# Data301

## Ismail Sarwari

***isa50***
**73712637**

***3/6/2024***

### Abstract

This project focuses on analysing beer reviews from the BeerAdvocate dataset , and diving deep into exploring beer style similarities and analysing words associated with each beer style . The dataset used is sourced from BeerAdvocate and consists of user reviews, with the primary research goal is to determine how users word varies with users reviews based on the similarity between different beer styles. The algorithm used for the project includes cosine similarity calculation and a priori . The intended result is to provide insights into beer style similarities and the most frequent terms associated with each style. This project can provide valuable information for both beer enthusiasts and brewers by offering a deeper understanding of popular beer types and people's opinions about them.With analysing the data the brewers can have a better understanding of which beer appeals to their target audience,as well as what words could attract their attention.

# Introduction

### Background

The project uses the dataset provided by the University of California San Diego (UCSD) research lab, which then has been imported and processed to extract relevant information for analysis. Specifically, the dataset contains user reviews and ratings in the BeerAdvocate data, offering detailed information into various beer products.

The algorithm used in this project operates within the Dask environment, a powerful tool for parallel computing. Dask helps with distributing tasks across multiple parts of the computer simultaneously, allowing it to 'break up large computations and route parts of them efficiently onto distributed hardware' (Dask Documentation, no date). Also, Dask development is supported by developer communities from Pandas, Numpy, Scikit-Learn, Scikit-Image, Jupyter, and others (Dask Documentation, no date).

### Motivation

I'm curious about how people's reviews of beer relate to similar beer styles. I want to analyse the data using two different algorithms - cosine similarity and a priori analysis - to identify similar beer styles and the most frequently used words in reviews for the similar beer style. This research could provide valuable insights into consumer preferences and help brewers understand what aspects of their beers are most appreciated by customers. These findings will offer a detailed exploration of the important characteristics of various beers, providing valuable insights into what makes a great beer. Such insights are not only beneficial for users seeking the best beers but also for industry stakeholders, as they can inform product development and marketing strategies.

### Research Question

“How users' feedback on the beer  word varies based on the similarity between different beer styles”,with this research question in mind , the plan is to  uncover patterns and relationships that may rise within analysing similar beer and words used in the  reviews of similar beer. Implementing algorithms like cosine similarity and a priori analysis will help answer this question by looking at the similarity between different beer styles based on the reviews they receive. Cosine similarity will allow us to measure the similarity of review patterns between beer styles, while a priori analysis will help us identify the most frequently occurring words in these reviews, providing insight into the key attributes that consumers consider when evaluating beers. By applying these algorithms to the dataset, we can find patterns and trends that can inform breweries and marketers about consumer preferences and guide product development and marketing strategies.

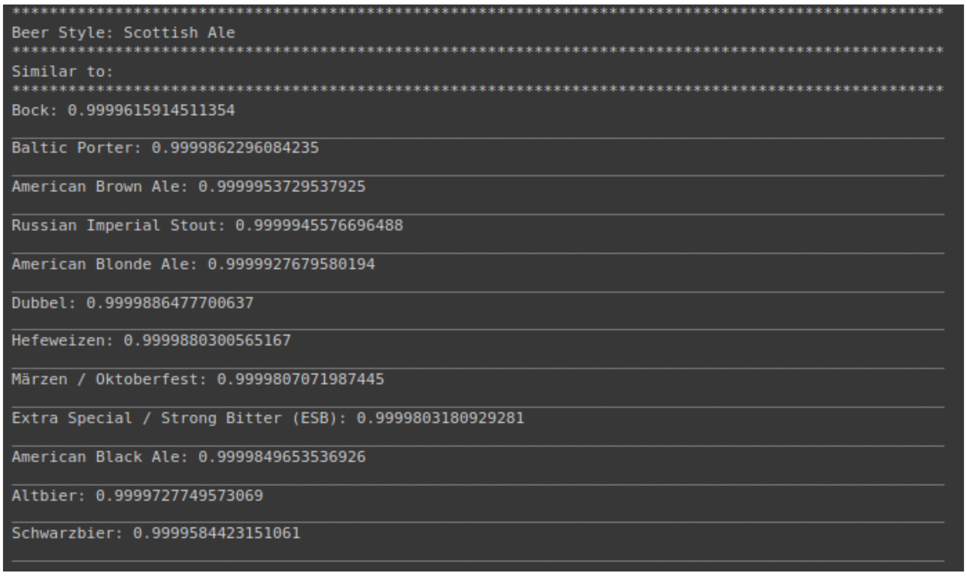
## Experimental Design and Methods

The project's approach involves a series of steps to achieve its goals efficiently. First, it retrieves the dataset from a URL and saves it locally, ensuring data integrity by converting it from ".gz" to JSON format, following conversion of JSONLINES. Then, it loads the data into a Dask Dataframe using Dask Bag, which helps handle large datasets effectively and splits it into manageable partitions for processing.

