Book Recommender System

A Cold Start Problem

How do we recommend books without past user's preference or information?

· 3 Parts Project·

Part 1	Data Collection
Part 2	Data Exploration
Part 3	Recommendation System

Stakeholders

Class of DSIF 4

A Big Thank You

Divya

Ben

Shao Quan

A good teacher can inspire hope, ignite the imagination, and instill a love of learning

– Brad Henry Love

Part 1: Data Collection

Step 1: Scrapy Spider

Using a scrapy spider created by: havanagrawal

Scrapped 28 list of books, grabbing the details of **175,406** books.

#	List Scrapped
1	Best_Page_Turners_with_Redeeming_Social_Value
2	Couldn_t_Put_The_Book_Down_
3	Books_you_wish_more_people_knew_about_Part_II
4	Best_Books_of_the_21st_Century
5	Books_that_Blew_Me_Away_and_that_I_Still_Think_About_of_all_types_
6	Best_Unknown_but_must_be_Known_books_
7	1001_Books_You_Must_Read_Before_You_Die
8	Books_That_Everyone_Should_Read_At_Least_Once
9	Lesser_Known_Authors
10	What_To_Read_Next
11	The_Most_Influential_Books
12	100_Books_to_Read_in_a_Lifetime_Readers_Picks
13	Books_That_Should_Be_Made_Into_Movies
14	Must_Read_Non_Fiction
15	I_m_glad_someone_made_me_read_this_book
16	Best_Books_Ever
17	Books_With_a_Goodreads_Average_Rating_of_4_5_and_above_and_With_At_Least_100_Ratings
18	Books_that_Changed_the_Way_You_View_Life
19	100_Mysteries_and_Thrillers_to_Read_in_a_Lifetime_Readers_Picks
20	Read_Them_Twice_At_Least
21	Books_You_Wish_More_People_Knew_About
22	Best_Young_Adult_Books
23	Interesting_and_Readable_Nonfiction
24	Best_Books_of_the_18th_Century
25	Best_Books_of_the_Decade_2000s
26	Best_for_Book_Clubs
27	Best_Science_Fiction_Fantasy_Books
28	Best_Books_of_the_Decade_1990s

Step 2: Scraping Wikipedia

Using a package called **wptools** to scrape Wikipedia's info-box for titles with null values

52 Null Values Filled.

#	Features Scrapped
1	URL
2	Title
3	Author
4	Number of ratings
5	Number of reviews
6	Average ratings
7	Number of pages
8	Language
9	Original publish year
10	Genres

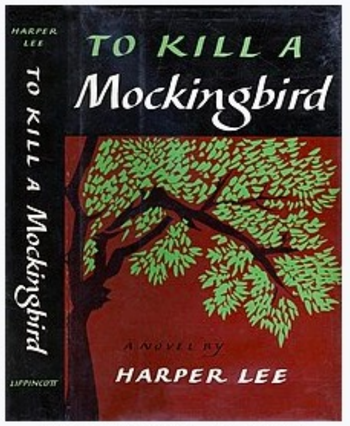
Issues

1. Multiple Duplicate books and URL in different list
After dropping the duplicates only **66,938** unique books remain

2. Null Values: **52,118**

3. Non-standardized data. Plural-non plural, word joined together, additional descriptions for some.

To Kill a Mockingbird



First edition cover – late printing

Author	Harper Lee
Country	United States
Language	English
Genre	Southern Gothic · Bildungsroman
Published	July 11, 1960
Publisher	J. B. Lippincott & Co.
Pages	281

