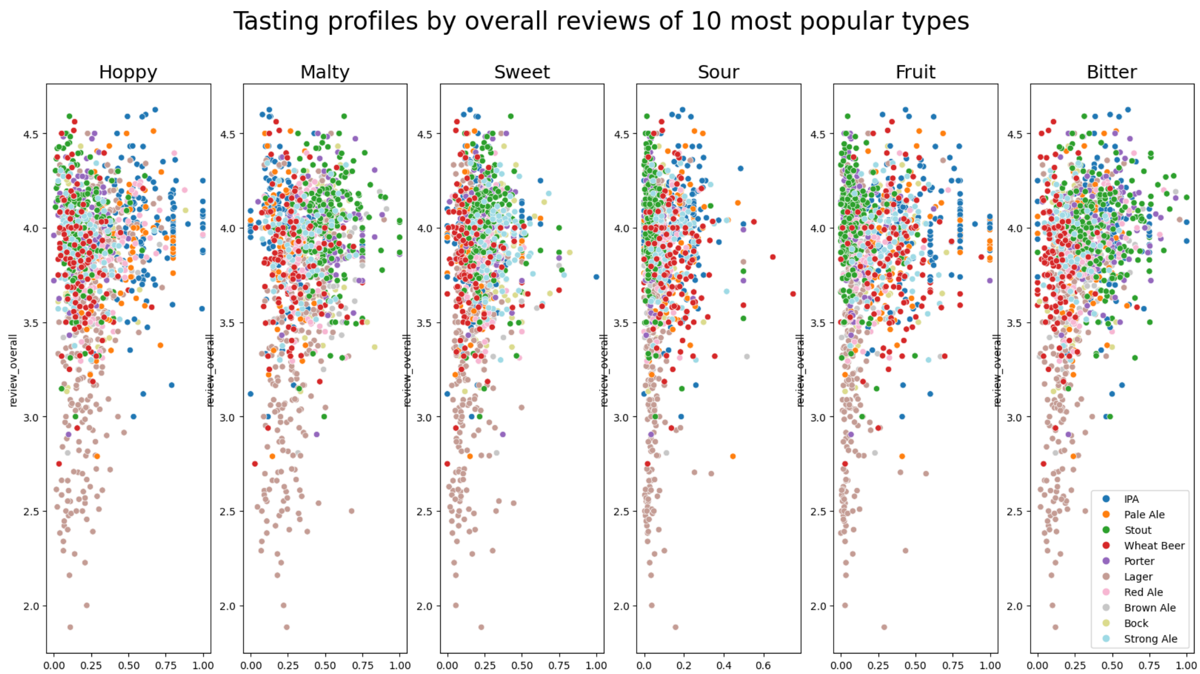


Fig. 10: *Distribution of taste rankings across review scores spectrum, colour-coded by type*

For the purpose of using tasting notes as a means to build the six-dimensional tasting profile of a new beer, the Word2Vec model proved useful. Preparing words from the dataset for Word2Vec involved determining a list of stopwords, i.e words not to be considered when building vector representations of each string. Besides the default list of English stopwords from Natural Language Toolkit (NLTK) library (Judah, 2021), the final list includes names of styles and types of beer, along with words relating to the beer production but not carrying any important for the project's purposes meaning (e.g. 'brewed', 'whirlpool', 'finishes', 'tastes').

Since the used version of Word2Vec made 300-dimensional vectors, dimensionality reduction was needed to see any patterns present. Adapting the approach outlined by Chambliss (2019), I used t-distributed stochastic neighbour embedding (t-SNE). For better illustration of patterns, t-SNE was used only on half the dataset. An example of the output is shown in Figure 11.

Fig. 11: *Mapping of the individual words from the tasting notes*

The next step was to create a vector for every tasting note rather than individual words in it. To accelerate the preprocessing custom functions outlined in Spathis (2021) proved useful. The visualisation of the dimensionality reduction on half the dataset is shown in Figure 12, with point labels identifying the names of the beer corresponding to the vectorised tasting notes.

