FIT5147 Data Exploration and Visualization
Assignment 3
Data Exploration Report

Student Name: **Vinh Phan**
Student Number: **27612937**

April 29, 2018

Contents

1 Terminologies

Terms	Description
Review	A review / feedback customers put in the feedback section of a product
ASIN	Amazon product identification system
Sales Rank	Amazon created a sales rank figure to rate a product
Helpfulness	In a review, we can "like / dislike" it. By using the feature, we can see how valuable the review is
Rating	The figure that customers rate for a particular product from the reviews
Bag words	The list of unique words in a paragraph. In the document context, we combine all review texts, then remove less informative words / symbols, then save it to a bag words
Single words	The word appears only one throughout the paragraph
Frequent pattern	A data mining problem, where we try to find the co-purchasing or correlated items from the transactional data

2 Introduction

2.1 Motivation

Amazon is one of the biggest e-commercial websites around the world. They have a lot of products, reviews, transactions happening every second. Improving future generations of products are one of the main targets of the company. This is because that can definitely increase sale revenue, company popularity, and help other businesses by giving them advice.

However, this gives challenging problems. In this project, we will discuss 2 aspects which can insist to answer a part of the issue

- By analysing customer reviews dataset, we can understand customer satisfaction. For example, whether they like the product or not, what do they think about the product after purchasing
- The relationship between items, this means people who buy product A will be more likely to buy product B. if we can understand that, we can give customers with a better deal in the future. For example, we can put correlated items into the same bucket for wholesale, or simply arrange them nearby in shelves

2.2 Questions

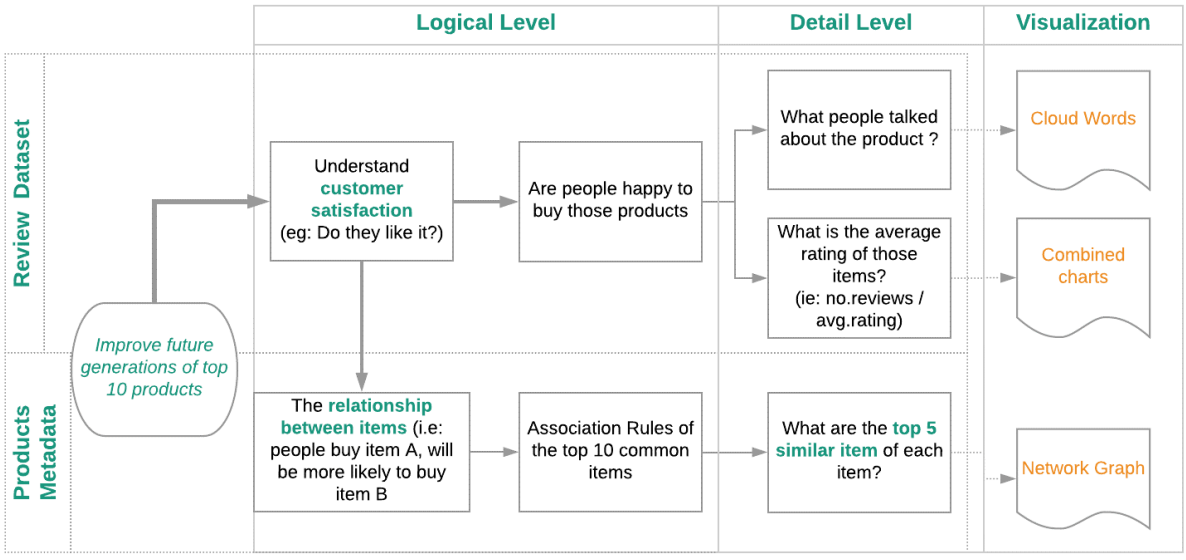


Figure 1: Thinking Flow: Brainstorming Map

The horizontal containers, in the flow chart, show 2 different datasets used in the project, and the vertical ones are the question hierarchy starting with a general logical level then drilling down to specific questions, finally ending up with visualizations

Throughout the document, we will provide images, explanation, and discoveries of the insight derived from those images

2.3 Problem Description

There are many problems we find when working on the project:

- **Dealing with big datasets** We have approximately **550,000 records each dataset**, we download data from SNAP Stanford University, but it is a raw dataset and contain a lot of redundant information
- **Variety of data types** such as JSON, CSV, Text, we have to do a lot of wrangling data to finally produce a formatted dataset
- **Slow performance queries** when we try to combine those dataset into a single table, it comes up with a huge number of records, over 7 million rows due to join tables. To solve it, we store all the data into Microsoft SQL Server database with indexing
- **Text processing** in the report, we try to see what customers put in the review. If they like the product, we will see some emotional words, such as "great", "good", "awesome", "affordable", "worthy"
- **Algorithms** in the project, we have to use Apriori algorithm to derive the insight from data. First, the frequent pattern mining, it helps us to build correlated items corresponding to a product
- **Network Visualization** we deal with a graph with 150 nodes and more than 700 edges for visualizing

2.4 Limitations

The product metadata dataset contains over 500,000 items, but to narrow down we only analyse on **"Music Digital"** category which is over 60,000 items from 1998 to 2014. Also, the purpose of this project is to improve top 10 most popular products which is based on the number of reviews, so the document will focus on analysing only the top 10 items

Even though we do a lot of wrangling data for text analysis, the cloud words seems to be less accurate than what we expect. In another word, the bag words do not represent completely the insight from reviews, some invaluable words still appear. Prefix, suffix words and collocations are not considered

3 Data Wrangling

3.1 Data Description

The datasets are downloaded from SNAP Stanford University:

- Source link: `http://jmcauley.ucsd.edu/data/amazon/links.html`

- Description: the data contains product metadata and review information
- Number of digital music items: **279,899**
- Relevant Reviews: **836,006**

3.1.1 A snapshot of data

Metadata Product

```
Id:   1
ASIN: 0827229534
  title: Patterns of Preaching: A Sermon Sampler
  group: Book
  salesrank: 396585
  similar: 5  0804215715  156101074X  0687023955  0687074231  082721619X
  categories: 2
   |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[122
   |Books[283155]|Subjects[1000]|Religion & Spirituality[22]|Christianity[122
  reviews: total: 2  downloaded: 2  avg rating: 5
    2000-7-28  cutomer: A2JW67OY8U6HHK  rating: 5  votes:  10  helpful:   9
    2003-12-14  cutomer: A2VE83MZF98ITY  rating: 5  votes:   6  helpful:   5
```

Figure 2: A record uncompressing from Product Metadata

Product Review

```
{
    "reviewerID":"AZPWAXJG9OJXV",
    "asin":"5555991584",
    "reviewerName":"bethtexas",
    "helpful":[0,0],
    "reviewText":"A clasically-styled and introverted album,.....
    "overall":5.0,
    "summary":"Enya at her most elegant",
    "unixReviewTime":991526400,
    "reviewTime":"06 3, 2001"
},
```

Figure 3: A record uncompressing from Review json file

3.1.2 Formatted Data Output

Source	Feature	Type	Description
Metadata	asin	bigint	ID of the product
""	title	varchar	Title of the product / product name
""	group	varchar	Product group (Book, DVD, Music)
""	salesRank	int	Amazon sales rank
""	similar	bigint	Similar items recommended by Amazon
""	categories	varchar	Location in product category hierarchy
""	reviews	varchar	Review summary
Review	reviewerId	varchar	ID of the reviewer
""	reviewerName	varchar	Name of reviewer
""	helpful	faction	Helfulness figure rated by visitors
""	reviewText	varchar	Customer review text
""	rating	decimal(2,2)	Customer rating for the product
""	reviewTime	date	Time of the review

3.2 Wrangling Tools

Language	Python 2 on Jupyter Notebook
Libraries	Json, Pandas, Regular Express RE, NLTK, Stop words data, Corpus, wordtokenize, d3Network, ODBC
Workbooks	Data Manipulation with Amazon Products Metadata.ipynb, Data Manipulation with Amazon Reviews Data.ipynb