Step 3: Filling in the Language Column

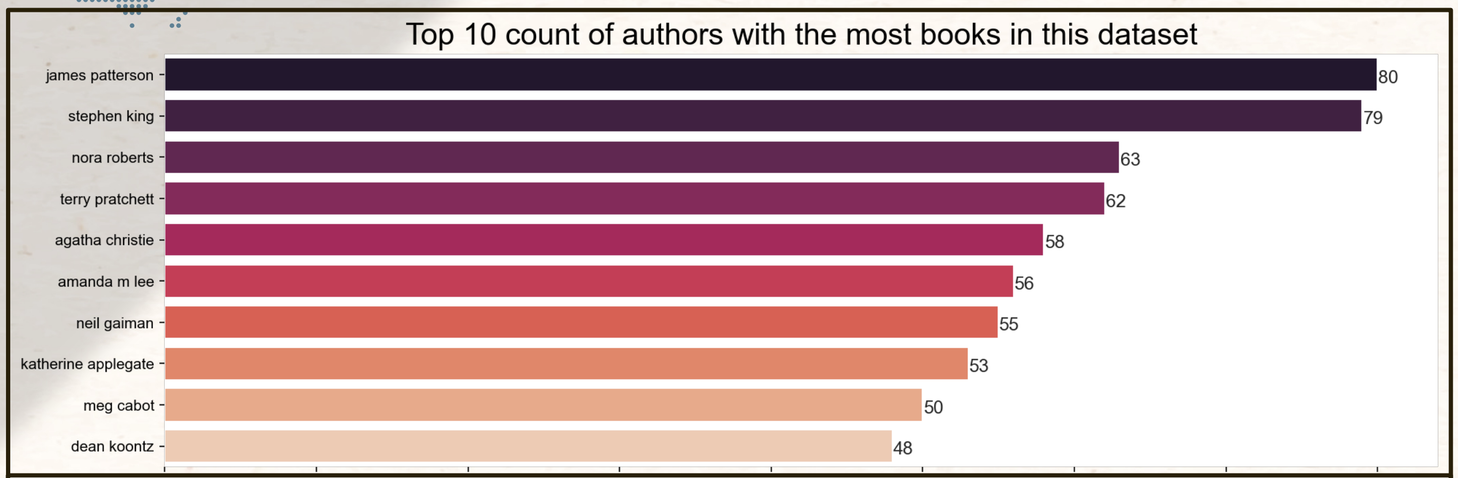
Of the 52,118 Null values, 10% of it was from the language column.

1. LangID to detect language of the title
2. Scrape Wikipedia ISO language code page using BeautifulSoup
3. Map the languages and fill in the null values

Part 2: Data Exploration

Authors

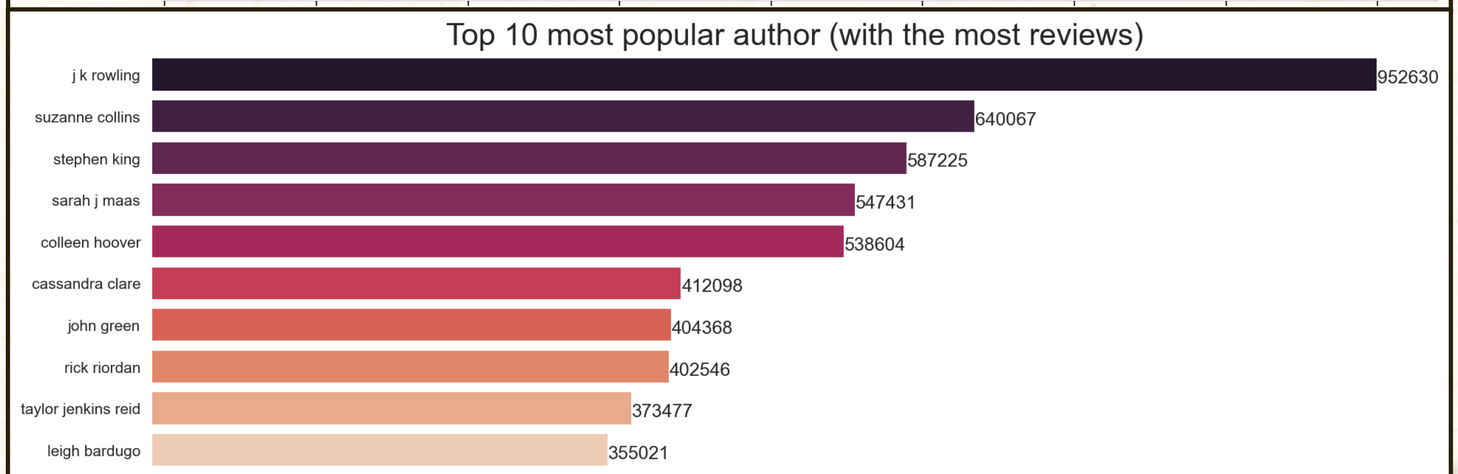
Exploring the Features: Authors, Languages, Titles, and Genres