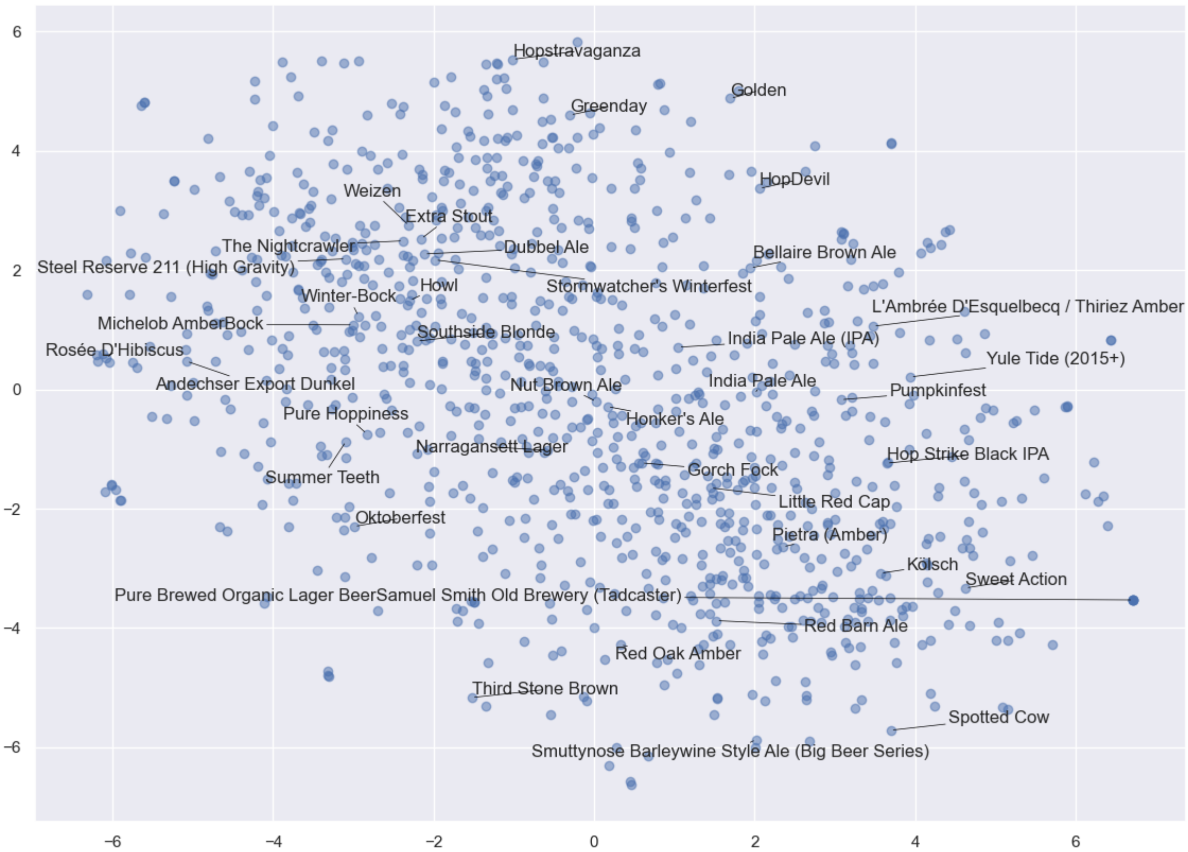


Fig.12: *Mapping of vectors of the tasting notes*

With the data ready and explored, the model could be trained. Using 80% of the data for training and 20% for testing, I trained the model so it would predict a value of how hoppy, malty, sweet, sour, fruity or bitter the beer is based on its tasting note, represented in a 10-dimensional vector. The mean squared error after fitting the model stood at 0.037 which is suitable for this context.

The purpose of this model is so that a user could put in a tasting note of any beer, perhaps the one they already enjoyed, and receive 10 beers most similar to it. To test out the model, I collected another dataset by scraping the Great British Beer Festival online programme. This dataset ('gbbf' in the notebook) only has the name, style, brewery, ABV and tasting notes of 455 beers. The notebook lets the user choose any beer from the dataset and use its tasting note to receive recommendations. Alternatively, a user can type in their own tasting note. The model created also leaves space to input ABV or style of beer and get recommendations based on those factors as well.

