Genre prediction from lyrics.com

Problem

According to Hypebot.com, in every 24 hours in 2018 roughly 20,000 musical tracks were uploaded to major streaming services. That amounts to approximately a million tracks every six weeks. This enormous amount of data is too much for any single human to sift through in a lifetime. Assuming three minutes a track, a million tracks would take a single individual almost six years to listen to.

In order for music to be made available to individuals it is necessary to perform some degree of automated classification. The task of genre classification (assigning similar music to groups) is one of the most fundamental of these tasks. Broad grouping of music along stylistic similarity allows increased ease of discovery of new artists for listeners, and for marketers it allows identification of potential consumers for music from new bands.

Data

Lyrics.com is a website featuring lyric data from roughly 150,000 artists across 1.3 million songs. Of these 1.3 million songs, roughly 15% have genre tag classifiers. Additional tagging is available for subgenre or audience-specific tags.

In order to acquire the lyric information, a web scraping application was created using BeautifulSoup and the Requests library for python. The 1.3 million songs represented on Lyrics.com were downloaded and stored in a SQLite database object for further analysis using SQLAlchemy.

Exploratory Data Analysis

Initial investigation of the data set showed that of 1.25 million songs scraped from the website, 382,600 had associated lyrics. The vast majority of songs represented appeared without lyric content, either because they represented instrumental tracks or because user catalogued lyrics had not been supplied.

Rock	56226
Pop	46089
Jazz	18897
Electronic	17622
Folk, World, & Country	17236
Hip Hop	14935
Funk / Soul	12355
Latin	7190
Blues	5270
Stage & Screen	3512
Reggae	2853
Classical	889
genre	850
Non-Music	367
Children's	349
Brass & Military	26
genres	8
Ce genre de Mec	5
Ce genre	2
Rapgenre	2
Buy This Song	1
Ce genre Lyrics	1
transgenres	1
Le genre féminin	1

Name: genre, dtype: int64

Figure: Genre representation in the lyrics.com dataset

Genre Curation

When considering genre taggings of songs in the database, 134,673 unique combinations of song title and artist were represented and genre tagged. Roughly 38,000 songs had multiple associated genre tags. Multiply tagged songs had as few as two or as many as eighty associated genres. Of tagged genres, Rock and Pop were by far the most highly represented, with Rock representing 27% of all singly-tagged songs and Pop representing 22%.

Focusing on a subset of the seven most highly represented genre tags further reduced the dataset to 94,000 tracks. By analyzing the distribution of these tracks it was possible to determine the degree of overlap between tracks. The resulting analysis showed the Pop tag to be highly promiscuous, associating with all other tags except Hip Hop. Rock was also highly promiscuous, associating with all tags except for Jazz. Of all the combinations the Rock/Pop tag combination was the most highly convolved.

In order to address this issue, the group of tracks was further decreased to the Rock, Jazz, Hip Hop, and Folk genres. This leaves us with approximately 85,000 tracks with a single genre track, roughly half of which are tagged as Rock tracks.

Language Analysis

We can also assess the language of each track, limiting to English-language tracks. The langdetect package provides tools for matching a large number of languages, English among them. By assessing the first fifty characters of each

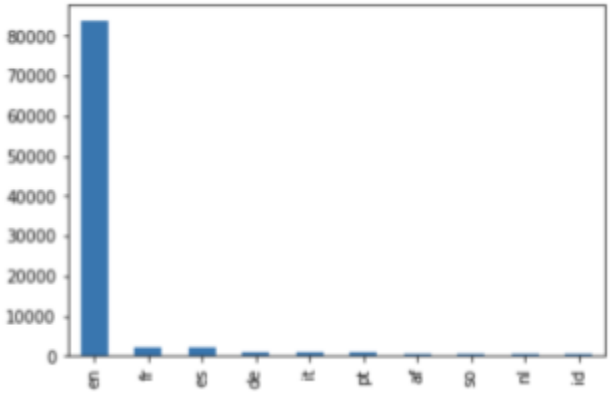

Figure: The genres represented in the dataset skew heavily toward English. Some 10,000 tracks are in other languages.

song we can arrive at a language classification for all tracks in roughly ten minutes of compute time.