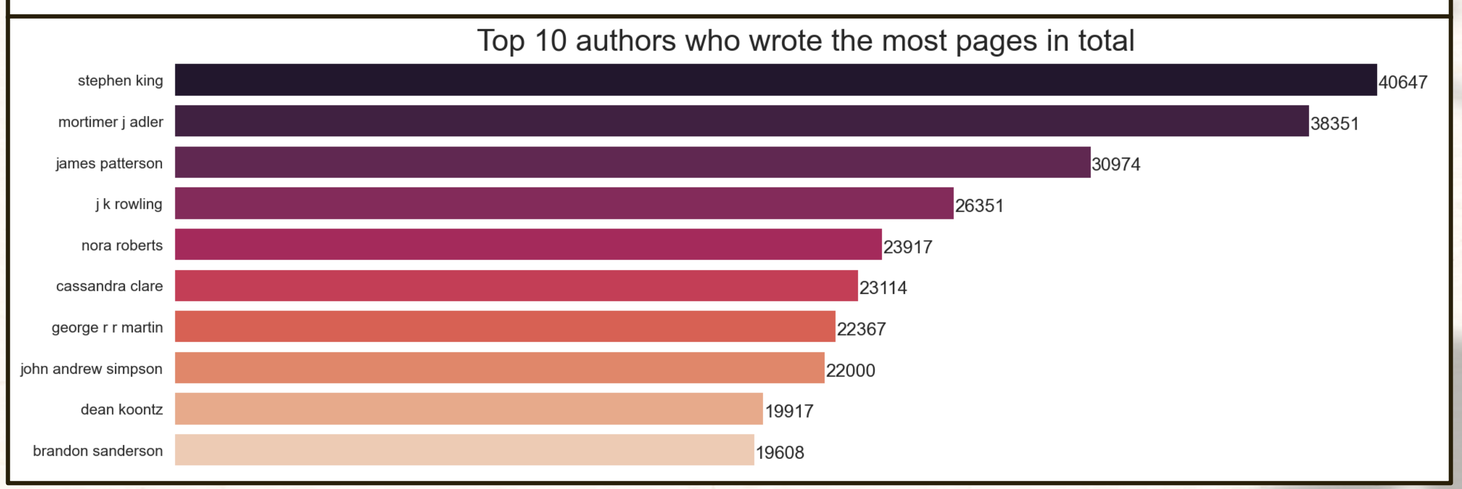
Anonymous is actually at the top with 94 books but they could be from 94 different authors or 1

Interesting Observation

The only Author that appear in all 3 top 10 list is Stephen King



J.K. Rowling is at the top with almost **1 million reviews**. The difference between the second author Suzanne Collins is almost **300,000**

