Music Artist Embeddings

Anonymous ACL submission

Abstract

This paper will describe the process of collecting a song lyric corpus by webscraping lyric and music genre data, and using it to extract BERT embeddings in a way that attempts to capture semantically relevant information about a music artist so that artist/genre similarity tasks may be performed using classic clustering techniques.

1 Intro

The question that this project addresses is that given a corpus containing lyric data for different music artists, is it possible to capture meaningful information about each artist by extracting BERT embeddings for an artist's lyrics, and creating an "artist embedding". This question is important because it would allow many different music recommendation tasks to be done using simple unsupervised learning algorithms as opposed to complex and hard to train neural networks.

2 Data

The data set that was used for this project was one that was scraped from the internet specifically for this project. It includes lyrics for 2000 different music artists' top 10 songs, as well as each artist's main music genre. The genres and music artists were scraped from https://en.wikipedia.org/wiki/Lists_of_musicians. The number artists scraped from Wikipedia is much higher than the 2000 included in this data set. This is because the amount of lyric data that was gathered was limited by time constraints. Additionally, only 8 different genres of artists were able to be gathered because of time constraints, (Punk, Emo, Dance Pop, Hip-hop, Hard Rock, Dream Pop, Acid Rock, Country, and Adult Alternative). The lyric data for each of the 2000 artists was gathered by scraping https://genius.com/. This data was stored in a csv file with columns for the artist name, genre, number of songs scraped, and lyrics.

3 Methods

This section will discuss the methods used to gather the song lyric corpus, methods used to create artist embeddings, and the methods used to evaluate the embeddings.

3.1 Data Gathering

A much larger portion of this project than anticipated was dedicated to finding a usable data set. There exist large data sets containing artist data, song lyrics, and other metadata. However, these data sets are either in the bag-of-words format, or they are restricted due to copyright issues. This being the case, for this project to be possible, it was necessary to gather a custom data set. As mentioned in the previous section, the artist and respective genre data was scraped from Wikipedia, and the lyric data was scraped from Genius.

Because multiple websites were used, the work needed to gather and clean this data was comprehensive. First over 100 lists of artists were scraped from Wikipedia, as well as the genres each artist belongs to. Because of the sheer number of artists gathered, it was impossible to gather lyric data on all of them. This resulted in roughly 2000 artists being selected to gather lyric data on.

Next, for each of the 200 artists, hyperlinks to lyric websites for their top 10 songs were gathered using the Genius API. Finally, lyric data was scraped from all 10 of the websites containing lyrics for the top 10 songs. The Genius API was the main factor that contributed the time constraints that were mentioned earlier, as it severely limited the number of requests that were able to be made for each artist. This resulted in lyric data only being able to be scraped at a rate of about 10 songs every 1-2 minutes.

3.2 Artist Embeddings

The artist embeddings used in this project were made using the BERT transformer model, (BERT-base-uncased). For each artist, an embedding was created by taking the lyric data, tokenizing it, running it through the BERT model, and extracting the hidden state of the BERT model. Because each artist has lyric data for their top 10 songs, the lyric data had to be split into several lists of tokens, as the BERT model can only take inputs with a maximum size of 512 tokens. This usually resulted in each artists having 4 different BERT embeddings.

These 4 embeddings were created by taking the average of the second to last hidden layer for each token for each of these token sequences. The second to last layer of the hidden state was used as in the original BERT paper it was found that this approach worked best for feature extraction tasks. These 4 embeddings were then averaged to create a single artist embedding. Two trials were than ran using these embeddings, one using 500 artists embeddings and one using all 2000 artist embeddings.

3.3 Clustering and Artist Similarity

To get and understanding of the semantic meaning captured by the extracted artist embeddings, a K-means clustering model was used to split artists embeddings into meaningful groups. Two K-means models were created, one using 500 artist embeddings, and one using the full 2000. To understand the clustering, histograms of genre counts for each cluster were created.

Additionally, the lists of artists in each cluster were investigated. Because the dataset used does not contain any metadata about the artists, these are the only ways to investigate the cluster characteristics. Additionally, artist similiarity was investigated for certain artists based on the cosine similarity between artist embeddings.

4 Results

For the first K-means model, (500 artist embeddings), 4 clusters were created with the following genre majorities, Cluster 1: Hip-hop - 66%, Cluster 2: Hip-hop - 40%, Cluster 3: Hip-hop - 83%, Cluster 4: Adult Alternative - 35%.