3.3 Text Data Processing

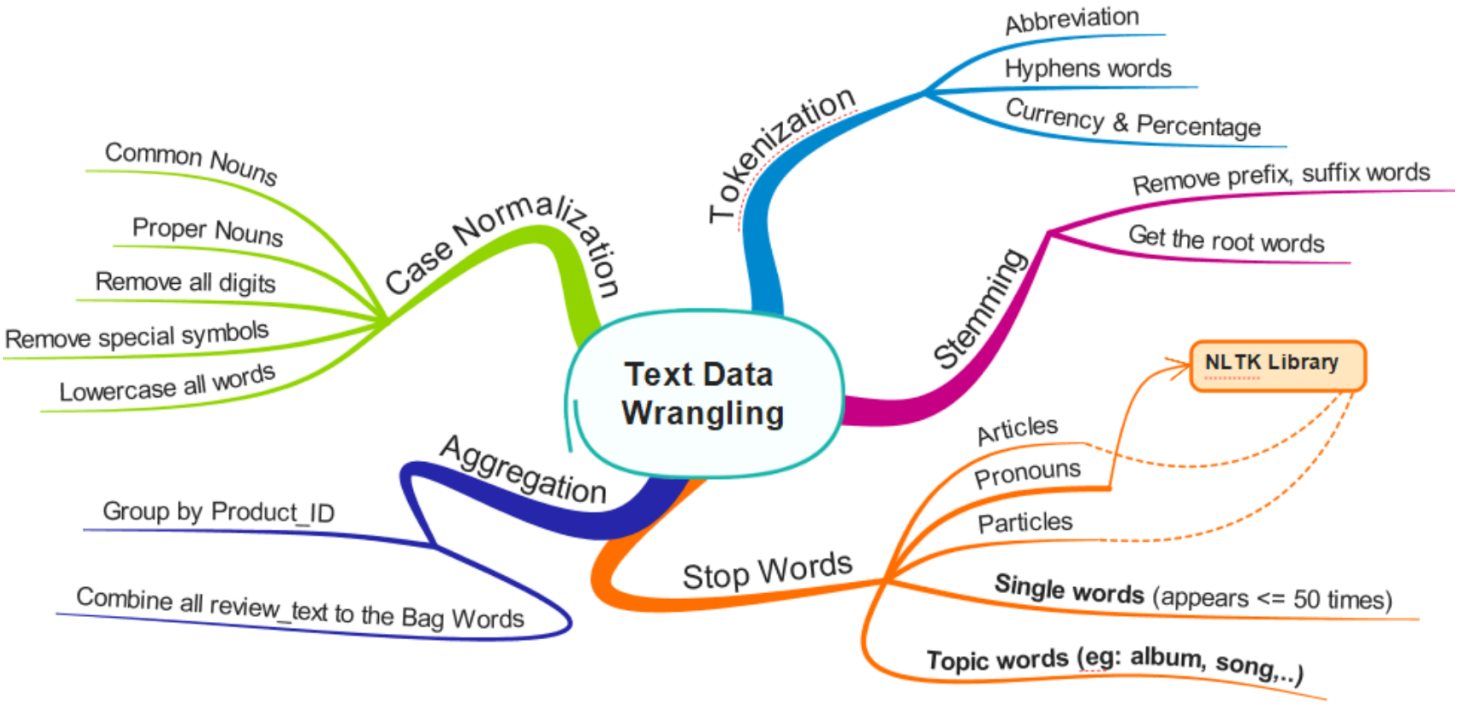


Figure 4: Mindmap: Steps for Text data wrangling

What we have done to wrangle the raw text data:

1. **Loading data**
    - Parsing json file
    - Normalizing data, save it in Pandas
2. **Formatting data**
    - Checking data format
    - Standardize "helpful" from faction into right format
3. **Tokenization**
    - Hyphen words removed
    - Currency
4. **Stop Words**
    - Stop words list from NLTK library
    - Single word (words appear one in the text)
    - Remove topic words, such as "album", "song", "instrument", "music"
5. **Aggregating**
    - Group data by product

- Combine all "review

6. **Performance enhancement**
    - Store all data into tables in MS SQL Server, because data is too big to export

