Apply Machine Learning to Recommend Beer

Stephen Polozoff

September 2020

1 Introduction

The brewing industry has grown and evolved in the United States since the Tied House Law ended in the late 1970s, a prohibition-era law that prohibited breweries from being able to sell their product. Not long after the law ended, there were only a couple of hundred breweries in the United States; now, there are over 8,000 breweries across 50 states with each brewery producing their lineup of beers [2]. These breweries can produce a wide variety of options, ranging from signature beers and other products available year-round to the microbrews that are only available seasonally or made as one-time craft experiments. At an average of 16 choices of beer to choose from produced by a brewery, that is over 128,000 options to choose from in the US alone.

The shelves showcasing beer available at local groceries, gas stations, and individual sellers can be overwhelming for those wanting to explore their options. With so many to choose from, it can be challenging for the curious beer connoisseur to try something new without background or history. Beer ranking websites such as RateBeer and BeerAdvocate provide a sense of validation and insight regarding a beer's properties readily available for those needing to research their potential purchase. However, this one size fits all method has its limits. Rankings and ratings may give a researcher a general idea of how the beer is from someone else's perspective. Still, it will not definitively tell them if the beer they are considering will meet their expectation.

Machine learning projects have been done to provide a sense of direction regarding the next beer they should pick up from the shelf. One project utilized the collaborative filtering method by analyzing the reviews left on a popular beer ranking website BeerAdvocate. They collected review data of the top 100 ranked beers in each beer style [4]. They utilized user behavior and posting data to find a connection or similarity with other users and products [4]. They also used the latent factor approach, narrowing the search down based on a few factors. Another project took data set from RateBeer of the top 25 to 50 beers from each state, regardless of beer style [3]. They took the top 100 most frequent words and top 100 distinct words by TF-IDF score and utilized a cursory LDA model, ending up with two topics of beers: dark beers of coffee and chocolate characteristics and light beers of fruity and hoppy characteristics [3].

Using machine learning techniques, we can reduce the shadow of doubt considerably. Being able to scrape the existing data sets of the beer's ratings and reviews regarding its characteristics across all beer styles, we will be able to further narrow down the choices of beers with the option of presenting new styles the user may like to try. We will utilize other projects' ideas and incorporate characteristics such as International Bitterness Units (IBU), Alcohol by Volume (ABV), and reviews and ratings of beer.

The phases of this research will be done incrementally, completing the current objective, and each result being validated before moving on to the next phase. The research will require a keyword data bank as definitions of the vectors in the spider graph, which will be used to compare with the keywords found in the reviews on BeerAdvocate.com. Once we have our data bank set up, the next step will be to scrape data of up to 50 beer from each beer style according to BeerAdvocate, including but not limited to beer names, styles, alcohol by volume, average rating, number of ratings, and user reviews. With the information we scrape, we will gather word counts from user reviews to compare with the data bank of words initially created. The keywords found in reviews compared to the keyword data bank will give an overall description of each beer's characteristics, applying an order of magnitude towards a vector on the spider graph. One challenge will be to isolate word combinations implying a negative feedback on an aspect of the beer to apply a negative order of magnitude to a vector on the spider graph. Other factors will also be included, such as international bitterness units and alcohol by volume. Training data provided by a user will indicate what beer will need to be scraped for data to build a profile of each user-preferred beers. We will map out the user's training data which will provide a comparable model to the data of beer we have previously scraped. We will cross-reference the training data results with the data that we have initially gathered. The ideal result will be to match similar profiles of beers from the training data with those initially compiled in bulk from BeerAdvocate in order to make an accurate list of recommendations that will appeal to the user.

Project Timeline:

Phase 1: Data Bank Built (October 2020) Data bank of keywords put together that describes a beer's profile, including but not limited to taste, aroma, and alcohol content.

Phase 2: Data of Top 50 Beer Options from Each Style Collected (October 2020) Web Scrape data from BeerAdvocate to help adjust keyword data bank and build profiles of individual beers and styles.

Phase 3: Build Beer and Beer Style Profiles (October – November 2020) Take the data that was web scraped and build profiles of each beer and beer styles.

Phase 4: Test Against Training Data (November 2020 – December 2020) Apply training data to the model, web scrape information needed to build a profile of the training data to compare with the profile collection.

Phase 5: Publication and Presentations (December – 2020) The research report will be compiled for peer and mentor review, and a presentation to discuss in depth the process of the project.

References:

[1] "National Beer Sales Production Data." Brewers Association, 17 Aug. 2020, www.brewersassociation.org/statistics-and-data/national-beer-stats/.

[2] "NINKASI: Beer Recommender System." NYC Data Science Academy, nycdatascience.com/blog/student-works/ninkasi-beer-recommender-system/.

[3] Xie, Medford. "What to Drink Next?-A Simple Beer Recommendation System Using Collaborative Filtering." Medium, Medium, 19 June 2019, medium.com/@medfordxie/what-to-drink-next-a-simple-beer-recommendation-system-using-collaborative-filtering-b65dd32b600d.