https://www.overleaf.com/project/5c55dade988738126042a646

Abstract

This paper studies the relationship between popular songs' lyrics and the economic condition of their respective era. The goal is to build a classification model where a song's lyrics is passed into the model, and the model can predict if the song is popular during the times of expansion, peak, recession, or recovery. This study might attract scholars from multiple disciplines such as sociology, economics, and humanities. With the classification model, we can extract the "mood" of a nation in a certain period of time. In addition, we might also predict the economic growth from the national "mood" reflected in music. We can also use the classifier to judge if a new song is "retro" or not.

1 Introduction

Each era has its distinctive genre of music. For example, the rock music in the 70s, the electronic music in the 80s, and the hip-hop music in the 90s. While enjoying a diverse set of music from many eras and different countries, I came up with the theory that popular song lyrics can reflect a country's economic health in that certain period. For example, most of the "City Pop" genre reflects the blooming economy and celebrates consumers of the late 80s in Japan. Thus I wish to use the project to study that if popular music can be used as a country's economy barometer. Specifically, I will use lyrics of Billboard Top 100 songs since 1950s as the corpus, divide the corpus into four categories based on the economic cycles as defined by National Bureau of economic research, extract textual features from the lyrics within each category, and build a classification model where it can categorize a song into the four pre-defined categories. With this classification model, we can classify the era or economic cycle when the music is written. Because the economic cycle data set is

generated in hindsight, if we use this model on the latest popular music, we might be able to detect changes in economic cycle for a certain country.

For economists, if we continuously running this model on the latest popular music, we can catch the moment that the "national mood" has shifted. Ideally, we can use this model to analyze popular music and use it as an indicator for a country's economic health.

For humanities scholars, this classification algorithm can identify "outliers" in each era. For example, during the time of economic expansion when everyone else is singing about prosperity, there might be an artist whose music is rather melancholic. Thus humanities scholars can find these "outlier artists" and study their work closely.

For further study, we can incorporate other features of music such as tempo and beats into the study. My hypothesis is that up-beat and bright tones are correlated with prosperity. However, at this time, I do not have a reliable source to scrape song files. In addition, I have to think about the best way to represent tempo and beats numerically, such the data can be used to train models.

2 Problem Definition

The goal of this project is to create a classification model, where the input is a song's lyrics, and the output is one of the classes Expansion,Peak,Recession,Recovery, which represent the economic cycle of the year that the song made to Billboard Top 100 chart.

3 Data

With all the data collected, I should have lyrics for Billboard Top 100 songs in every year since 1958, and I should also have U.S. economic cycle information for every year since 1958. Joining the two tables together, I will have (each song's lyrics-

economic cycle at the time when the song is on Billboard Top 100) data set. The economic cycle will be the "class" for each lyric, and the model is supposed to predict the economic cycle given a song's lyrics.

Song	Artist	Lyrics	Year
99 Revolutions	Green Day	There's a trouble	2012
Uptown Funk	Mark Ronson	This hit, that ice	2014
Baby	Justin Bieber	Ooh whoa, ooh	2010

Figure 1: Lyrics data set

Year	Economic Cycle
1990	Expansion
1991	Expansion
1992	Peak
1993	Peak
1994	Recession

Figure 2: Economic cycle data set

4 Related Work

1. https://ieeexplore.ieee.org/abstract/document/1394328

2. https://dl.acm.org/citation.cfm?id=1101255

3. https://ieeexplore.ieee.org/abstract/document/4725050

4. https://ieeexplore.ieee.org/abstract/document/5363083

5 Methodology

For creating representation of each song, I plan to use the following methods to create word-level representation first:

1. Word2Vec

2. Character-level CNN

3. Forward LSTM

4. Backward LSTM

5. 4-layer ELMo which takes the above mentioned methods as layers

After creating the representation for each song, I plan to use the following frame work for building the classification model:

1. Support Vector Machine

2. Naive Bayes

3. Logistic Regression

4. Single-layer neural network

5. Convolutional Neural Network

In total, there will be 5*5, 25 combinations of models and embedding methods.

6 Evaluation and Results

This problem is a standard supervised classification problem, where the output labels are provided. Thus to evaluate the accuracy of the model, I will be using F2 score to compare the prediction with the ground truth. In addition, I will use a confusion matrix to show the type I and type II errors.

7 Discussion

You can leave this section blank.

8 Work Plan

1. Scraping song titles and artists information for Billboard Top 100 songs since 1950s

   For this task, I plan to use this Wikipedia page:

   https://en.wikipedia.org/wiki/List_of_Billboard_top-ten_singles

   One challenge I might encounter during this process is the inconsistent formating of wikipedia tables. Between all pages for each year, the column name might not be exactly the same. For example, column "artist" might be spelled as "artists" in some other pages. Also, wikiepdia table is known for the practice of stacking column names together as shown in Figure 1. This makes it harder for me to parse the table.

2. Scraping song lyrics based on song titles and artists information

   For this task, I plan to use this website:

   https://www.lyrics.com/

Clock speeds			Fillrate			
Base core clock (MHz)	Boost core clock (MHz)	Memory (MT/s)	Pixel (GP/s)[f]	Texture (GT/s)[g]	Size (GiB)	Ba
1151	1379	2100	18.41	27.6		
1227	1468	6000	19.6	29.4	2	
1354	1455		43.3	54.2		
1392	1518	7000	33.4	66.8	3	

Figure 3: Tables that have stacked column heads

While this website seems to have simplistic design, the challenge for this task is to identify if the resulting song is indeed the song we are looking for. There are plenty of songs with the same name, see Figure 2. Thus it becomes challenging to pick the right lyric that matches the song we are looking for, see Figure 3.

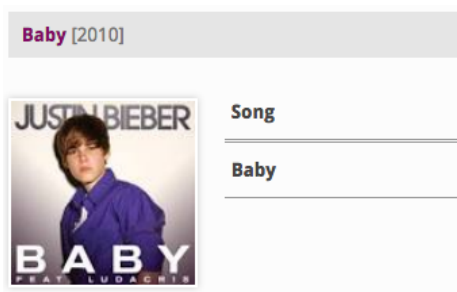Figure 4: So many songs named "Baby"

Figure 5: Yet we are looking for this song by Justin Bieber

One potential solution is to use the artist name and time period as a part of the search query. If the website returns multiple results, then I will use FuzzyWuzzy in Python to find the most likely candidate for the song.

3. Scraping US economic cycle data from National Bureau of economic research

   For this task, I will be using the data from this web page https://www.nber.org/cycles.html

   Besides the formatting of the table in this website, one major challenge for me would be understanding the nuance behind the data and think of a way to visualize the economic cycle.

4. Tokenize lyrics and prepare for building models

   Because I restrict my song selection to Billboard Top 100, it is very likely that the song is written in English, thus making it easier to use the pre-existing package for tokenization.

5. Testing different classification algorithms and compare performance

   One challge in this step is having enough data. Songs are shorter than the typical text corpus such as news paper. Thus I'm worried about not having enough tokens for building the model. Also, song writers choose their vocabulary not only based on the idea they want to express but also based on rhymes and beats. Thus there might be some noise in the vocabulary, see Figure 4.

Figure 6: Some lyrics are highly repetitive

Acknowledgments

If you got help from anyone or had substantive discussions, please acknowledge those people here and describe how they contributed. The work you do for your project should be entirely your own.

References

Note that you must cite all your references

A Supplemental Material

If you want to put longer examples of data and code, put it here in the appendix.