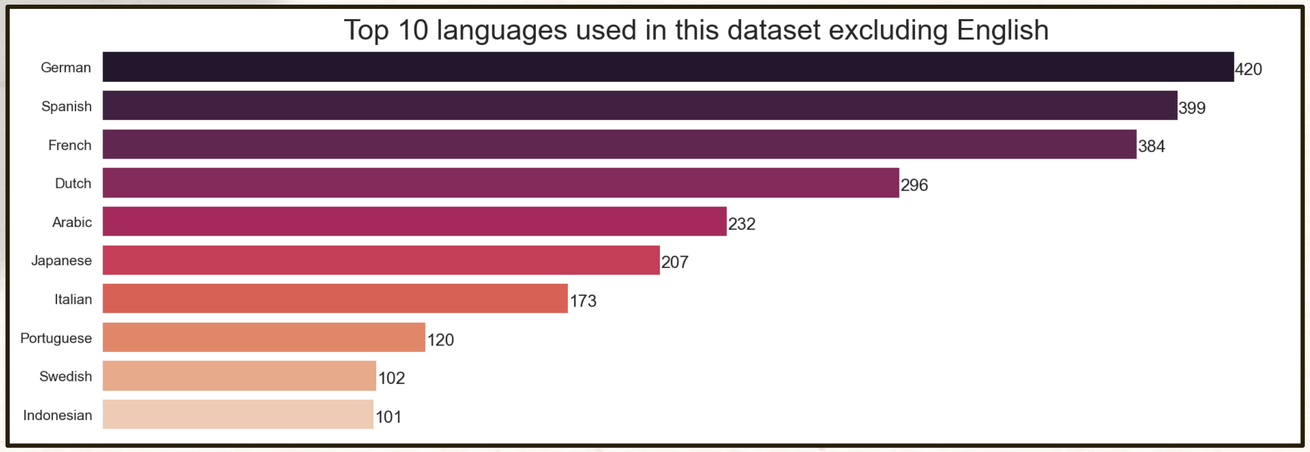


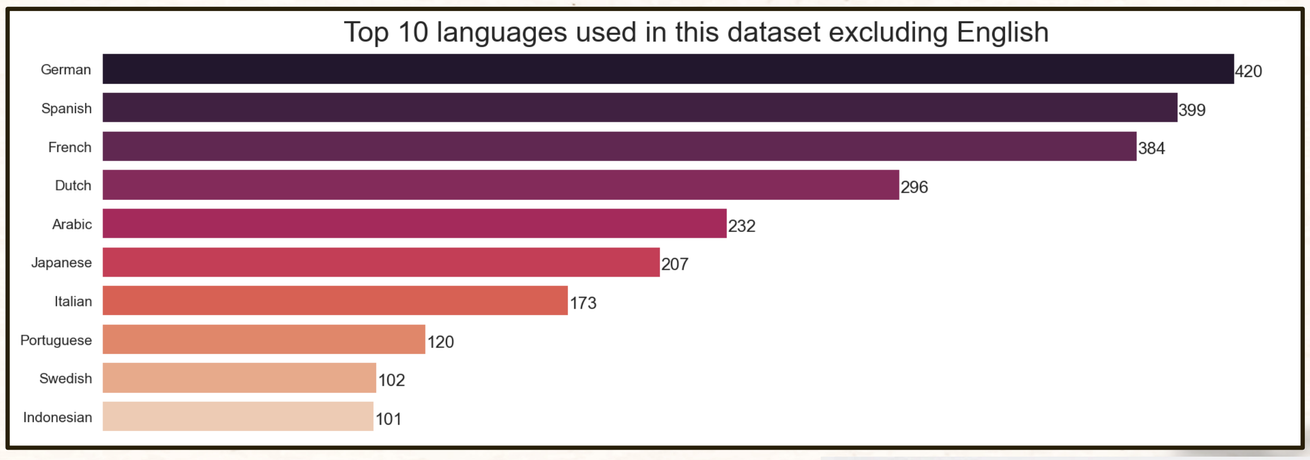
We can see that on average the top 10 authors who write the most tend to have written an average of 25,000 pages

Part 2: Data Exploration

Languages

Exploring the Features: Authors, Languages, Titles, and Genres





Interesting Observation

There are a total of 63 unique languages in this dataset

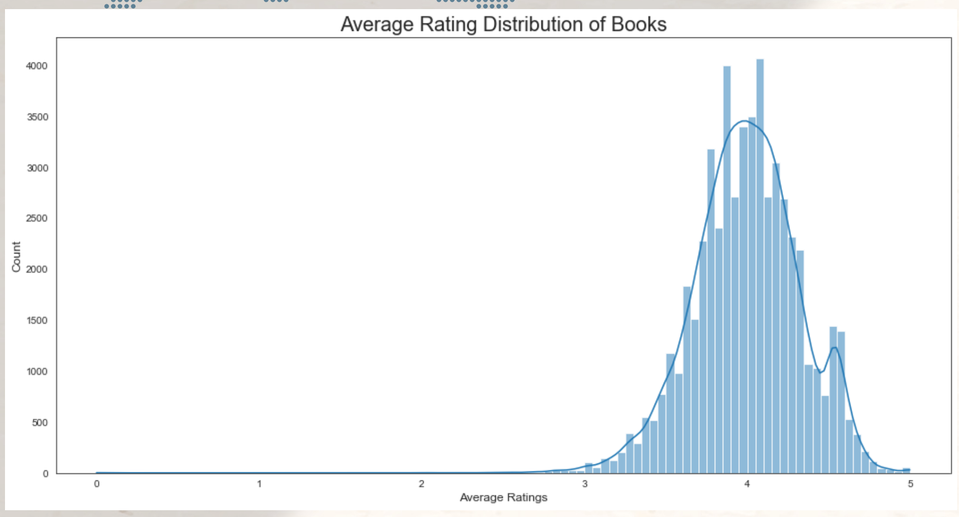
93.98% of the books in this dataset are written in English