Finally, we have a clean dataset stored in Microsoft SQL tables

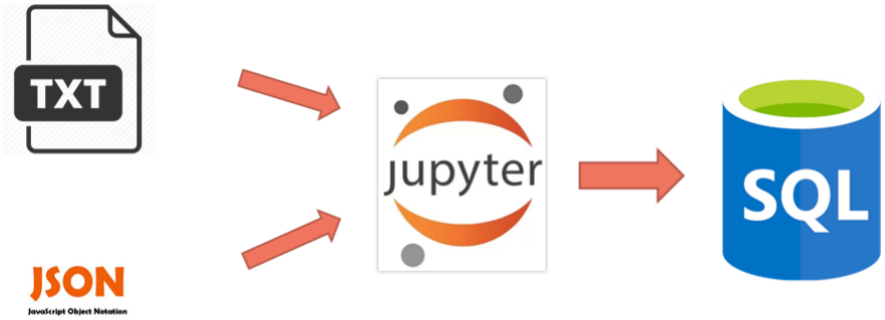


Figure 5: Mindmap: Steps for Text data wrangling

3.4 Text Analysis

Problem 1: after combining all review texts of the first item. We see a lot of less informative words, such as "quote", "album", "song", "one", "music"

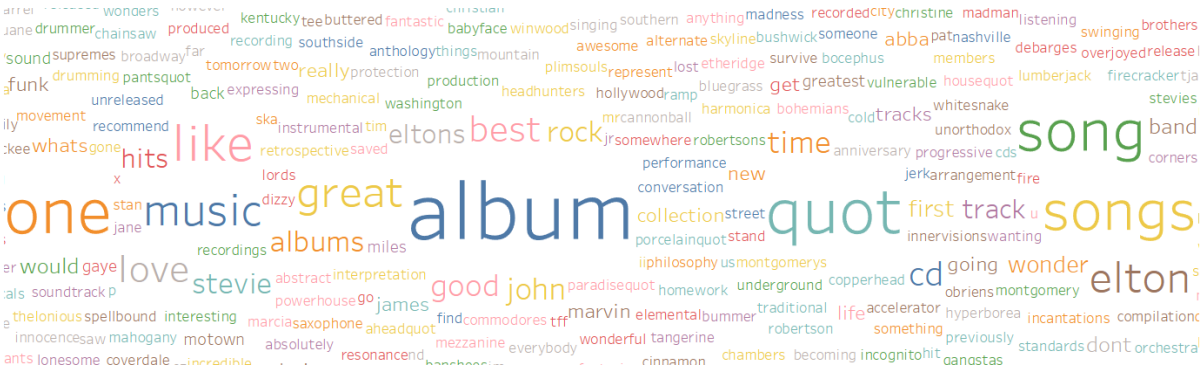


Figure 6: Cloud word before removing topic words

Solution 1: We collect topic words, and insert them into stop words list. Single words are also considered

Problem 2: Aggregating data for network graph visualization is not easy. We have to convert relational data into nodes, and edges format, so that the *"d3Network"* library can process.

Also, we have a lot of edges and nodes at the end, this is not practical to visualize **Solution 2:** Removing all nodes have less than 5 connections, and use D3js to visualize data

