Aron Müller Project: Data Analysis 23.12.2022

Task 1: Use NLP techniques to analyze a collection of texts

The dataset I plan to use is the amazon musical instrument reviews (https://www.kaggle.com/datasets/eswarchandt/amazon-music-reviews) I found on kaggle.com. I know the topic of musical instruments and for further understanding, it is good to know some specific words like brands or technical terms in the context of musical instruments. The data will be saved locally on my computer in a .csv format. I will be using Python Jupyter Notebook for the data analysis. To get the data into the Jupyter Notebook environment I will be using the pandas library. To get a first overview of the dataset I will use the pandas function df.describe(). For the first raw data cleaning, I will also use the pandas library, after that I can use the NLTK library for removing punctuation, emojis, stopwords, lemmatization, stemming, and Tokenization of the review.

To get the data into a numerical form I need to tokenize the words. There are multiple approaches to tokenizing unstructured texts. One basic approach is o use the Pythons inbuild method text.split(). With this approach, every word in a sentence will be separated. Another approach is to use the NLTK word_tokenize. After tokenizing we need to bring the token in a numerical form, for this, there are multiple ways to do so. One simple method is o show the word count, this means each word occurrence is counted. To extract the most prevalent topics of a text one can choose multiple ways to do it. One method is a more graphical, intuitive method. One can create a Wordcloud to show the most used words in a text. To simply create a wordcloud the python library wordcloud is used. A more logical approach is o sort the most counted words and analyze this list. For further analysis, one can test the correlation between wordcounts by using NumPy and matplotlib or seaborn library. There are multiple ways to show statistical dependencies between the words.