Indonesian and Japanese made it to the top 10 list of languages, missing are the languages Chinese, Hindi, and Russian. One hypothesis is that this could be a reflection of the users of the site rather then an actual global consensus.

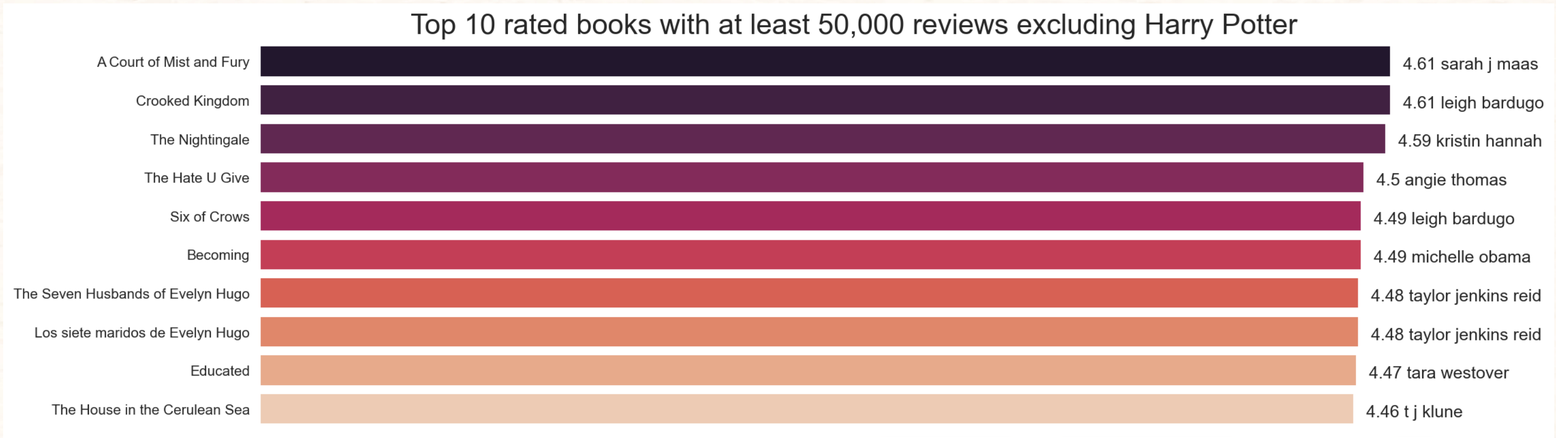
Part 2: Data Exploration

Titles

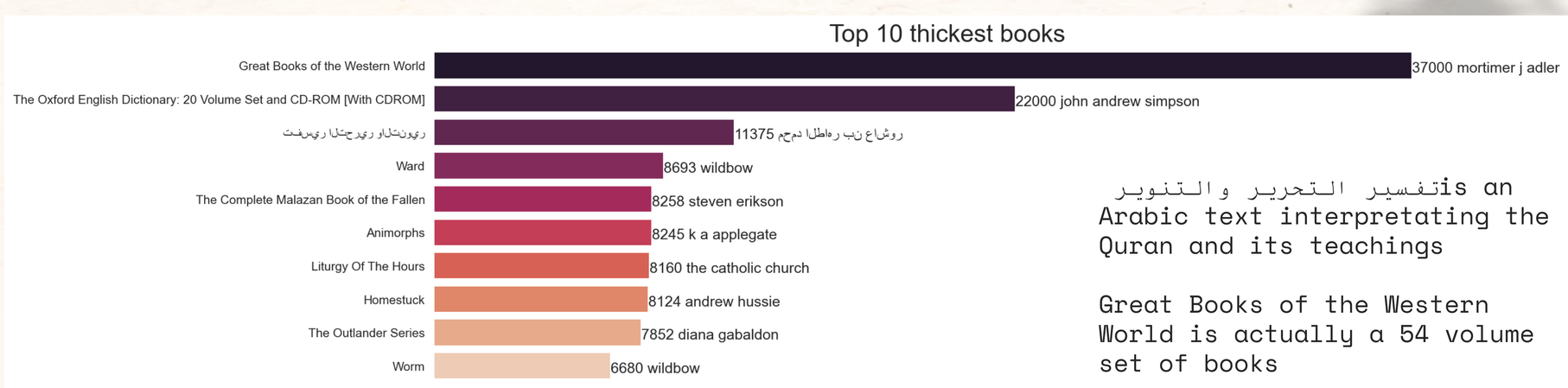
Exploring the Features: Authors, Languages, Titles, and Genres