4 Data Checking

4.1 Data Distribution with Histogram

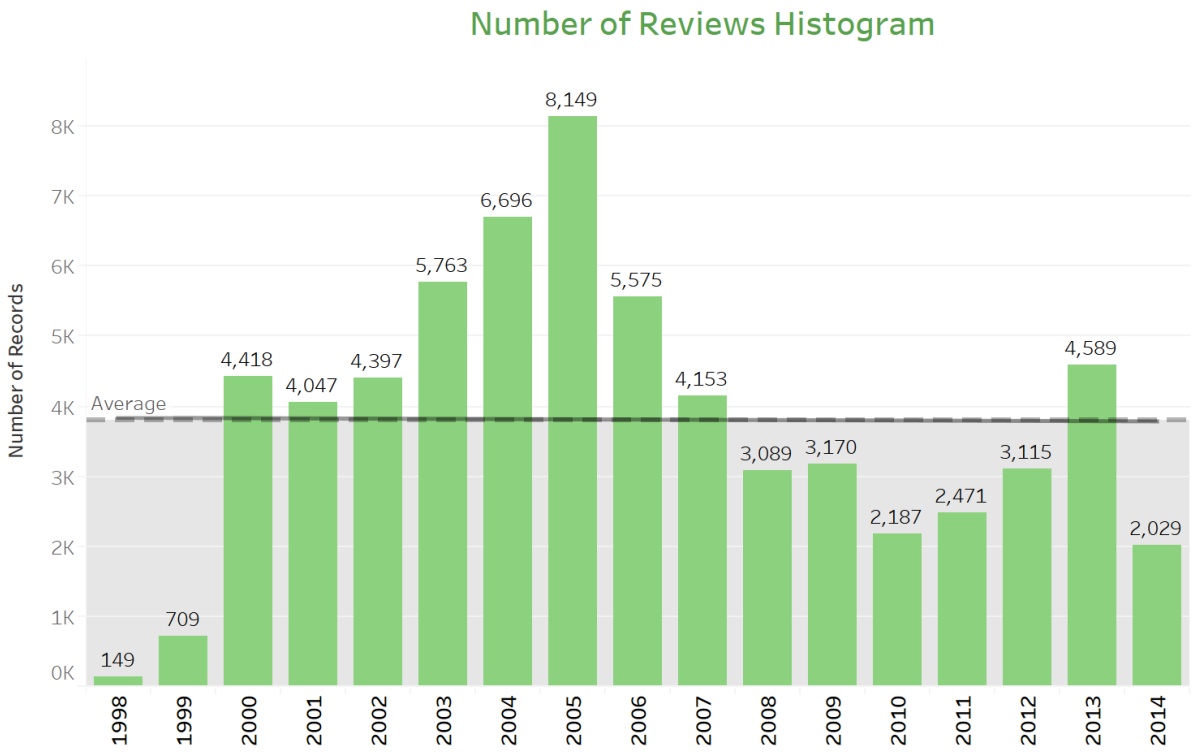


Figure 7: Histogram: number of reviews from 1998 to 2014

The graph illustrates the distribution of reviews over years from 1998 to 2014. The dash line stands for the average review number which is nearly 4000 over 17 years period

We can see that the maximum number in 2005 which is 8.149 reviews, but in 2010 the figure went down significantly to 2,187 due to the financial crisis since 2008

At the beginning in 1988, Amazon didn't have many products, therefore the number was very low

5 Data Exploration

5.1 What are the top 10 highest reviews?

The diagram shows top 10 items that have highest number of reviews from 1998 to 2014. In the container, there are 3 labels that are ASIN, Title of product, number of reviews respectively

