Mini Project

Book Recommendation System

Agenda

- Introduction
- Exploratory Data Analysis
- Build Model
- Conclusion

Introduction

- Tập dữ liệu: Book-Crossing: User review ratings từ Kaggle
- Tập dữ liệu là dữ liệu được cung cấp từ 278.858 người dùng, cung cấp 1.149.780 đánh giá (explicit/implicit) về 271 379 quyển sách
- Dữ liệu có cột thuộc tính: Số thứ tự bản ghi, ID người dùng, địa chỉ, tuổi, ID sách, đánh giá, tiêu đề sách, tác giá, năm xuất bản, nhà xuất bản, ảnh, ngôn ngữ sử dụng, thể loại, thành phố, liên bang, đất nước

Exploratory Data Analysis

- Thực hiện một số khảo sát về dữ liệu
  - Missing value
  - Scatter
  - Distribution

  …

Exploratory Data Analysis

Tập dữ liệu chứa 231.210 bản ghi và 18 cột thuộc tính.

Các loại dữ liệu của các thuộc tính bao gồm 3 thuộc tính kiểu số nguyên rời rạc,2 thuộc tính số thực liên tục và 14 thuộc tính kiểu đối tượng.

Sử dụng không gian bộ nhớ ít nhất là 33.5 megabyte (MB).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 231210 entries, 0 to 231209
Data columns (total 19 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   Unnamed: 0           231210 non-null  int64
 1   user_id              231210 non-null  int64
 2   location             231210 non-null  object
 3   age                  231210 non-null  float64
 4   isbn                 231210 non-null  object
 5   rating               231210 non-null  int64
 6   book_title           231210 non-null  object
 7   book_author          231210 non-null  object
 8   year_of_publication  231209 non-null  float64
 9   publisher            231209 non-null  object
 10  img_s                231209 non-null  object
 11  img_m                231209 non-null  object
 12  img_l                231209 non-null  object
 13  Summary              231209 non-null  object
 14  Language             231209 non-null  object
 15  Category             231209 non-null  object
 16  city                 228883 non-null  object
 17  state                226817 non-null  object
 18  country              223120 non-null  object
dtypes: float64(2), int64(3), object(14)
memory usage: 33.5+ MB
```

Exploratory Data Analysis

	Unnamed: 0	user_id	location	age	isbn	rating	book_title	book_author	year_of_publication	publisher	img_s	
0	0	2	stockton, california, usa	18.0000	0195153448	0	Classical Mythology	Mark P. O. Morford	2002.0	Oxford University Press	http://images.amazon.com/images/P/0195153448.0...	http://images.amazon.com/images/P/019
1	1	8	timmins, ontario, canada	34.7439	0002005018	5	Clara Callan	Richard Bruce Wright	2001.0	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/000
2	2	11400	ottawa, ontario, canada	49.0000	0002005018	0	Clara Callan	Richard Bruce Wright	2001.0	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/000
3	3	11676	n/a, n/a, n/a	34.7439	0002005018	8	Clara Callan	Richard Bruce Wright	2001.0	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/000
4	4	41385	sudbury, ontario, canada	34.7439	0002005018	0	Clara Callan	Richard Bruce Wright	2001.0	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.0...	http://images.amazon.com/images/P/000

Bản ghi trong tập dữ liệu

Exploratory Data Analysis

