Stefan Sebastian, 242

Abstract

Contents

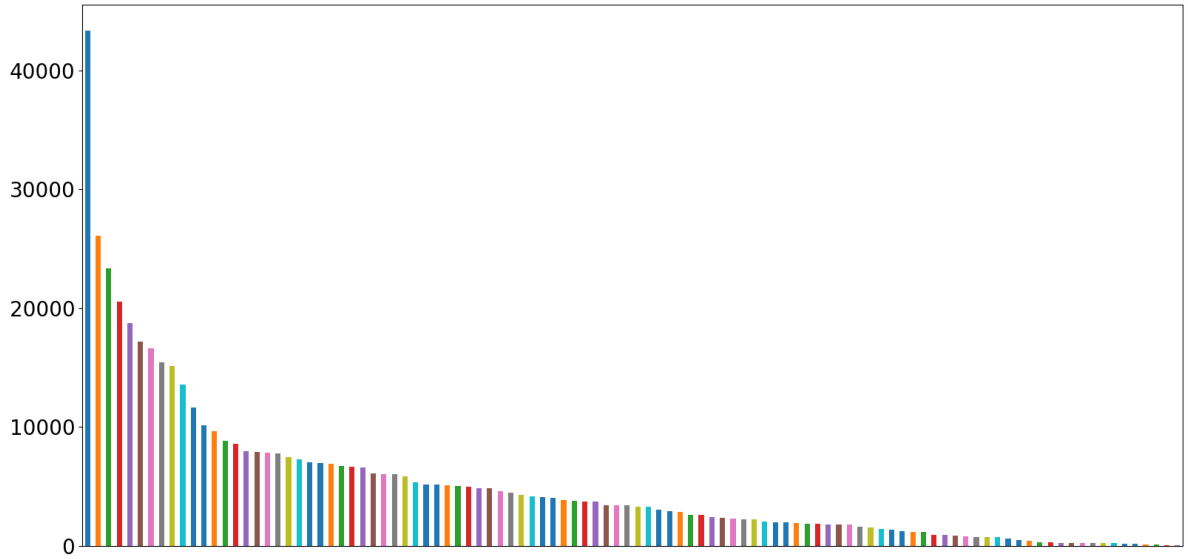

Figure 1: A plot of the number of reviews per beer style

1 Dataset

The dataset used for this experiment is a collection of beer reviews taken from the BeerAdvocate website. The original dataset was made up out of 1.5 million user reviews, from 33387 different users, collected between 1998 and 2011. It is not available at this time but I found a subset of around 500 thousand reviews on data.world[1].

The data is in csv format, each row containing various information like: beer name, beer style, alcohol content, scores for taste, appearance, aroma and a textual review. The only columns considered for this experiment were the beer style and the text review.

On exploratory data analysis, 407 duplicate and 119 missing reviews were found. Also there are 104 distinct beer styles and the amount of data for each of them is quite imbalanced as can be see in figure 1, where each bar represents a beer style with a size proportional to the number of reviews for it.

The first step in preprocessing this dataset was to remove rows containing duplicates and missing values in the columns that are relevant for the experiment: beer style, text review. Next, some of the styles for which there wasn't a lot of data available were dropped, as they would be outliers for clustering. The threshold for this cut was chosen arbitrarily at 7000 reviews.

Upon inspection of the beer styles present in the dataset, a fine granularity was observed. For example, the difference between American Pale Ale and American India Pale Ale is minimal and not very rigorous. This and other similar cases would only confuse the classifier. For this reason, a mapping was made for each style to a broader category, using the taxonomy published by BusinessInsider[2].

The final mapping can be seen in figure 2. Finally, a balanced dataset is built by selecting the minimum value for which we have an equal distribution of reviews per beer style, which is around 9500 per style. The analyzed dataset is made through random sampling of 9500 values for each of the remaining beer styles.

Figure 2: Mapping of beer styles from the original dataset

References

[1] "Beer advocate reviews." `https://data.world/petergensler/beer-advocate-reviews`, 2011.

[2] M. Stanger and S. Gould, "Everything you need to know about beer, in one chart." `https://www.businessinsider.com/different-types-of-beer-2014-12`, 2014.