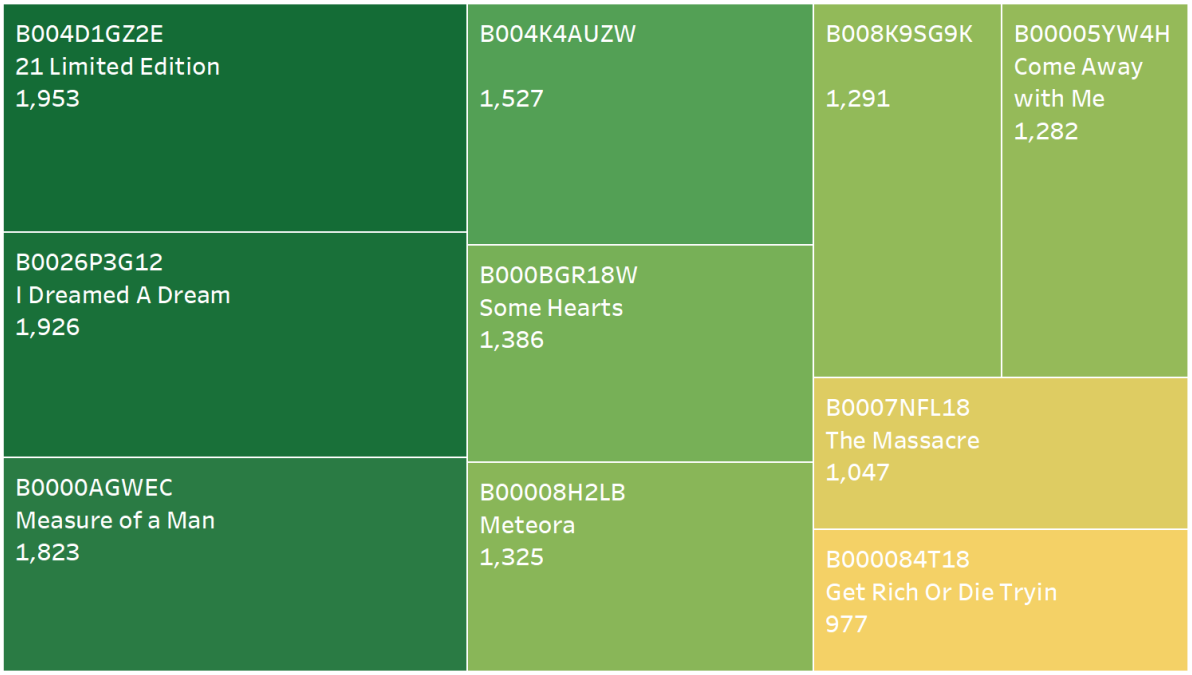


Figure 8: Top 10 albums have highest reviews from 1989 to 2014

5.2 Are people happy to buy those items?

The exploration process for this question includes 2 main steps:

- Firstly, building a bag of words, we can see what keywords they use to describe the product, whether they are positive or negative words
- Secondly, we look at the average rating by customers against the number of reviews. This shows what rate customers give for each product

5.2.1 What did they talk about those products?

To concentrate on analysing at detail, we will look at the top 1 album which is *"B004D1GZ2E",Adele: 21 Limited Edition album,* in the ASIN Amazon product identification

Data Preparation Process

- We clean the whole "review" dataset by removing symbols, digits
- Merging all review texts to a single paragraph
- Tokenizing the paragraph, meaning that we split it by word and store them in a long list
- Creating a "stop words" list which is defined by
  - Articles, pronouns, particles which are supported by NLTK library in Python
  - Single word which is the word appears only one in the paragraph
  - Look at top 100 most frequent words only
- Saving the bag words result in MS SQL tables

- Using Tableau to visualize data

Statistical Summary table

Feature	Value
Total number of Reviews	1,953
Total number of words in praragraph	112,868
Total number of stop words removed	62,291
Total number in Bag words	13,924