```
Unnamed: 0                0
user_id                   0
location                  0
age                       0
isbn                      0
rating                    0
book_title                0
book_author               0
year_of_publication       1
publisher                 1
img_s                     1
img_m                     1
img_l                     1
Summary                   1
Language                  1
Category                  1
city                   2327
state                  4393
country                8090
dtype: int64
```

Thống kê số giá trị null

Exploratory Data Analysis

- Xóa bỏ các cột thuộc tính: Số thứ tự, địa chỉ, độ tuổi, ID của sách, năm xuất bản, ảnh, thành phố, liên bang, ngôn ngữ, đất nước
- Xóa các bản ghi chứa giá trị null

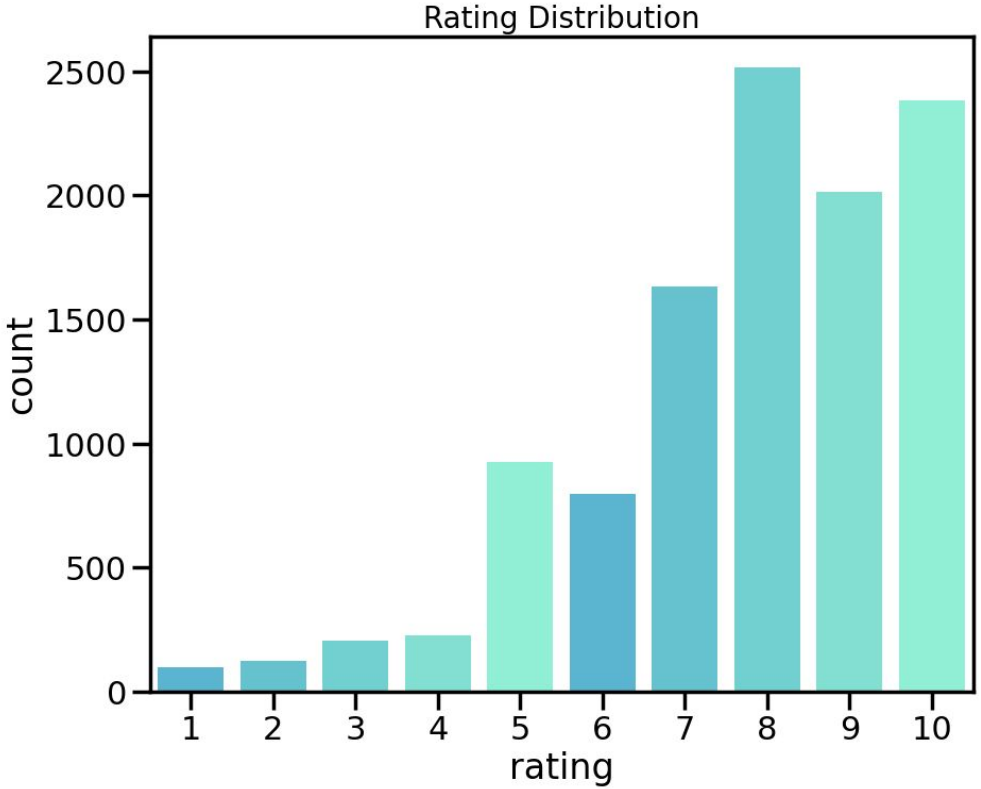
```
df.dropna(inplace=True)
df.reset_index(drop=True, inplace=True)