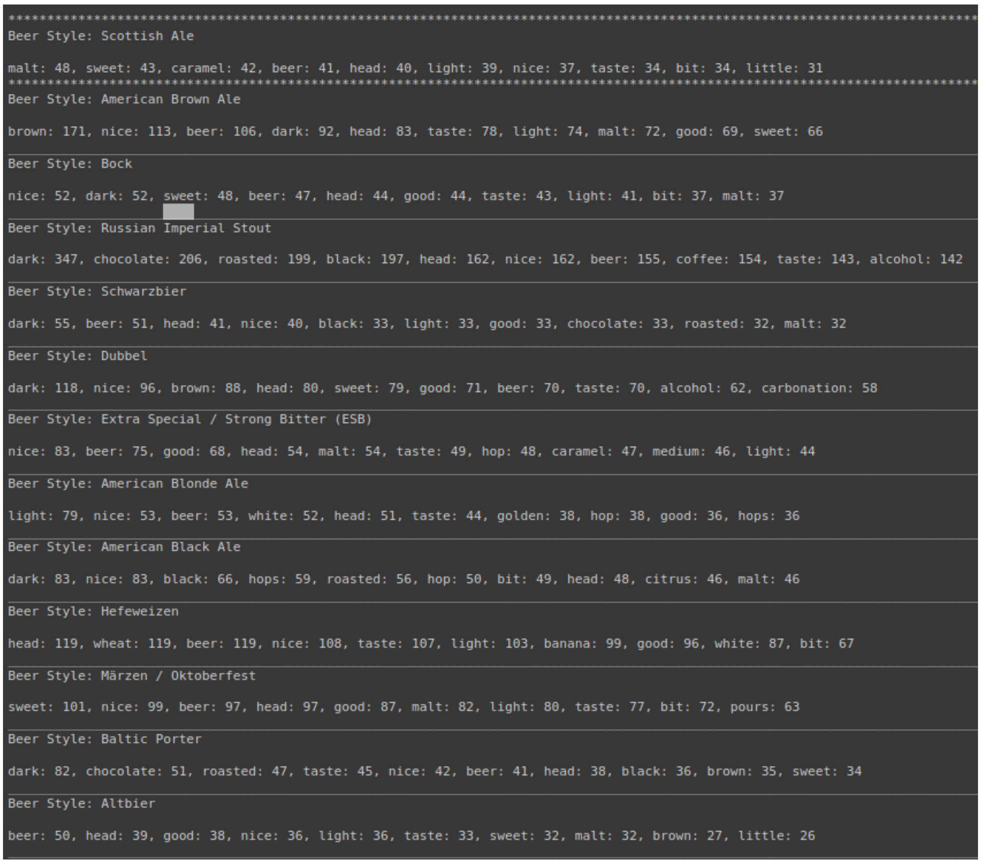
Then the project computes the data to find similarities between different beer styles and key words in the reviews. It does this by parsing the review data, extracting key information like beer style and review attributes such as appearance and taste. Then, it calculates the average attributes for each beer style. Using cosine similarity, it measures how similar different beer styles are based on their attributes. Additionally, the project follows by removing unnecessary words (stopwards) from the reviews and identifies the most frequent ones for each beer style. These processes are done with Dask for efficient handling of large data volumes.

Libraries and Codes Used
1. urllib.request: Used to retrieve data from a URL.
2. gzip and shutil: Used for decompressing ".gz" files.
3. jsonlines and json: Used for converting data to the JSONLines format and handling JSON data.
4. dask.bag: Used for efficient processing of large datasets and performing transformations.
5. numpy: Used for numerical operations and calculations.
6. random: Used for generating random samples from the dataset.
7. nltk: Library used for natural language processing tasks such as handling stop words.
8. create_sample: used for creating small amounts of data from the dataset.
9. extract_beer_style_and_rating: used to extract the reviews of users and beerstyle
10. beer_style_vectors - Used for making a mean vector for each beer style.
11. cosine_similarity: used for calculating the similarity between two vectors.
12. calculate_cosine_similarity: used for calculating the similarity of beer style to other beer styles.
13. a_priori_step1: used for calculating the most frequent word in the reviews
14. beer_style_vectors: used for finding the most frequent words in each beer style.