Visualization

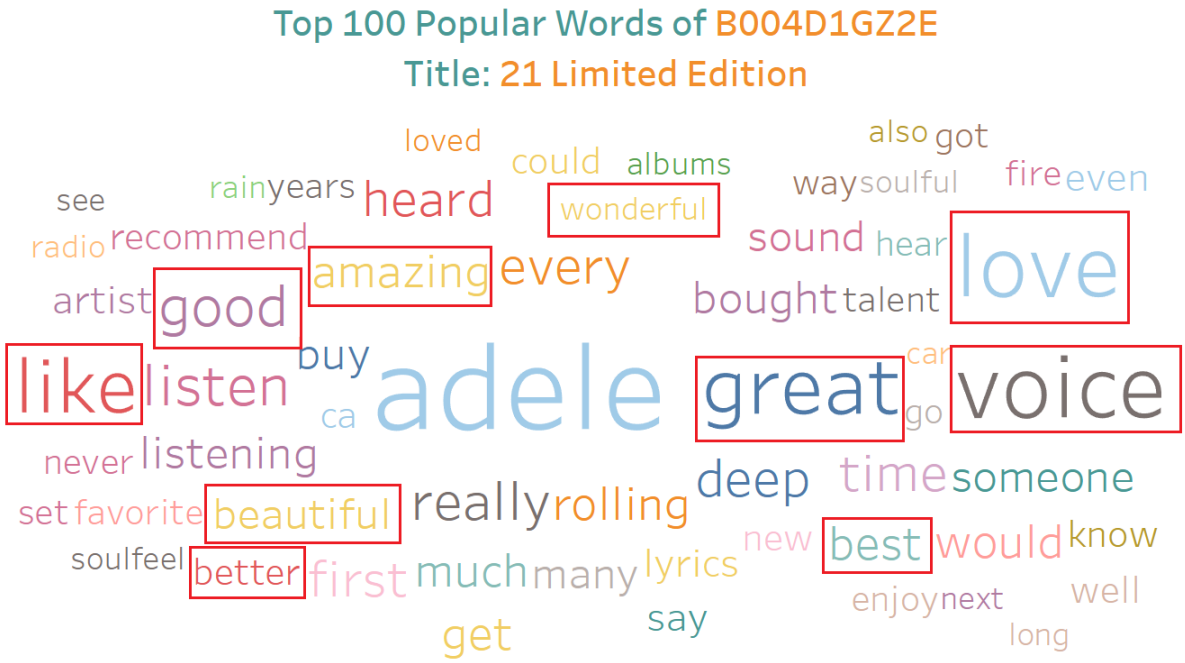


Figure 9: Cloud Words of album "Adele: 21 Limited Edition" with Tableau

Discoveries

- Looking at the graph, we can see that people say “good”, “amazing”, “wonderful”, and “adele”
- Indeed, “21 Limited Edition” is the best album of Adele since 2011. It achieved a lot of awards, such as Grammy Awards for album of the year
- By looking at the visualization, we can see that audiences like her album, and one of the reasons is because of her voice

Limitations

- The result is not as good as what we expect. This is because we couldn’t get rid of prefix, suffix of a word to get the root. For example, “song” and “songs” are the same meaning, so they should be combined into one word
- Collocation should be considered. At the moment, we just consider only single word

5.2.2 What is the average rating against reviews?

We want to see the correlation between "average rating" and "number of reviews" for top 10 most popular items in 2014. Whether or not the albums with highest reviews were also high ranking

Visualization

Figure 10: Dashboard: Average review rating against No.Reviews

Discoveries

- The graph demonstrates top 3 items have the highest rating, meaning that they are very good product with high sales, high review and high ranking
- However, the "Come Away with Me" album had a very high ranking and rating, but the reviews were low. It is also a worthy product to buy
- "The Massacre" had 1,047 reviews in 2014, but the rating was just 2.73. it seems to me that it's not a really good album to keep

Limitations

- The number of features available is short, so that we can't build a correlation matrix to see a bigger picture of data

5.3 What the relationship looks like between items?

There is a real demand that customers have demands, but they don't know what they should buy actually. By using recommendation system, we can recommend them which products are similar to what they are looking for

In the section, we introduce the results of **"Apriori"** algorithm by visualizing it on an interactive Network Graph in R with D3js

Exploration Process

- We have over 80 thousand items in the database, this is not practical to put all of them into the network graph. Thus, we select only top 15 items of highest reviews
- From the metadata product, we can extract the frequency of each item (ie: asin code). All items with frequency less than 5 will be removed. In another word, we just keep nodes with more than 5 edges. At the end, we will have a list of itemset
- Store it into 2 tables in MS SQL Server for visualization
- In R, we used **"d3Network"** which generates interactive network visualizations using D3 javascript library