df.drop(columns = ['Unnamed: 0','location','isbn',
                   'img_s','img_m','city','age',
                   'state','Language','country',
                   'year_of_publication'],axis=1,inplace = True)
```

Exploratory Data Analysis

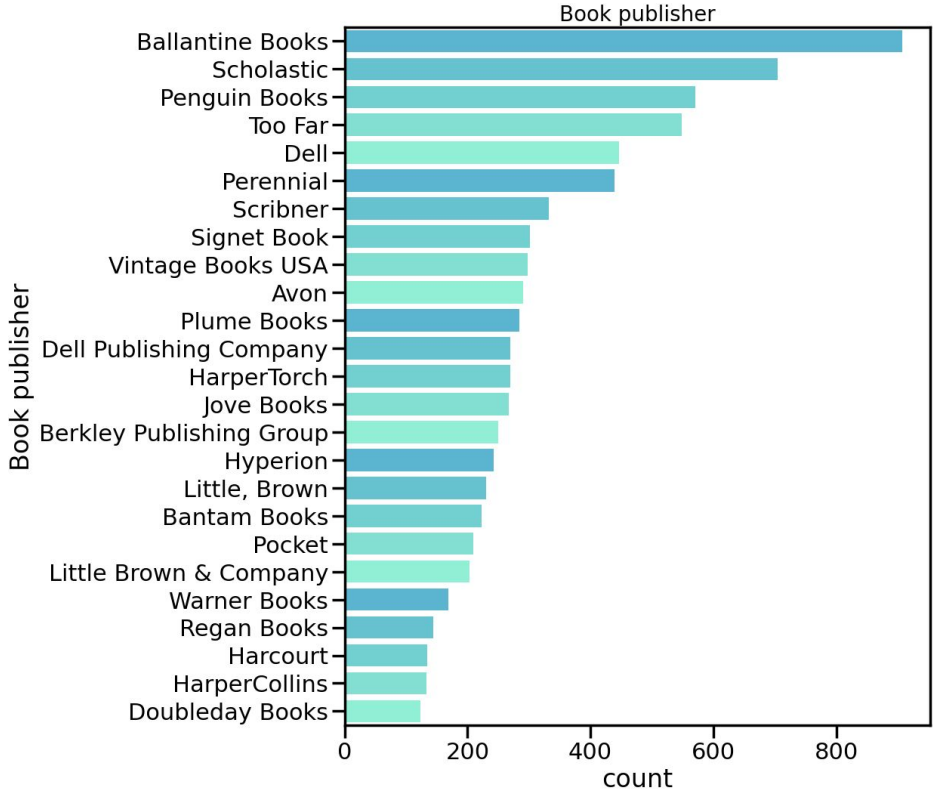
- Loại bỏ các bản ghi có đánh giá implicit, các giá trị đặc biệt trong cột thuộc tính

```
df.drop(index=df[df['Category'] == '9'].index, inplace=True) #remove 9 in category

