Ridley Arp

NLP/Unsupervised Learning Project Write Up

Metis Bootcamp March 22

**Abstract**

This project was created to use unsupervised learning models to extract topics and commonly used terminology/trends from song lyrics based on Genre/Time Frame. The resulting vocabulary could be used by soundtrack composers and lyricists to make more authentic lyrics to fit more naturally into the world of the story(either thematically or etymologically). This information was extracted via a web-scraping pipeline, with each genre independently and collectively analyzed for topics, with the strongest representatives of each topic/genre returned as a vocabulary separated by genre.

**Design**

The project is designed as a tool to be utilized by lyricists, composers, and musicologists to explore the trends and topics of music lyrics as differentiated by decade and genre. The resultant vocabulary would consist of the most used words from the most common topics from each genre and timeframe. This vocabulary could be used to explore and analyze the musical lexicon of different time periods, as well as aid in creating more authentic pieces to fit within a specific time period.

**Data**

The full dataset consisted of 32 genres, with the final corpus consisting of 1200 songs across 12 categories, with the top songs from each represented group. Time-based segments include each decade from 1950’s-2010’s, along with the most popular songs of the most popular genres including Rock, Country, Pop, etc.

**Algorithms**

*Feature Engineering*

* Elimination of problematic songs by combination and elimination of genres
* Isolated additional stop words through EDA to aid in preprocessing
* Custom tokenizer created to implement spaCy

*Models*

* CountVectorizer and TfidfVectorizer used as a basis for modeling with Tfidf offering the most promising results
* LSA and NMF tested with CV and TFIDF on multiple hyperparameters, with NMF chosen for easier readability.
* spaCy token models included to allow for further POS and syntax analysis

*Evaluation and Selection*

* created a pipeline to train and test hyperparameters for models on every genre in the corpus individually, as well as an overarching topic model, outputting a vocabulary for each topic for each segment.

**Tools**

* Numpy and Pandas for data manipulation and transformation
* Sklearn for modeling and tokenization
* Beautiful Soup for data acquisition
* spaCy and Regex for text preprocessing and analysis

**Communication**

* Presentation with variable pipeline created for further exploration