As all these books made it into a list of books readers would recommend for one reason or the other, The ratings are clustered around the scale 4



The Harry Potter series dominates the top 10 highest average rating with 6 out of 10 being from that series.



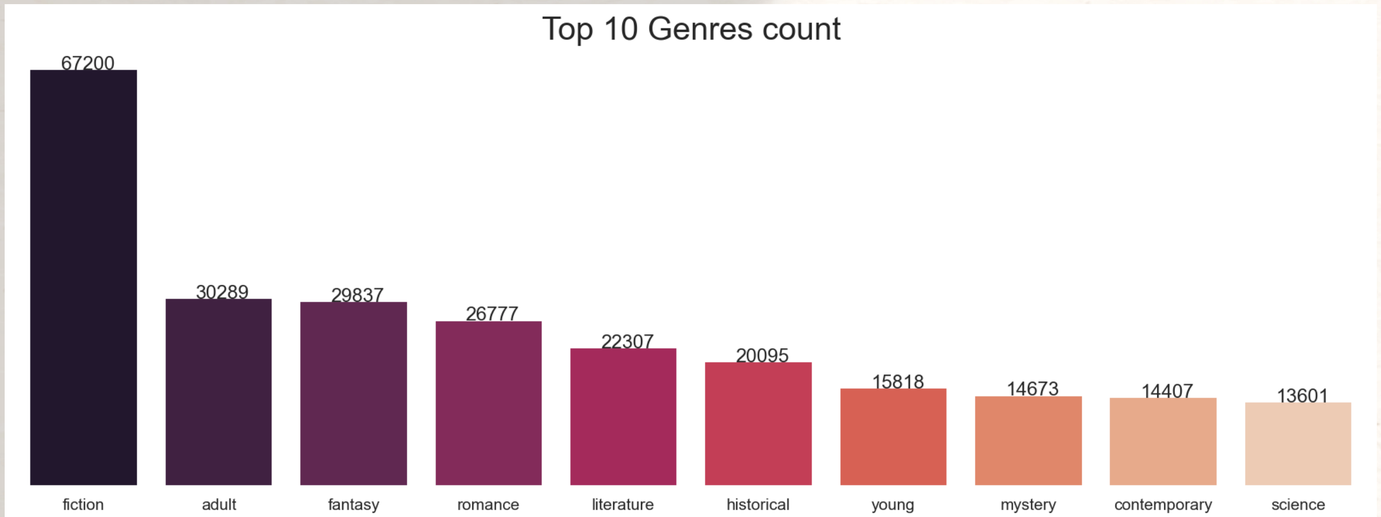
تفسير التحرير والتنوير is an Arabic text interpretating the Quran and its teachings

