Amazon Digital Music Recommendation System

Jiali Sun, Yaojun Qin, Yunzhu Liu, Ke Chen

Motivation and Goal

Digital Music plays an important part in our daily life, however, it's usually hard for customers to find the music of their own taste. Amazon is one of the most popular online-shopping website and lots of people usually buy digital music album on Amazon.

We noticed that the although Amazon website provides many kinds of the Digital Music Products, there's few product description, and Amazon does not provide powerful digital music recommendations mechanism based on users preference.

Thus, in this project, we are going to utilize the data analytic skills obtained from class and external resources to solve this problem.

Dataset

We obtain our dataset from http://jmcauley.ucsd.edu/data/amazon/links.html, and we get the data access permission by emailing Professor R. He, J. McAuley. **[Appendix 1]**

Overall, we have two datasets in "json" format. The first dataset (dataset A) mainly provides each product's information, including product ID, reviews, helpfulness, and overall rating. The second dataset (dataset B) provides the relations between products such as "people bought product A also bought product B".

Techniques

Overall, we leverage the following skills: data preprocessing, text mining, sentiment analysis, web scraping, Doc2Vec representing techniques, network analysis, user interface design and other basic data processing skills.

Procedure

First, we use dataset B to explore the detail relationship between every products by using Networking Technique. We consider every product and a node and process the feature "also bought" to discover the edges between nodes. In the history records, If people bought product A also bought product B, there will be an edge exist between node A and B. This relationship is useful for exploring connection between the products. **[Appendix 2]**

Secondly, we rebuild the score measure for each product. There are two reasons bring us to rebuild the score system. The first reason is that there are multiple overall score from different users for each product according to the historical data records and we consider to summarize those scores. The second reason is we noticed there is a feature called "helpful" in our dataset A, we considers this feature measures the quality of score, and it helps to avoid bias score rating. Eventually, we calculated weighted average score for each products. The weighted average formular is in **[Appendix 3].**

Thirdly, we get text reviews from dataset A. We first did data cleaning by using NLTK package; in specific, we remove useless information by deleting stopwords,extra space, and lemmatize. Moreover, we remove the low-frequency and high-frequency words by using CountVectorizer from sklearn package.

In addition, we did text data segmentation by calculating polarity score using textblob package. We discovered that over 75% of polarity scores equal to zero, which means that most of the reviews are just like description instead of showing preference. As a result, we considers the user reviews as the products descriptions. We apply Doc2Vec representing model on review text data and convert each review into a 100 dimensional vector. And we calculate the similarity based on cosine-similarity.

Finally, we design a user interface to show our recommendation results. In this step, we leverage the web scraping technique. Once we input a product ID, our system will provide 5 recommendation products shown on the user interface with product informations and images scraping from amazon website. The recommendation algorithm will be introduce in next part. **[Appendix 4], [Appendix 5]**

Recommendation Algorithms

After all the procedures, we designed an algorithm to produce the recommendation products.