## Results

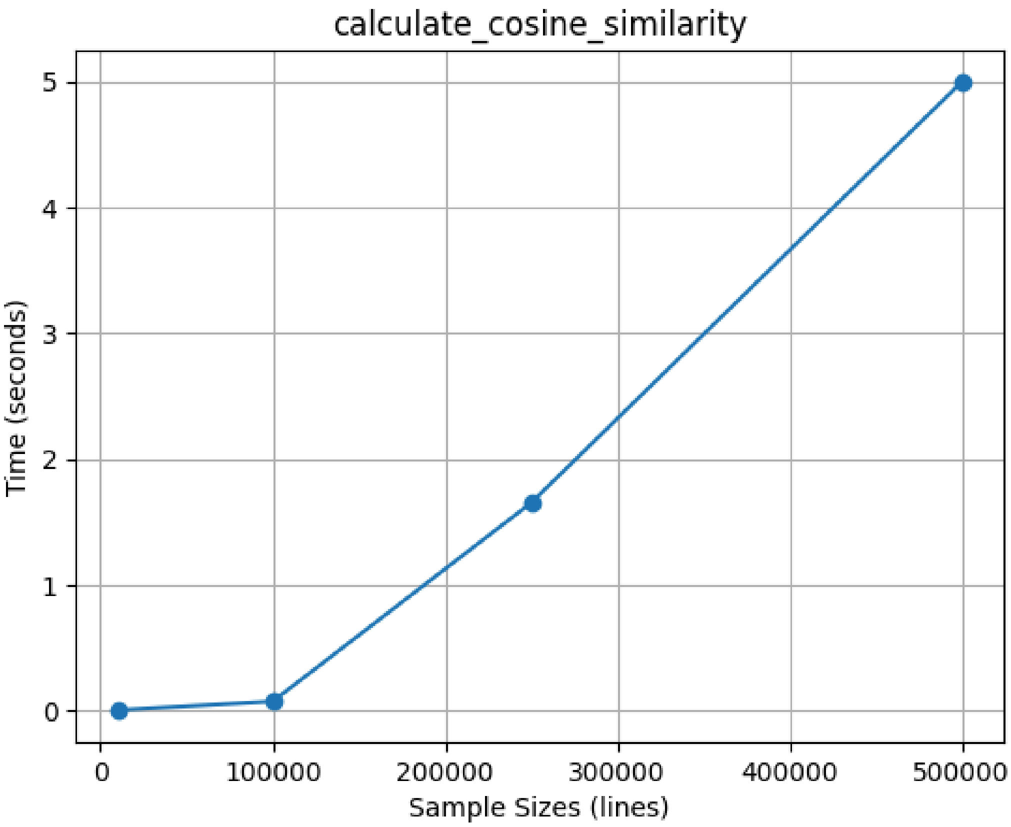


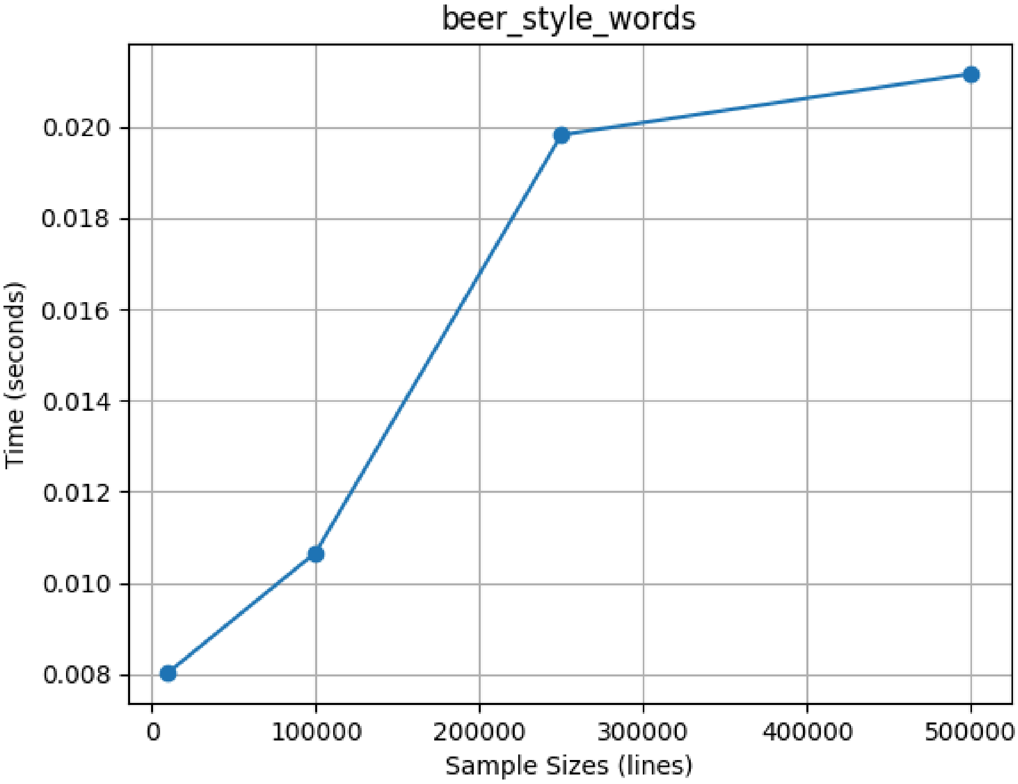
```
****************************************************************************************************
Beer Style: Scottish Ale
****************************************************************************************************
Similar to:
****************************************************************************************************
Bock: 0.9999615914511354

Baltic Porter: 0.9999862296084235

American Brown Ale: 0.9999953729537925

Russian Imperial Stout: 0.9999945576696488

American Blonde Ale: 0.9999927679580194

Dubbel: 0.9999886477700637

Hefeweizen: 0.9999880300565167

Märzen / Oktoberfest: 0.9999807071987445

Extra Special / Strong Bitter (ESB): 0.9999803180929281

American Black Ale: 0.9999849653536926

Altbier: 0.9999727749573069

Schwarzbier: 0.9999584423151061
```



```
**************************************************************************************************************************
Beer Style: Scottish Ale

malt: 48, sweet: 43, caramel: 42, beer: 41, head: 40, light: 39, nice: 37, taste: 34, bit: 34, little: 31
**************************************************************************************************************************
Beer Style: American Brown Ale

brown: 171, nice: 113, beer: 106, dark: 92, head: 83, taste: 78, light: 74, malt: 72, good: 69, sweet: 66

Beer Style: Bock