Chapter 4: Conclusion

The model I created lets a user get recommendations for any beer they liked based on similarity. It is a highly adaptable approach as it can include any beers available online or offline, and there is possibility to keep extending it and training on more data to achieve greater results. The goals of the project outlined in Chapter 2 can be considered achieved. The model is easy to use and a user only needs to input one value (a tasting note). The model can also work with non-traditional types of beers and there is possibility to include that filter in the final results, so that a user, for instance, receives only gluten free or vegan recommendations. The recommendations provided by the model are varied, as can be seen in the example in the accompanying notebook, where the top 10 most similar beers to the one chosen from GBBF dataset are all of different styles, to encourage user's exploration of them.

However, while recommendations may make sense on paper, the 'you'll-probably-like-this-too' approach does have its limitations, as mentioned in Chapter 1. The satisfaction of the users with my model's recommendations may not be in direct causation of their quality, as human experience is highly subjective. Yet, the main goal of this project was to create an adaptable approach that would promote exploring different beers without the need of fancy terminology, and it could be argued that it has been successful in that.

Future work on this approach could include creating a user-friendly interface to the model, such as an app, collaborating with bars and breweries to promote their products and including non-traditional beers into the dataset and recommendations.

Bibliography

Al-Rifaie, M.M., Cavazza, M., 2022. *Multi-Objective Evolutionary Beer Optimisation.* arXiv preprint arXiv:2204.02752. [Accessed 27 August 2022].

Arbor, A., 2022. [online] Available at: <https://www.mlive.com/news/ann-arbor/2022/09/michigan-brewery-map-app-adds-hard-ciders-meads.html> [Accessed 20 August 2022].

Bobadilla, J., Ortega, F., Hernando, A., Bernal, J., 2012. *A collaborative filtering approach to mitigate the new user cold start problem*. Knowledge-Based Systems. 26: 225–238. Available at: <doi:10.1016/j.knosys.2011.07.021> [Accessed 25 August 2022].

Brown, P., 2017. *The Story of Craft Beer*, Beer52. [Accessed 20 August 2022].

Conway, J., 2022. *Beer vendor market share worldwide 2021 | Statista*. [online] Statista. Available at: <https://www.statista.com/statistics/257677/global-market-share-of-the-leading-beer-companies-based-on-sales/> [Accessed 26 August 2022].

Chambliss, C., 2019. *Using word2vec to Analyze News Headlines and Predict Article Success*. [online] Medium. Available at: <https://towardsdatascience.com/using-word2vec-to-analyze-news-headlines-and-predict-article-success-cdeda5f14751> [Accessed 24 August 2022].

Datadoume, 2018, *Beer reviews* dataset, Available at: <https://www.kaggle.com/datasets/rdoume/beerreviews> [Accessed 5 June 2022].

Edmonton International Beerfest. 2022. *Lovers of Beer: 5 Best Beer Apps - Edmonton International Beerfest*. [online] Available at: <https://www.internationalbeerfest.com/lovers-of-beer-5-best-beer-apps/> [Accessed 21 August 2022].

Ferment magazine, 2021-2022, issues 74-81, published by Ferment & Beer 52, Edinburgh, Scotland. [Accessed 24 August 2022].

Furnari, C., 2020. *Beer Advocate Assets Sold To Untappd Parent Company Next Glass*. [online] Forbes. Available at: <https://www.forbes.com/sites/chrisfurnari/2020/02/26/beer-advocate-assets-sold-to-untappd-parent-company-next-glass/?sh=8be8dad34576> [Accessed 25 August 2022].

Gutierrez, A., 2022 *How do you create a successful SXSW experience? — Seek and you will find.*. [online] Seek and you will find. Available at: <https://www.iaskalotofquestions.com/work/2017/5/7/sxsw-2016 > [Accessed 27 August 2022].

Hop Culture. 2022. *The Beer Apps You Need To Have on Your Phone • Hop Culture*. [online] Available at: <https://www.hopculture.com/the-best-phone-apps-for-drinking-beer/> [Accessed 15 August 2022].

Judah, B., 2021. *Removing stop words with NLTK library in Python*. Medium. Available at: <https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1> [Accessed: 26 August 2022]

Kaggle.com. 2022. *Beer Profile and Ratings Data Set*. [online] Available at: <https://www.kaggle.com/datasets/ruthgn/beer-profile-and-ratings-data-set> [Accessed 23 August 2022].

