Lyrics Analysis - Sureel.ai Task Assignment

Task Introduction

The task consists of developing a method to identify similarities between newly created lyrics (e.g., AI-generated lyrics) and the writing styles of popular artists.

Method

I have developed a baseline proof of concept based on *artist embeddings*. The core idea is to compute a distinctive embedding for each artist that encodes the patterns, structures, and topics frequently used by the author, allowing for comparison with a similarly computed embedding for new lyrics. The metric used is cosine similarity.

To achieve this, I followed these steps:

- Downloaded the dataset from https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information, which includes millions of lyrics uploaded to the Genius platform. The dataset is tagged with artist, year, genre, language, and other metadata.
- Preprocessed the data. For this proof of concept, I selected English lyrics and developed *artist embeddings* for 12 prolific, popular artists known for writing their own songs, ensuring consistency in style.
- Computed *artist embeddings* as the average of their lyrics embeddings. The model used is available in the SentenceTransformers library and is **all-MiniLM-L6-v2**, a lightweight model typically used for semantic search embeddings.
- Generated a validation dataset using the OpenAI API (model **chatgpt-4o-latest**). This dataset consists of synthetic lyrics mimicking the style of each of the selected artists. The lyrics were generated with a prompt that included two randomly sampled examples of the artist's lyrics. The evaluation dataset contains 60 AI-generated lyrics (5 per artist).
- Developed a final script that loads these embeddings and computes cosine similarity against provided lyrics to output similarity metrics.

Results

To evaluate the baseline I have analyzed two metrics: top-1 accuracy and top-3 accuracy. Achieving a top-1 accuracy of 42.37% and a top-3 accuracy of 79.66%. These results provide an intuition that the embeddings are significant for the task to some extent. When taking a look closer to the results, we can see that some artists are consistently failing. This happens, for example, in the case of The Beatles. I have three reasons for this behavior: first, their lyrics use a simple English and talk about typical pop topics that are also present in other artists like Johnny Cash or Ed Sheeran. Second, quality of AI generated lyrics is not perfect, and hence is not able to reproduce exactly the style. To study in depth this phenomenon, I have validated the method against real lyrics from each artist that were left out when computing the embeddings. In this experiment both accuracies increase between a 3% and 5%. Lastly, the embeddings used are meant for semantic search, and they are produced by a lightweight, less powerful model. In this task, semantics is not the only factor. Style, syntax and rhythm play a key role and some of these characteristics might be overlooked by the model.

Possible Improvements

- Fine-tune embedding model with contrastive learning: it could be interesting to train with a Contrastive Learning Loss in order to separate artist styles in the embedding space.
- Select representative sentences: instead of creating an artist embedding with all their lyrics. Some threshold metric (e.g perplexity) can be used to filter characteristic sentences.

- Other models: ongoing research projects (e.g https://arxiv.org/pdf/2406.15231) proposes LLM2Vec models to perform this task. It seems like GPT based architectures are better to encode style features.
- Combine features: there are validated classical text features that can be combined with DL embeddings to achieve better results. For example the Universal Authorship Representation (UAR).