nice: 52, dark: 52, sweet: 48, beer: 47, head: 44, good: 44, taste: 43, light: 41, bit: 37, malt: 37

Beer Style: Russian Imperial Stout

dark: 347, chocolate: 206, roasted: 199, black: 197, head: 162, nice: 162, beer: 155, coffee: 154, taste: 143, alcohol: 142

Beer Style: Schwarzbier

dark: 55, beer: 51, head: 41, nice: 40, black: 33, light: 33, good: 33, chocolate: 33, roasted: 32, malt: 32

Beer Style: Dubbel

dark: 118, nice: 96, brown: 88, head: 80, sweet: 79, good: 71, beer: 70, taste: 70, alcohol: 62, carbonation: 58

Beer Style: Extra Special / Strong Bitter (ESB)

nice: 83, beer: 75, good: 68, head: 54, malt: 54, taste: 49, hop: 48, caramel: 47, medium: 46, light: 44

Beer Style: American Blonde Ale

light: 79, nice: 53, beer: 53, white: 52, head: 51, taste: 44, golden: 38, hop: 38, good: 36, hops: 36

Beer Style: American Black Ale

dark: 83, nice: 83, black: 66, hops: 59, roasted: 56, hop: 50, bit: 49, head: 48, citrus: 46, malt: 46

Beer Style: Hefeweizen

head: 119, wheat: 119, beer: 119, nice: 108, taste: 107, light: 103, banana: 99, good: 96, white: 87, bit: 67

Beer Style: Märzen / Oktoberfest

sweet: 101, nice: 99, beer: 97, head: 97, good: 87, malt: 82, light: 80, taste: 77, bit: 72, pours: 63

Beer Style: Baltic Porter

dark: 82, chocolate: 51, roasted: 47, taste: 45, nice: 42, beer: 41, head: 38, black: 36, brown: 35, sweet: 34

Beer Style: Altbier

beer: 50, head: 39, good: 38, nice: 36, light: 36, taste: 33, sweet: 32, malt: 32, brown: 27, little: 26
```

