Choose a job you love, and you will never have to work a day in your life.

(Confucius)

Lab 1
Unsupervised Approaches

In this lab, we apply unsupervised approaches to a dataset of song lyrics. The dataset consists of song title, song lyrics, the respective artist, and musical genre per song.

The aim of the Lab is to implement an approach that structures (e.g. clusters) this data based on the song lyrics alone. That is, we'll use the genre and artist information only to validate the approach.

Tasks:

- Cluster the song lyrics into the numer of genres or artists. Do the clusters correspond to the genres/artists? (kMeans)
- Generate a dendrogram of the songs. Do songs from the same genre/artists appear in the same branches? Do higher level nodes group similar artists/genres as their children? (hierarchical clustering)
- What are the most distinct terms within a genre or per artist? (TF IDF)
- Create a 2D plot that shows the similarity of the artists/genres. (PCA)

Technology:

Python – SciPy, sklearn, numpy, pandas, matplotlib