Visualization and Discoveries

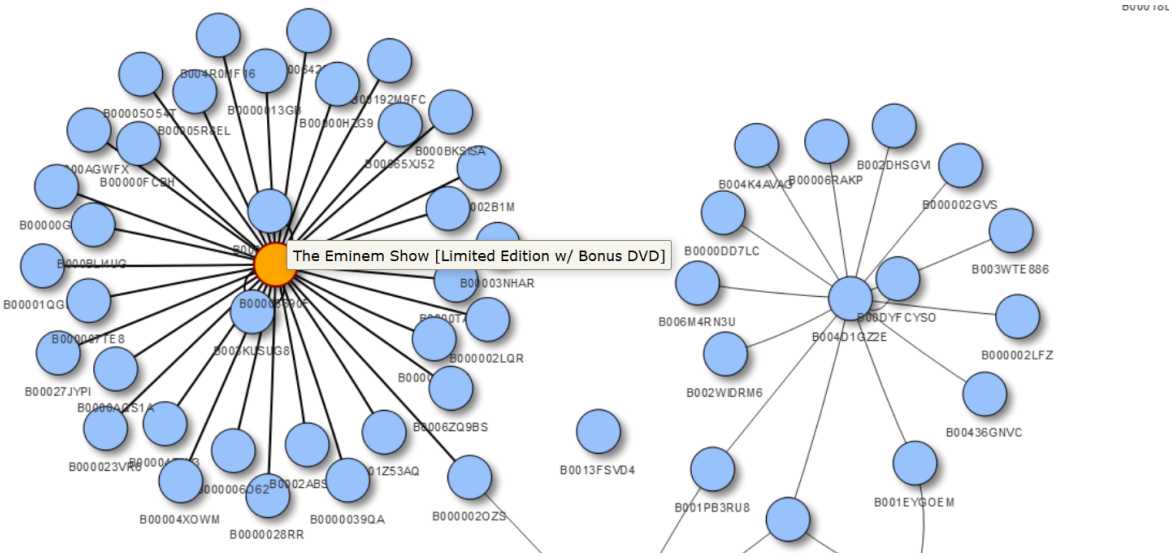
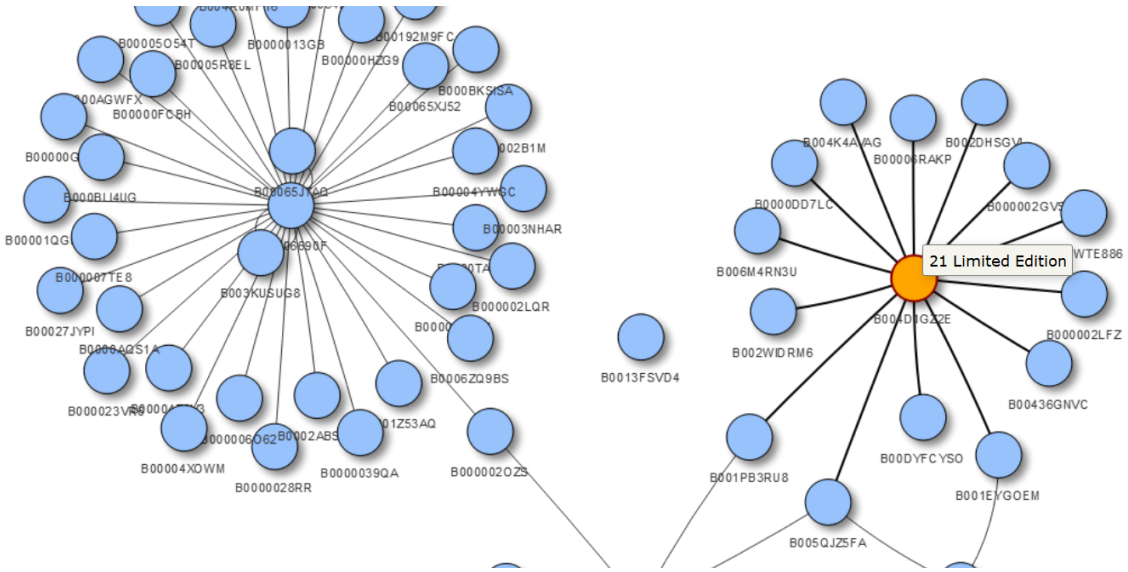


Figure 11: The network graph, the centroid is node which is linked to other similar node in the network

"The Eminem Show" has most connections from other nodes. This means it is the most popular item in Amazon in 2014



The second one is "21 Limited Edition" of Adele. Although it has less connection than "The Eminem Show", it has the most reviews in 2014

By looking at the big image, we can see what albums customers pay attention in 2014. Also, those items had high demand in the past

6 Conclusion

I have worked in different stages in the process of analysing data from the scratch. I have to read different materials to choose a right data source which fits the project's tasks, and my capability

At the beginning, it was difficult to deal with a big dataset. Also, the data is not in the right format, which requires me to wrangling a lots

After the ideas to analyse data, it's challenging as well, because most of the ideas is not suitable with the dataset. However, if i merge many different datasets, it will be not enough time to finish the project

Visualization tools is very interesting, especially when i work with D3js and Shiny. I have learnt a lot of new things from the project by going through different stages of analysing data. Now it is the time to create a web-based with interactive visualizations to communicate the results

7 References

- D3Network. (n.d.). Retrieved from `http://christophergandrud.github.io/d3Network/`
- Apriori algorithm. (2018, February 17). Retrieved from `https://en.wikipedia.org/wiki/Apriori_algorithm`

- Static and dynamic network visualization with R. (n.d.). Retrieved from `http://kateto.net/network-visualization`
- 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