When analysing Scottish Ale beer style, it's noticeable that the word "sweet" frequently appears in the reviews. Interestingly, upon examining other beer styles, it becomes apparent that many of them also contain mentions of sweetness in user reviews.


calculate_cosine_similarity
Time (seconds)
0
1
2
3
4
5
0
100000
200000
300000
400000
500000
Sample Sizes (lines)


beer_style_words
Time (seconds)
0.008
0.010
0.012
0.014
0.016
0.018
0.020
0
100000
200000
300000
400000
500000
Sample Sizes (lines)

The function that finds the most frequent words for each beer style scales well with larger datasets, which is to be expected. However, when it comes to calculating cosine similarity, it doesn't work as efficiently with larger amounts of data. This might be because it uses a lot of memory, especially when sorting the data and selecting the top 5 items. Currently, the function uses list type to get top 5 frequent beer styles, instead of the optimised 'topk' function from Dask Bag. To improve this process, could enhance it by working with the Dask environment, which is better suited for handling large-scale data operations.

## Conclusion

The analysis made in this project gives significant insights into users preferences and ratings within the BeerAdvocate dataset, effectively addressing similar words that appear with each type of beer style. Many data processing and analysis techniques were used, such as calculating the average values for each beer style and measuring similarities through cosine similarity analysis. The cosine similarity takes in factors like taste, smell, looks, and feel and gives a similar beer style based on those factors.Additionally, the A Priori algorithm was used to discover the most frequently occurring words. Combining the outcomes of these two algorithms allows for a comprehensive understanding of user preferences and the identification of common themes in beer reviews.

The implications of these results are significant for both beer enthusiasts and industry stakeholders. For users, the insights gained from this analysis can enhance their beer selection process by providing a better understanding of the attributes that contribute to their enjoyment of different beer styles. Additionally, industry stakeholders, including breweries and marketers, can leverage these findings to inform their product development strategies and marketing campaigns. By focusing on attributes that resonate with consumers and drive positive ratings, breweries can tailor their offerings to better meet consumer preferences, leading to increased customer satisfaction.

Looking ahead, there are a few ways the project can be analysed further. One idea is to dig deeper into user reviews to spot specific words or phrases that show whether people liked or disliked a beer. find what people like, what they want and what their major concerns are (Parthvi Shah, no date) This could help us understand what people enjoy within each beer style, what they want more of, and what they're not so happy about. Use of fancy built in library TextBlob tools to find out feelings of users from the text review and get a better idea of what users like or dislike about the beer style and how these carried on with similar beer style.

## Critique of Design and Project

One aspect of the project design that could have been improved is the method used for calculating cosine similarity between beer styles. While the approach of calculating cosine similarity between mean vectors of beer style attributes is a decent technique at best , it may not fully capture the complexity and intentions of user preferences and ratings. The method assumes that each beer style can be represented by taking the mean vector of attributes, which may oversimplify the diversity within each beer style category, and may miss the users preferences, potentially leading to inaccuracies in similarity assessments.

To address these limitations, a more sophisticated approach could have been implemented, such as creating user ratings or review with  weights in the calculation of cosine similarity. By assigning weights to attributes based on their importance to individual users or specific reviews, the similarity measure could better reflect the varying preferences and perceptions of users. Furthermore, adding the users overall rating with higher weights added, could potentially have a better impact on the similarity of beer style. These approaches would require more time as well as computational resources and data preprocessing but could lead to more robust and accurate results in capturing similarities between beer styles.

## Reflection

1. Data retrieval and manipulation: Understanding how to fetch data from a URL, handle different data formats (such as JSON ,JSONLINES".gz"), and manipulate large datasets efficiently using Dask.
2. Data processing and analysis techniques: Using various methods learned in the quiz like Using cosine similarity to measure similarity between vectors  cosine similarity analysis, and Apriori algorithm to find frequent words.
3. From a sample data project removing stop words using NLTK.

From this project, I learned the importance of combining different data processing and analysis techniques to derive meaningful insights from large datasets. By exploring concepts like mean vector calculation and cosine similarity analysis, I gained a deeper understanding of how to quantify similarities and patterns within complex datasets. Additionally, working with text data taught me valuable skills in text preprocessing and analysis, enabling me to extract useful information from textual reviews effectively. Overall, this project enhanced my knowledge and skills in data analysis, text processing, and statistical techniques, providing a practical application of concepts learned in coursework.

## References

Dask documentation
https://docs.dask.org/en/latest/why.html

TextBlob, Parthvi Shah

https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524