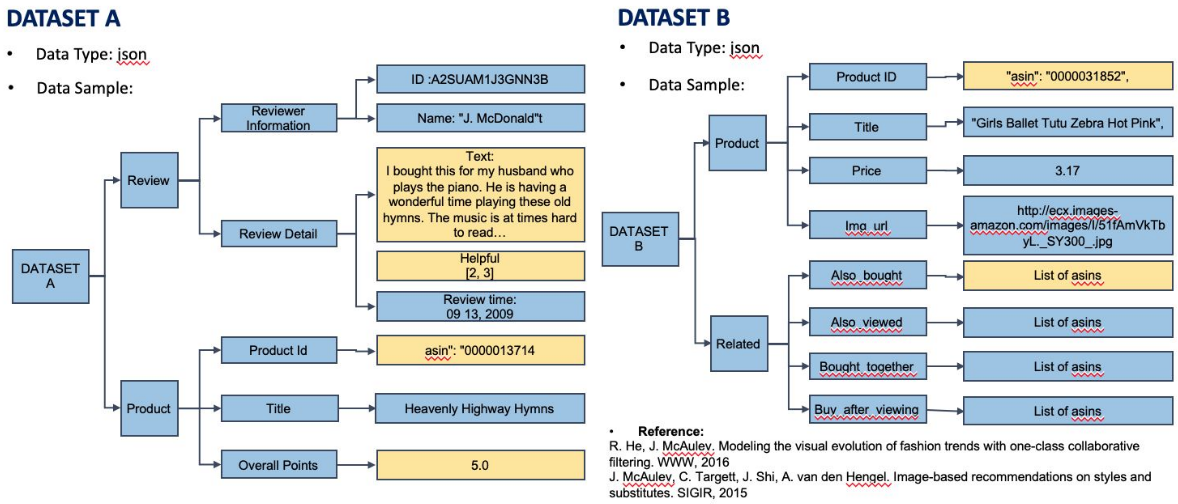
Once given a product id, the system will first consider this product as a target node and feed it into the networking we designed. The algorithm will output all the connected nodes which represent the related products (List A). Moreover, we apply two filters on items in List A. In the first filter, the algorithm will select at most 10 products based on doc2vec similarity score. In the second filter, the algorithm will select top five products sorted by their scores which are from our score re-building system. Eventually, the system will scrap the product information including album name, musicians and album cover picture from Amazon website and displays on our graphical user interface.

Summary

To summarize, aiming to build a more accurate digital music recommendation system through utilization of analytics skills including web scraping, text mining, polarity analysis, doc2vec similarity analysis, graphing, and data preprocessing techniques, we developed an algorithm to calculate similarities between products, assign better scores to products and select products that not only have the most in common with the product a customer input but has the highest scores as well. Since we take both music taste and quality review into consideration, it is hopeful that our system could provide recommendations of a decent quality.

Because of limited time and capabilities of computers, we only pick one of the four relations between products in dataset B to calculate the edges in the network. For further improving, including more relations and assigning different weight to relations might be a legitimate direction.

Appendix:



DATASET A

- Data Type: json
- Data Sample:

DATASET A

Review

Reviewer Information

ID :A2SUAM1J3GNN3B

Name: "J. McDonald"t

Review Detail

Text:
I bought this for my husband who plays the piano. He is having a wonderful time playing these old hymns. The music is at times hard to read…

Helpful
[2, 3]

Review time:
09 13, 2009

Product

Product Id

asin": "0000013714

Title

Heavenly Highway Hymns

Overall Points

5.0

DATASET B

- Data Type: json
- Data Sample:

DATASET B

Product

Product ID

"asin": "0000031852",

Title

"Girls Ballet Tutu Zebra Hot Pink",

Price

3.17

Img_url

http://ecx.images-amazon.com/images/I/51fAmVkTbyL._SY300_.jpg

Related

Also_bought

List of asins

Also_viewed

List of asins

Bought_together

List of asins

Buy_after_viewing

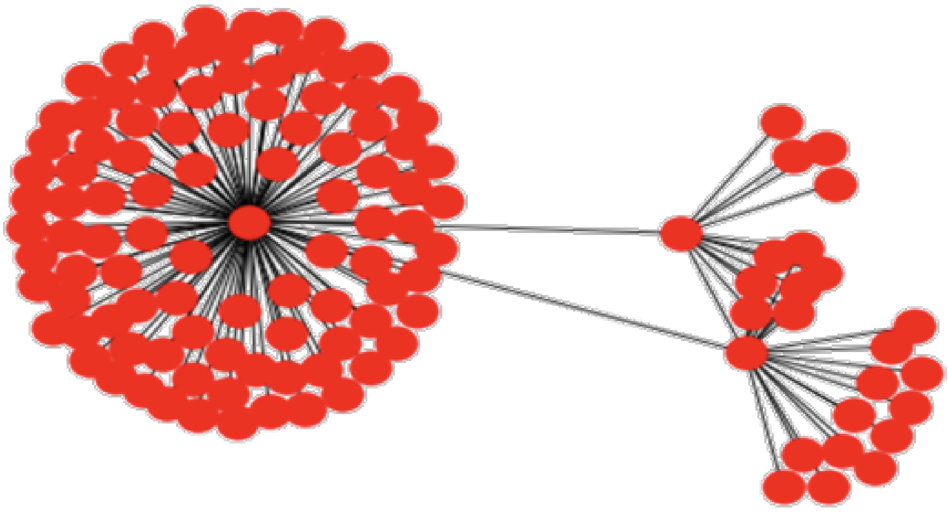
List of asins

- **Reference:**

R. He, J. McAuley. Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW, 2016

J. McAuley, C. Targett, J. Shi, A. van den Hengel. Image-based recommendations on styles and substitutes. SIGIR, 2015

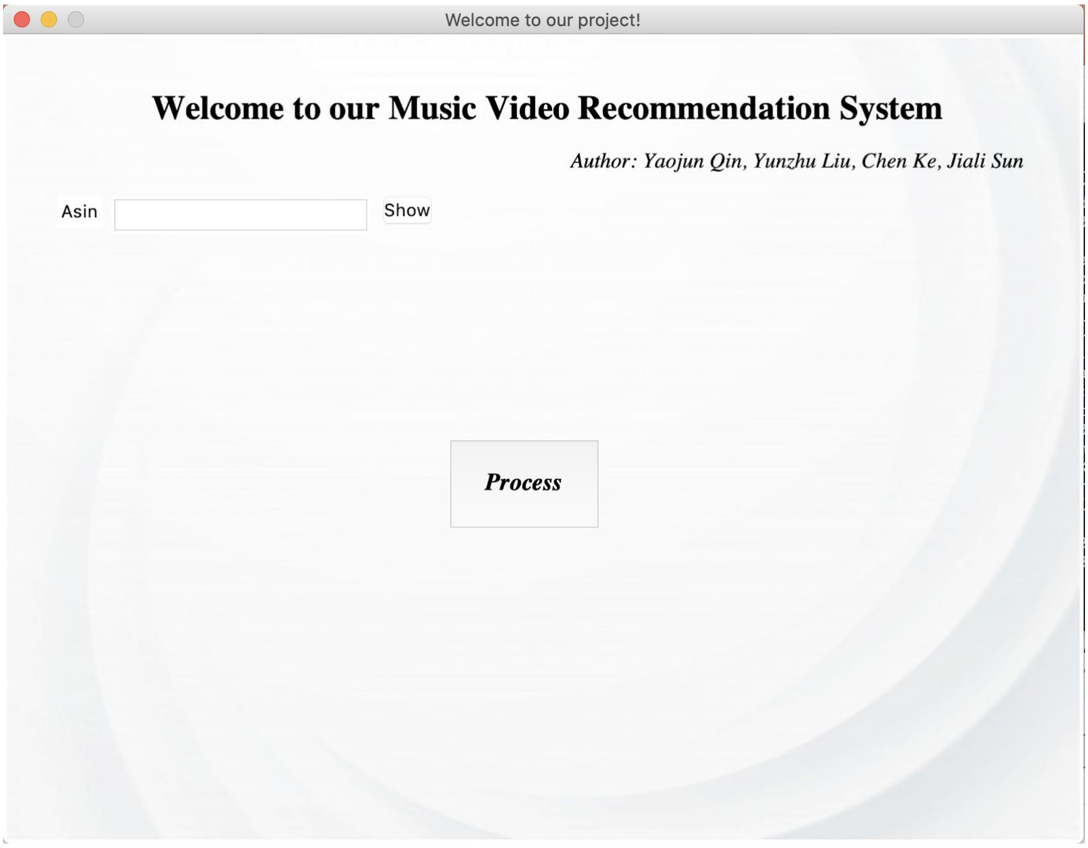
Appendix [1]



Appendix[2]

$$Final\ Score\ for\ Each\ Product = \frac{\sum Base\ Score * (1 + \frac{\#\ of\ Helpfuls}{\#\ of\ Helpful + \#\ of\ Not\ Helpful})}{\sum(1 + \frac{\#\ of\ Helpfuls}{\#\ of\ Helpful + \#\ of\ Not\ Helpful})}$$