This reduces the training set to 74,000 tracks, eliminating 10,000 tracks, most of which are Spanish, French, Dutch, and Italian.

Song Deduplication

In order to check for song lyrics that are represented multiple times in the data, a custom approach was used to iteratively compare sub-blocks of the data. An exhaustive approach using bit vectors to compare all pairwise similarity scores among the 85,000 tracks was found to take nearly six hours to run on one core, but by comparing blocks of 20,000 or fewer records using cosine similarity, this time can be shortened to minutes at the cost of the possibility of a small number of duplicate records remaining in the dataset.

Data is first sorted by track name, then blocks of 20,000 records are compared for pairwise similarity. This process is then repeated multiple times until the number of duplicates converges. Validation with exhaustive duplicate detection shows this method to remove 99% of duplicates in the lyrics.com catalogue. In cases where duplicates were detected, the earliest dated entry was retained and all other entries were discarded.

The resulting set of tracks comprises 41,422 unique tracks, indicating that in fact more than half of the data is duplicated, representing four genres. In previous runs, Jazz was found to be the most highly duplicated of all the genres, with nearly three quarters of tracks comprising covers or standards.

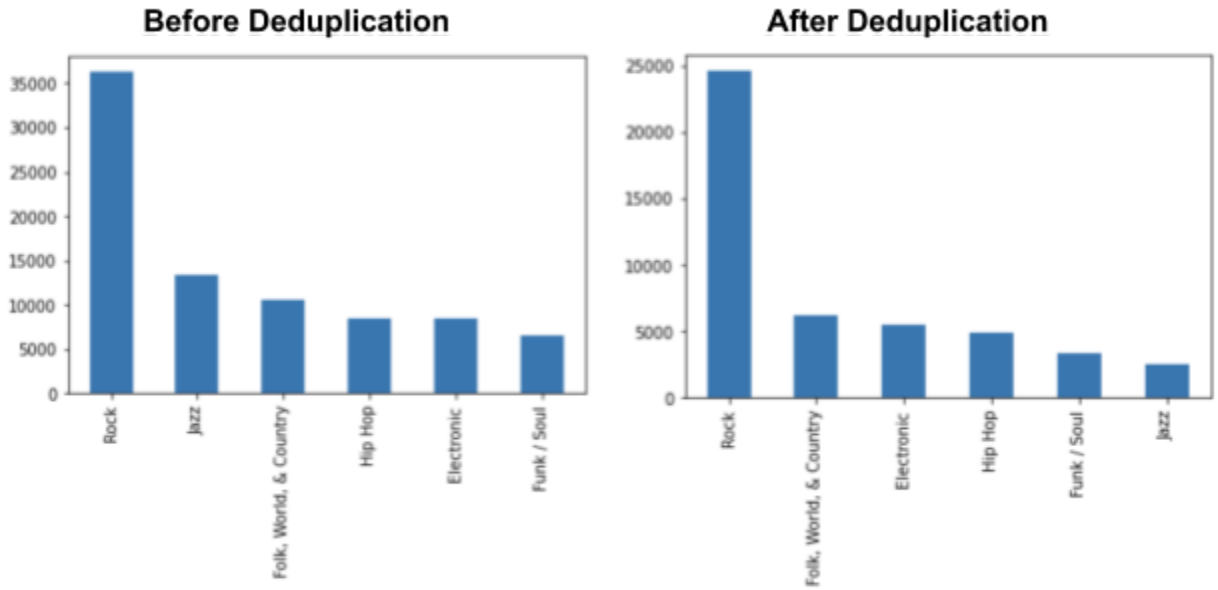

Figure: Genre counts before and after deduplication. Jazz decreases the greatest amount, while Electronic has the lowest ratio of duplicate tracks.