df.drop(index=df[df['rating'] == 0].index, inplace=True) #remove 0 in rating

df['Category'] = df['Category'].apply(lambda x: re.sub('[\W_]+',' ',x).strip())
```

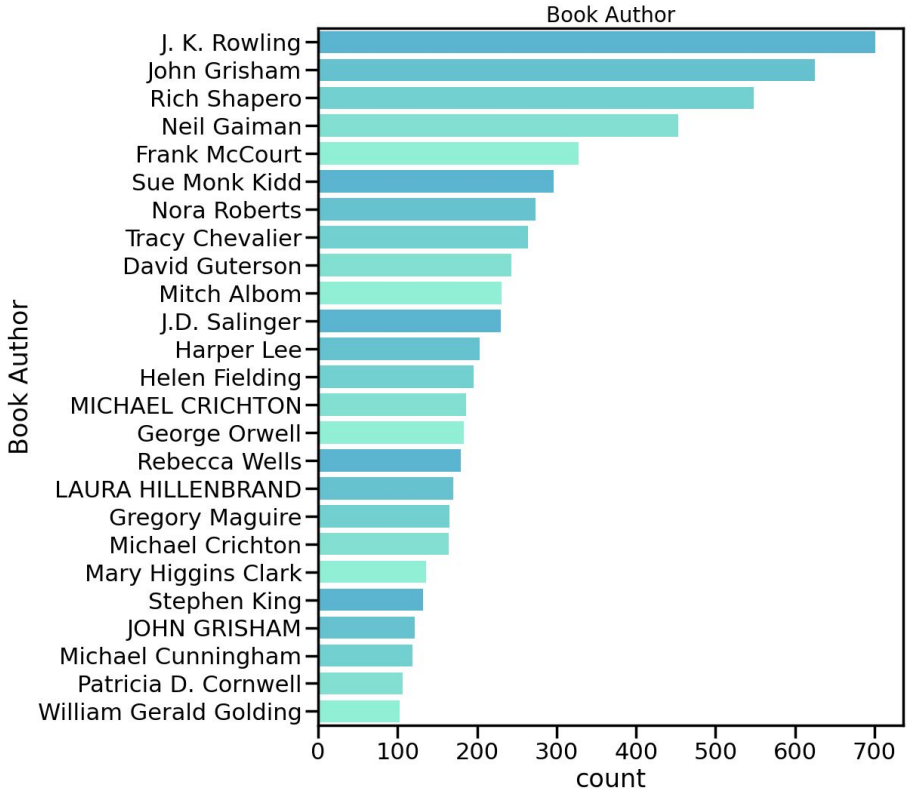
Exploratory Data Analysis



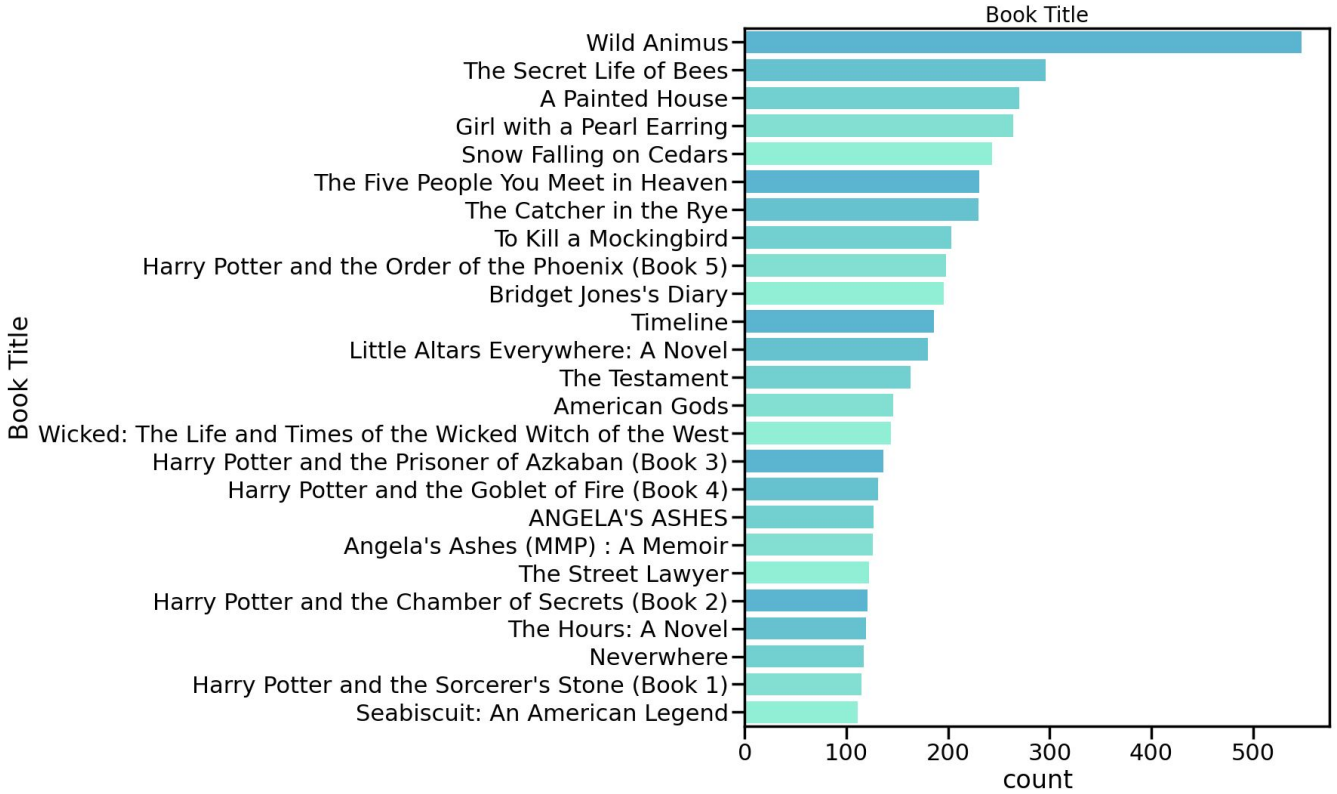
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Build Model: CF Item - Based

- Chia tập sách thành 2 tập: sách phổ biến và sách hiếm gặp