Appendix[3]

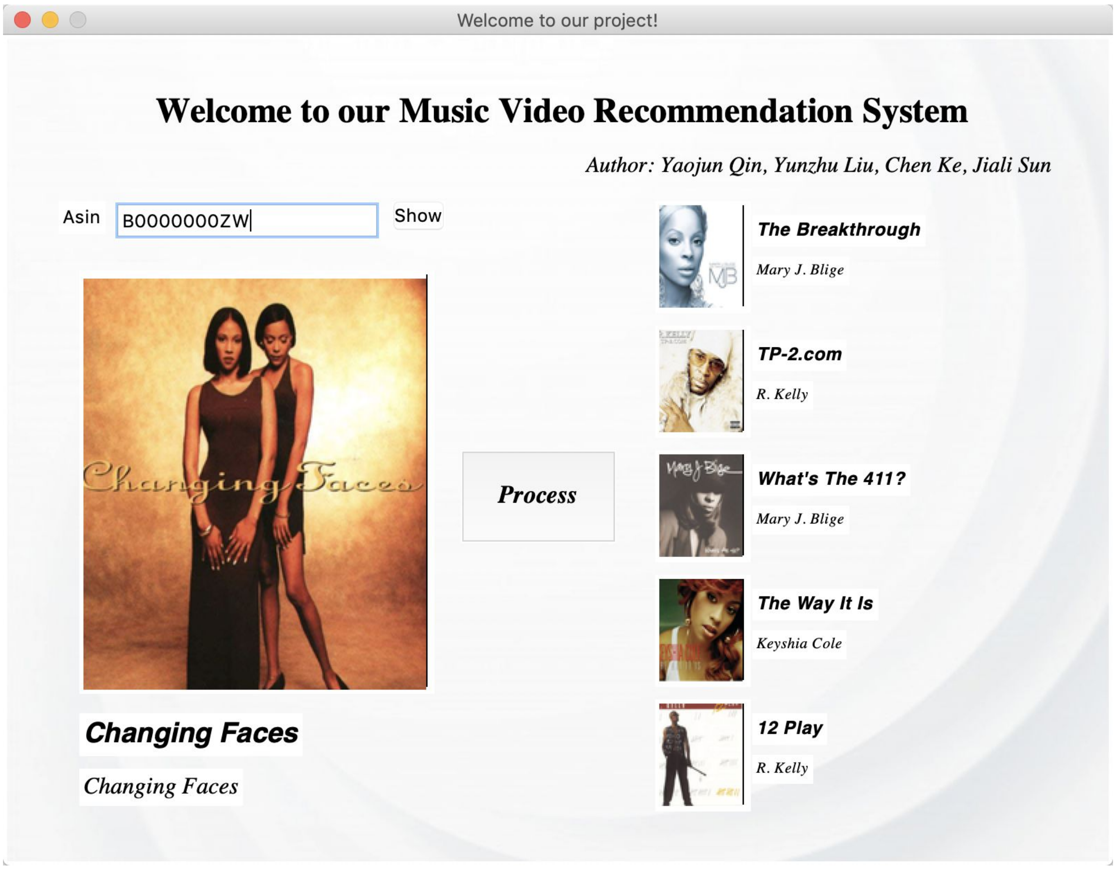


Welcome to our project!

Welcome to our Music Video Recommendation System

Author: Yaojun Qin, Yunzhu Liu, Chen Ke, Jiali Sun

Appendix[4]



Welcome to our project!

Welcome to our Music Video Recommendation System

Author: Yaojun Qin, Yunzhu Liu, Chen Ke, Jiali Sun

Asin B0000000ZW

Changing Faces

Changing Faces

The Breakthrough

Mary J. Blige

TP-2.com

R. Kelly

What's The 411?

Mary J. Blige

The Way It Is

Keyshia Cole

12 Play

R. Kelly

Appendix[5]