Investigating Cleaned Data

We can interrogate the data now looking for interesting trends between the four genres of interest.

When considering the length of lyrics in each song, Rap has by far the longest average song length at 2800 characters, while Jazz is the shortest at roughly 750.

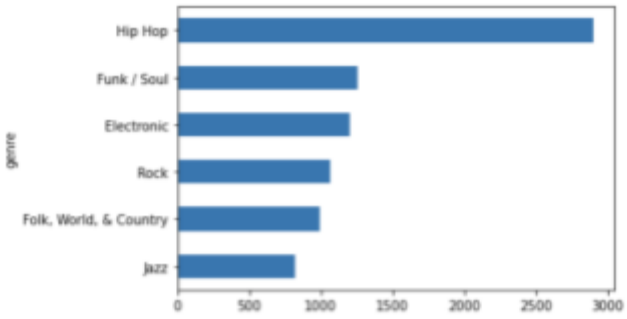

Figure: Hip Hop has the longest lyric length by character count, while Jazz has the shortest. The majority of songs range between 1000-1500 characters.

The number of unique words in each genre roughly approximates this distribution. In order to compensate for the Rock category's much larger data set, ten bootstrap samples of 2000 songs were taken from each genre and the number of unique words were computed. Rap again predominates with an average of roughly 70,000 unique words per 2000 song subset, while Rap and Electronic come in at an average of 22,000 unique words per song analyzed, and Folk comes in last with 18,000.

We can also look within each genre to see which artist has the largest number of unique songs associated with their name in the Lyrics.com dataset. Elvis leads the pack as the biggest contributor to the Rock genre with 149 distinct songs; Johnny Cash leads Folk with 103; Depeche Mode leads Electronic with 62; and Jay-Z is the top Rap contributor with 52 songs.

Word Clouds

Finally, to provide an overall visualization of the differences between the various genres, we can use the wordcloud python library to create word cloud representations of each genre. From these we can see some intuitive words that are associated with each, including Rap's favorite pronoun.

Figure: A subset of genre word clouds. "Love" and "know" feature prominantly, as do words like "got" and "want". Hip Hop features more slang terms and vernacular phrasing.

Conclusions

The Initial investigation using observational visualizations and EDA demonstrates that the dataset has limitations associated with genre tagging and duplicated tracks. Despite this, we can infer a number of interesting things about the data.

The discrepancy in Rap's character length relative to other songs is not unexpected, but is striking. Jazz's low complexity may indicate this class will cause difficulties when attempting to classify. Regardless, this discrepancy is a useful metafeature to consider.

genre	artist	
Electronic	Depeche Mode	67
	Madonna	55
	Pet Shop Boys	51
Folk, World, & Country	George Jones	103
	Johnny Cash	92
	Marty Robbins	71
Funk / Soul	James Brown	54
	Aretha Franklin	51
	Earth, Wind & Fire	47
Hip Hop	Jay-Z	53
	Nas	50
	Eminem	40
Jazz	Frank Sinatra	111
	Nat King Cole	66
	Ella Fitzgerald3	62
Rock	Elvis Presley	145
	The Rolling Stones	104
	Elton John	81

Figure: Top contributing artists by genre. By observing the three top artists in number of unique songs within each genre label, we can form intuition about what constitutes a song in each classification group.

We can also observe differences in the word clouds generated from each artist. Core vocabulary differences can be observed, but the similarities are again the most striking thing about these visualizations. Hip Hop again seems to assert its individuality from the older music styles.

Next steps are to create a classifier to assess the ability to assign labels to tracks that have not been observed. The ultimate goal of this investigation is to arrive at a system for classifying songs into their respective categories in an automated manner. The real question is whether these labels are based in something that can be derived from lyrical context, or simply a product of instrumental differences.