Kane, F., 2018. Building Recommender Systems with Machine Learning and AI. Sundog Education. [Accessed 23 August 2022].

Mikolov, T., Chen, K., Corrado, G., Dean, J. 2013. *Efficient Estimation of Word Representations in Vector Space.* Available at: <https://arxiv.org/abs/1301.3781> [Accessed 23 August 2022].

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. 2013. *Distributed Representations of Words and Phrases and their Compositionality.* Available at: <https://arxiv.org/abs/1310.4546> [Accessed 22 August 2022].

Misfud, C., 2018. *Why Beer Is the World's Most Beloved Drink*. [online] Time. Available at: <https://time.com/5407072/why-beer-is-most-popular-drink-world/> [Accessed 17 August 2022].

Mudambi, S. and Schuff, D., 2010. *Research Note: What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com on JSTOR*. [online] Jstor.org. Available at: <https://www.jstor.org/stable/20721420> [Accessed 26 August 2022].

Myers, Q., 2022. *The Inside Story of iBeer, the Underdog Beer App That Made Millions*. [online] MEL Magazine. Available at: <https://melmagazine.com/en-us/story/ibeer-app-history> [Accessed 26 August 2022].

Nye, Mary Jo, 2016. *Speaking in Tongues: Science's centuries-long hunt for a common language. Distillations*. 2 (1): 40–43. Retrieved 22 March 2018. [Accessed 24 August 2022].

Otter, D., Medina, J., Kalita, J. 2018. A Survey of the Usages of Deep Learning in Natural Language Processing. Available at: <https://arxiv.org/abs/1807.10854> [Accessed 25 August 2022].

Spathis, D., 2021, *average-word2vec* repository, Available at <https://github.com/sdimi/average-word2vec/blob/master/notebook.ipynb> [Accessed 20 August 2022].

Shelton, A., 2015. *Can Tech Replace the Human Element of Beer Recs? We're Skeptical*. [online] Outside Online. Available at: <https://www.outsideonline.com/food/introducing-ultimate-beer-recommendation-app/> [Accessed 26 August 2022].

Singhal, A., 2001. *Modern Information Retrieval: A Brief Overview*. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43. [Accessed 23 August 2022].

SMI Aware. 2017. *Investigator's Guide to Investigating Untappd - Social Media Information | SMI Aware*. [online] Available at: <https://smiaware.com/blog/investigators-guide-untappd/> [Accessed 20 August 2022].

Sp1222, 2020, *Beer Information - Tasting Profiles* dataset, Available at <https://www.kaggle.com/datasets/stephenpolozoff/top-beer-information> [Accessed 5 June 2022].

The Beer Connoisseur. 2016. *Best Beer Apps for iOS and Android*. [online] Available at: <https://beerconnoisseur.com/articles/best-beer-apps-ios-and-android> [Accessed 21 August 2022].

The Great British Beer Festival - Latest Updates. (n.d.). *The Great British Beer Festival* - Latest Updates. Available at: <https://gbbf.realaledb.com/LiveBar.php> [Accessed 25 August 2022]

Twardzik, T., 2021. *The 5 Best Apps for Finding, Rating, and Sharing Craft Beer*. [online] MUO. Available at: <https://www.makeuseof.com/best-apps-for-finding-rating-sharing-craft-beer/> [Accessed 19 August 2022].

Untappd, I. and Team, T., 2022. *Untappd - Drink Socially*. [online] Untappd. Available at: <https://untappd.com> [Accessed 19 August 2022].

Wilson, R.,ThankBeer: *A Beer Recommendation Engine Final Report*. Available at: <http://cs229.stanford.edu/proj2012/Wilson-BeerRecommendationEngine.pdf > [Accessed 23 June 2022].