```
book_title = str(book_title)
if book_title in df['book_title'].values:

    rating_counts = pd.DataFrame(df['book_title'].value_counts())
    rare_books = rating_counts[rating_counts['book_title'] <= 180].index
    common_books = df[~df['book_title'].isin(rare_books)]
```

Build Model: CF Item - Based

- Với tập sách hiếm gặp: gợi ý các item một cách ngẫu nhiên
- Với tập sách phổ biến: tính độ tương đồng các item dựa trên ma trận đánh giá

```
[130] item_based_recommender('Fahrenheit 451')

      There are no recommendations for this book
      Try:

      Bridget Jones's Diary

      The Secret Life of Bees
```

```
[145] item_based_recommender('The Street Lawyer')

      You may also like these books
      1                                The Secret Life of Bees
      2                             Girl with a Pearl Earring
      3    Tuesdays with Morrie: An Old Man, a Young Man,...
      4                                          The Testament
      5                  The Red Tent (Bestselling Backlist)
      Name: book_title, dtype: object
```

Build Model: Content - Based

- Với tập sách hiếm gặp: gợi ý các item một cách ngẫu nhiên
- Với tập sách phổ biến: chọn các thuộc tính “book_title”, “book_author”, “publisher’ , “category” làm đặc trưng quan trọng.
  - Tạo cột mới “combined_features”
  - sử dụng Countvectorizer để chuyển chuỗi văn bản trong “combined_features” thành ma trận count_matrix
  - Sử dụng độ tương đồng cosine để tính sự tương đồng

Build Model: Content - Based

```
[235] content_based_recommender('Harry Potter and the Order of the Phoenix (Book 5)')
```

```
You may also like these books
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Wicked: The Life and Times of the Wicked Witch of the West
```

```
[239] content_based_recommender('The Street Lawyer')
```

```
You may also like these books
The Testament
A Painted House
The Catcher in the Rye
Harry Potter and the Order of the Phoenix (Book 5)
Wicked: The Life and Times of the Wicked Witch of the West
```

Build Model: Content - Based

- Với tập sách phổ biến: chọn các thuộc tính “summary” làm đặc trưng quan trọng.
  - Loại bỏ các kí tự đặc biệt, chuyển đổi chữ in hoa thành chữ thường, tách từ, loại bỏ từ dừng (stop words)
  - Sử dụng CountVectorizer để chuyển chuỗi văn bản thành ma trận count_matrix
  - Sử dụng độ tương đồng cosine để tính sự tương đồng

Build Model: Content - Based

```
[241] content_based_recommender2('The Street Lawyer')

      You may also like these books
      American Gods
      The Testament
      Wild Animus
      Timeline
      To Kill a Mockingbird
```

```
[240] content_based_recommender2('Harry Potter and the Order of the Phoenix (Book 5)')

      You may also like these books
      Harry Potter and the Prisoner of Azkaban (Book 3)
      Harry Potter and the Goblet of Fire (Book 4)
      Harry Potter and the Sorcerer's Stone (Book 1)
      Harry Potter and the Chamber of Secrets (Book 2)
      Timeline
```

Build Model

```
item_based_recommender('Harry Potter and the Order of the Phoenix (Book 5)')
```

```
You may also like these books
1                    Timeline
2        Snow Falling on Cedars
3       The Secret Life of Bees
4               A Painted House
5    Girl with a Pearl Earring
Name: book_title, dtype: object
```

```
[242] content_based_recommender2('Harry Potter and the Order of the Phoenix (Book 5)')
```

```
You may also like these books
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Harry Potter and the Chamber of Secrets (Book 2)
Timeline
```

```
[244] content_based_recommender('Harry Potter and the Order of the Phoenix (Book 5)')
```

```
You may also like these books
Harry Potter and the Chamber of Secrets (Book 2)
Harry Potter and the Prisoner of Azkaban (Book 3)
Harry Potter and the Goblet of Fire (Book 4)
Harry Potter and the Sorcerer's Stone (Book 1)
Wicked: The Life and Times of the Wicked Witch of the West
```

Thank you