Project Proposal: Big Data Analysis of Music

Team Members

Peizhou Guo	pg1534
Xiaojie Zha	xz1776
Yang Shi	ys2843

Background

Music has already been a part of people's daily life. You can find people who wear earbuds almost at anywhere and anytime. For music sharing platform (like Last.fm and Spotify), one important function is to find the pattern of users' tastes and the connection among different type of music.

So, we are planning to provide an analytical service around music using the dataset provided in NYU class:

The Million Songs Collection is a collection of 28 datasets containing audio features and metadata for a million contemporary popular music tracks.

Yahoo! Music dataset contains several subsets, including Yahoo! Music User Ratings of Musical Artists, Yahoo! Music User Ratings of Songs with Artist, Album, and Genre Meta Information, Yahoo! Music ratings for User Selected and Randomly Selected songs, Yahoo! Movies User Ratings and Descriptive Content Information, Yahoo! Delicious Popular URLs and Tags, etc.

The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing.

Amazon product dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 - July 2014.

Objectives

Using different combination of these datasets, we can analyze music from different aspects:

(1) MSD & Yahoo review & Emotion (Director: Peizhou Guo)

Links (dataset sources):

- MSD: https://labrosa.ee.columbia.edu/millionsong/
- Yahoo review: http://webscope.sandbox.yahoo.com/catalog.php?datatype=r
- Emotion: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

1. MSD provides several sub-dataset, including SecondHandSongs dataset (cover songs), musiXmatch dataset(lyrics), Last.fm dataset (song-level tags and similarity), Taste Profile subset (user data), tagtraum genre annotations (genre labels).

Using these data, we can try to get several patterns of users. For example, for a specific group of people who love electronic and R&B, 60% also like new metal, 30% like Blues rock, and only 10% like classic music. This can be considered as three patterns. When company like Spotify or Last.fm want to recommend some music to their users, these analysis result could be helpful. Base on a user's personal music data, we can consider which pattern does this user fit. The more data collected from users, the more precise the recommends could be.

2. For each music genre, we can make a list for top artists, this result can be used as a part of service music recommendation service.

3. Using Yahoo review dataset, tagtraum genre annotation (genre labels), musiXmatch dataset and Emotion dataset, for each genre of music, we can analyze the basic emotions which expressed by the lyrics. Also, using the same method, we can analyze the emotion felt by users after listening each genre of musics. For example, although the lyric of most Melodic death metal are negative, however most metal fans feels positive after listening these songs.
We can use score to replace each emotion word, calculate the score for every single music track.

4. Using musiXmatch dataset, we can also analyze the most several frequent words in each music genre.

5. Using SecondHandSongs dataset, we can know which kind of music are easy (or popular) to be covered by other people, the list of these songs can be sold to those karaoke companies to help them save storage space, and gain profit in the same time.

6. For each music genre, we can make a list for top artists, this result can be used as a part of service music recommendation service.

(2) Amazon reviews & Yahoo Music keyword analysis & Google Map (Director: Xiaojie Zha)
Links (dataset sources):
- Yahoo review: http://webscope.sandbox.yahoo.com/catalog.php?datatype=r
- Amazon(metadata/reviews/genre): http://jmcauley.ucsd.edu/data/amazon/links.html

1. Using R and certain APIs, the music genres' and music lovers' location information can be easily presented on the map. We can then analysis the popularity of a genre in different states in the U.S, which state is more suitable for a musician to introduce his/her new albums.

2. With the same resources and the keywords in songs, we can see which city is more likely to be appeared in the songs, and show trends of a certain genre.

3. By grouping the customer music reviews and other purchases records by customer id, we are able to make common conclusions on the purchase habits of different types of music lovers. For example, people who like classic music are more likely to purchase pianos, etc.

(3) MSD & Google Trends (Director: Yang Shi)
Links (dataset sources):
- MSD: https://labrosa.ee.columbia.edu/millionsong/
- Google trends API: https://github.com/GeneralMills/pytrends

1. Find out which genres of songs can make the songs more popular and has higher business value and also raise the popularity of the singers.

2. Mining the trends of the popularity of different genres of songs refer to years. And try to notice if there is anything we can predict. For example, a certain genre of songs become popular once again after ten years since it set off a popular trend.

3. By grouping MSD and google trends dataset, we can get the result of when it is the best time in a year to release new album so that the profit is maximized.

4. Find out the effect of a successful song. If there would be an increase of similar songs issued and how are the profits of such songs, after a successful song is issued.

Tools

Hadoop Pig / Apache Spark / R
JAVA / Python / SQL
NYU HPC (Hadoop Dumbo Spark)

Outcome

1. Write a business report to introduce the music analytical service.
2. Build a website to present results (in charts and diagrams) and draw conclusions.

Resources

[1] Million Song Dataset: https://labrosa.ee.columbia.edu/millionsong/
[2] Google trends: https://github.com/GeneralMills/pytrends
[3] Emotion: http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[4] Amazon (metadata asin/reviews/genre): http://jmcauley.ucsd.edu/data/amazon/links.html
[5] Yahoo http://webscope.sandbox.yahoo.com/catalog.php?datatype=r