For the second K-means model, (2000 artist embeddings), 9 clusters were created with the following genre majorities, Cluster 1: Adult Alternative - 30%, Cluster 2: Hip-hop 24%, Cluster 3: Adult Alternative - 38%, Cluster 4: Hip-hop - 84%, Cluster 5: Hip-hop - 65%, Cluster 6: Adult Alternative - 40%, Cluster 7: Hip-hop - 85%, Cluster 8: Hip-hop - 66%, Cluster 9: Hip-hop - 63%. The complete genre percentages for each cluster model are available in the clustering.ipynb and large_clustering.ipynb notebooks.

Additionally the 5 nearest artists were calculated using cosine similarity for the following artists in descending order: Shakira - (Beck, Mexicano 777, Neo, The Black Eyed Peas, Banghra), Bob Dylan - (Luke Doucet, Blue Öyster Cult, Key, Brett Dennen, Grateful Dead), Katy Perry - (Jamiroquai, Emily Osment, Janelle Monáe, Fall Out Boy, Lorde), Lil Wayne - (Dreezy, Isaiah Rashad, Nicki Minaj, Quando Rondo, Santana).

5 Discussion

The full story of the artist clusterings is not told by the somewhat unremarkable genre majorities. Judging by those numbers it seems that the embeddings did not capture enough meaningful information about the artists to create meaningful clusters, otherwise the genre distributions of the clusterings would be more homogeneous. This however is untrue, and a qualitative analysis of the clusters reveals semantically meaningful trends in each of them.

One thing that an be noticed in both the 500 and 2000 artists clusters is that Hip-hop tends to be the majority genre. Based on those majority percentages, it would seem that the hip-hop clusters are not well defined, or that there are multiple different hip-hop sub genres that are, lyrically, radically different enough to constitute different clusters for each. The second explanation is more true than the first, because in the 2000 artist model, Clusters 4, 5, 7, 8, and 9 are comprised of American, Asian, French, Latin, and European/African Hip-hop artists respectively. This is a very interesting result because the BERT model used to create these artists embeddings was only trained with an English corpus. Yet the extracted embeddings capture information about the different languages artists speak. Furthermore, these clusters are very homogeneous in their language classifications. I have no statistics for this, but a qualitative analysis of the artist lists for each cluster gives this impression.

Why though are there only language clusters for the Hip-hop genre? This is most likely caused by the methods used to gather the data set used to create the artist embeddings. As mentioned in section

2, the list of artists was scraped from Wikipedia. The artists lists on Wikipedia are not all of the same format, and most likely, the list of Hip-hop artists included artists who speak languages besides English, while the other lists were restricted to English speaking artists.

In addition to the langauge clusters, it also appears that the other clusters have meaning trends. From the 2000 artist model, Cluster 2 appears to contain more underground/indie artists compared to other clusters in the model. Cluster 3 appears to contain artists that align better with classic American radio programming. And Cluster 6 appears to contain artists whose lyrical content is focused more on emotions. To be clear, these observations are purely subjective and based on my own music domain knowledge, as I have no metadata to quantify these findings. However, I have confidence that given proper metadata for each of these artists, a statistical analysis would reinforce these obeservations.

Again, this lack of metadata makes it hard to quantify the semantics captured by the artist embeddings, but looking at the artist similarity rankings in section 4 show that these embeddings do capture some meaning. As the similarity rankings are consistent with each artists' style, (minus the Beck and Santana). For example, the artists most similar to Katy Perry, (Jamiroquai, Emily Osment, Janelle Monáe, Fall Out Boy, Lorde), are all artists in the Pop genre.

6 Implications and Next Steps

The results of this project are very promising, despite the difficult to quantify results. This project has shown that meaningful information about a music artist can be extracted into a dense embeddings, given only a small corpus of that artist's lyrics. With additional data, model fine-tuning, and more advanced unsupervised machine learning models, the potential of this approach for genre classification and music recommendation systems could be very high.

Additionally, this project has unintentionally shown that this embedding approach can be used for language classification. The fact that the embedding model was able to capture information on the different languages that artists use, without even being trained to perform such a task is quite extraordinary to me, and shows that this approach could have potential for many different classification tasks.

The immediate next steps for this project would be to gather a proper music artist data set that contains more metadata on each artist in addition to genre and lyrics. Having more quality data alone could significantly improve the artist embeddings, and make clustering algorithms an even more viable method of classification. Additionally, expanding this project to attempt different kinds of classification tasks, (language classification, song sentiment classification, etc.) would give insight on where this approach could be used elsewhere.

7 References

McCormick, Chris. BERT Word Embeddings Tutorial, mccormickml.com/2019/05/14/BERT-word-embeddings-tutorial/.

"Lists of Musicians." Wikipedia, 29 May 2023, en.m.wikipedia.org/wiki/Lists_of_musicians.

Devlin, Jacob, et al. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv.Org, 24 May 2019, arxiv.org/abs/1810.04805.