# CS464 Machine Learning
# Project Proposal

## "Song Lyrics Decade and Popularity Classification"

**Group Members**

Elif Kurtay 21803373 Cansu Moran 21803665 Atakan Dönmez 21803481
Öykü Irmak Hatipoğlu 21802791 Elif Gamze Güliter 21802870

### 1. Description of the data

For our project, we will extract song lyrics from AZLyrics.com website using AzAPI[1]. In addition, we will use the Spotify-Data 1921-2020 dataset available on Kaggle[2]. This dataset includes 169k songs from the year 1921 to 2020 with labels such as song name, artist name, danceability[0-1], energy[0-1], popularity[1-100], tempo[50-150], liveness[0-1] and year[1921-2020]. By using AzAPI and Kaggle dataset, we aim to create our own dataset where each song lyric has their corresponding release decades and popularity measures.

### 2. What we aim to answer with this data

With this data, we aim to perform two different classification tasks. First, we aim to classify song lyrics into decades to observe what kind of distinction the song lyrics have in different decades, for example 70s, 80s or 90s. After training our model to classify the songs' release decades based on their lyrics, we can predict the decade in which a song may have been released based on the lyrics. The second classification task we aim to perform after the decade prediction is based on the popularity of the songs in each decade. On the dataset, the popularity metric has a value in range 1 to 100. Since having a class for each popularity value will not be sensible, we will perform regression to predict popularity as a continuous value. We will train a different regression model for each decade and we will try to observe the predictions on how popular a song would be if it was released in each decade. Hence, we will be able to compare a song's predicted popularity in its predicted release decade versus other decades as a result of both classification tasks.

### 3. What we plan to achieve by the milestone

By Progress Report, we want to create our dataset and process it according to the features we selected while getting rid of any unnecessary data samples such as any non-English songs or samples that have missing features. We aim to experiment on different text vectorization techniques such as Word to Vector(Word2Vec), Global Vector(GloVe), Term Frequency-Inverse Document Frequency(TF-IDF), and Bag of Words. In addition, we plan on starting implementing our classification models. Currently we plan to use Recurrent Neural Networks (RNN) with Long Short-Term Memory (LSTM) as a Deep Learning method and Naive Bayes classifier as a Machine Learning method for both classification tasks.

By the Final Report, we aim to optimize our models and compare the optimized results of different techniques to observe which learning-based technique is best fit for our problem. Furthermore, we want to explore the results of our models by predicting how popular a song would be in each decade and which year that song should have been released for high popularity.

[1] elmoiv, "Azapi," *PyPI*, 12-Feb-2021. [Online]. Available: https://pypi.org/project/azapi/. [Accessed: 30-Oct-2021].

[2] E. Negi, "Spotify-data 1921-2020," *Kaggle*, 29-Aug-2020. [Online]. Available: https://www.kaggle.com/ektanegi/spotifydata-19212020. [Accessed: 30-Oct-2021].