Yakushova, A., 2022, *PML mini project* repository, Available at: <https://git.arts.ac.uk/21038970/PML_mini_project> [Accessed 1 September 2022]

Yakushova. A., 2022, Thesis repository, Available at: <https://git.arts.ac.uk/21038970/thesis>

Appendix

```
df_BA['Hoppy'].describe()

count    1849.000000
mean        0.266970
std         0.179415
min         0.000000
25%         0.126667
50%         0.233333
75%         0.380000
max         1.000000
Name: Bitter, dtype: float64
```

```
df_BA['Malty'].describe()

count    1849.000000
mean        0.336301
std         0.172058
min         0.000000
25%         0.200837
50%         0.322176
75%         0.460251
max         1.000000
Name: Malty, dtype: float64
```

```
df_BA['Sweet'].describe()

count    1849.000000
mean        0.231958
std         0.127808
min         0.000000
25%         0.140684
50%         0.212928
75%         0.296578
max         1.000000
Name: Sweet, dtype: float64
```

```
df_52['Hoppy'].describe()

count    106.000000
mean       0.609434
std        0.279294
min        0.000000
25%        0.400000
50%        0.750000
75%        0.800000
max        1.000000
Name: Hoppy, dtype: float64
```

```
df_52['Malty'].describe()

count    106.000000
mean       0.419811
std        0.267922
min        0.000000
25%        0.250000
50%        0.250000
75%        0.687500
max        1.000000
Name: Malty, dtype: float64
```

```
df_52['Sweet'].describe()

count    106.000000
mean       0.353774
std        0.213720
min        0.000000
25%        0.250000
50%        0.250000
75%        0.500000
max        1.000000
Name: Sweet, dtype: float64
```

```
data['Hoppy'].describe()

count    1955.000000
mean        0.277601
std         0.206038
min         0.000000
25%         0.116279
50%         0.220930
75%         0.389535
max         1.000000
Name: Hoppy, dtype: float64
```

```
data['Malty'].describe()

count    1955.000000
mean        0.340829
std         0.179480
min         0.000000
25%         0.205021
50%         0.322176
75%         0.464435
max         1.000000
Name: Malty, dtype: float64
```

```
data['Sweet'].describe()

count    1955.000000
mean        0.238563
std         0.136618
min         0.000000
25%         0.144487
50%         0.216730
75%         0.304183
max         1.000000
Name: Sweet, dtype: float64
```

```
df_BA['Sour'].describe()

count    1849.000000
mean        0.127244
std         0.139030
min         0.000000
25%         0.038732
50%         0.080986
75%         0.154930
max         1.000000
Name: Sour, dtype: float64
```

```
df_BA['Fruit'].describe()

count    1849.000000
mean        0.238473
std         0.194047
min         0.000000
25%         0.074286
50%         0.182857
75%         0.371429
max         1.000000
Name: Fruit, dtype: float64
```

```
df_BA['Bitter'].describe()

count    1849.000000
mean        0.266970
std         0.179415
min         0.000000
25%         0.126667
50%         0.233333
75%         0.380000
max         1.000000
Name: Bitter, dtype: float64
```

```
df_52['Sour'].describe()

count    106.000000
mean       0.149764
std        0.223886
min        0.000000
25%        0.000000
50%        0.000000
75%        0.250000
max        1.000000
Name: Sour, dtype: float64
```

```
df_52['Fruit'].describe()

count    106.000000
mean       0.565094
std        0.309236
min        0.000000
25%        0.250000
50%        0.600000
75%        0.800000
max        1.000000
Name: Fruit, dtype: float64
```

```
df_52['Bitter'].describe()

count    106.000000
mean       0.391509
std        0.247051
min        0.000000
25%        0.250000
50%        0.500000
75%        0.500000
max        1.000000
Name: Bitter, dtype: float64
```

```
data['Sour'].describe()

count    1955.000000
mean        0.128465
std         0.144915
min         0.000000
25%         0.035211
50%         0.080986
75%         0.158451
max         1.000000
Name: Sour, dtype: float64
```

```
data['Fruit'].describe()

count    1955.000000
mean        0.256183
std         0.214997
min         0.000000
25%         0.080000
50%         0.194286
75%         0.400000
max         1.000000
Name: Fruit, dtype: float64
```

```
data['Bitter'].describe()

count    1955.000000
mean        0.273722
std         0.185793
min         0.000000
25%         0.126667
50%         0.240000
75%         0.386667
max         1.000000
Name: Bitter, dtype: float64
```

Appx 1: *Distribution of scaled tasting scores in two datasets before concatenation and after*