Great Books of the Western World is actually a 54 volume set of books

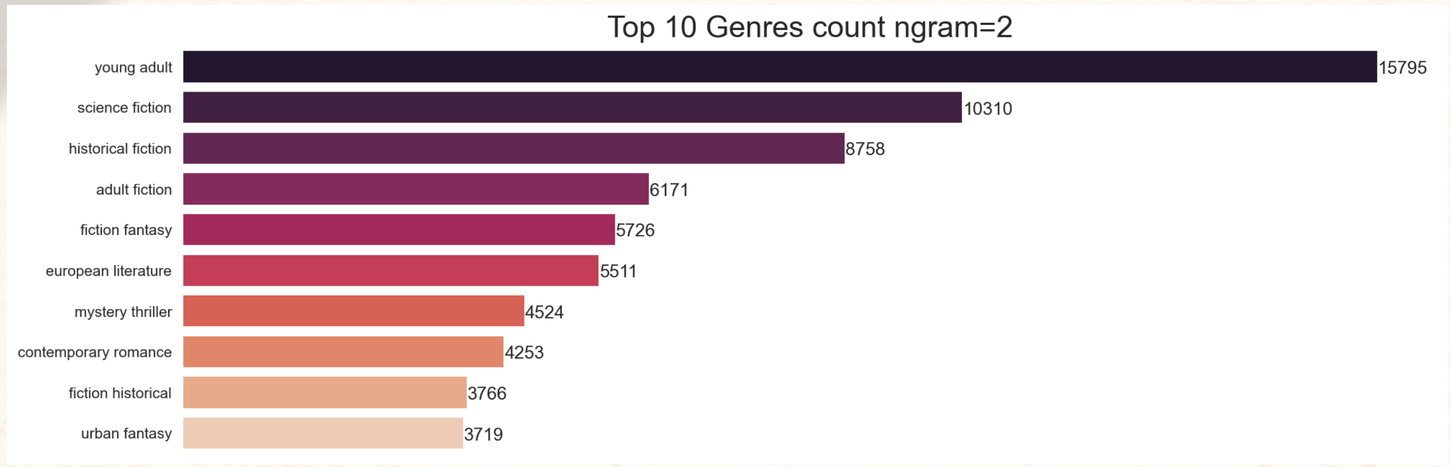
Part 2: Data Exploration

Genres

Exploring the Features: Authors, Languages, Titles, and Genres



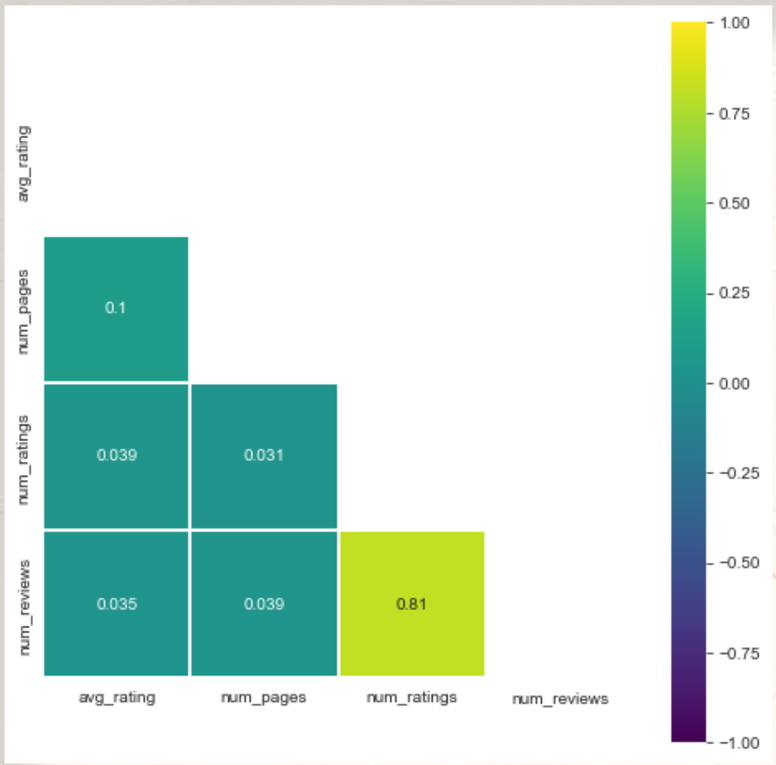
Fiction of all types dominate the genres



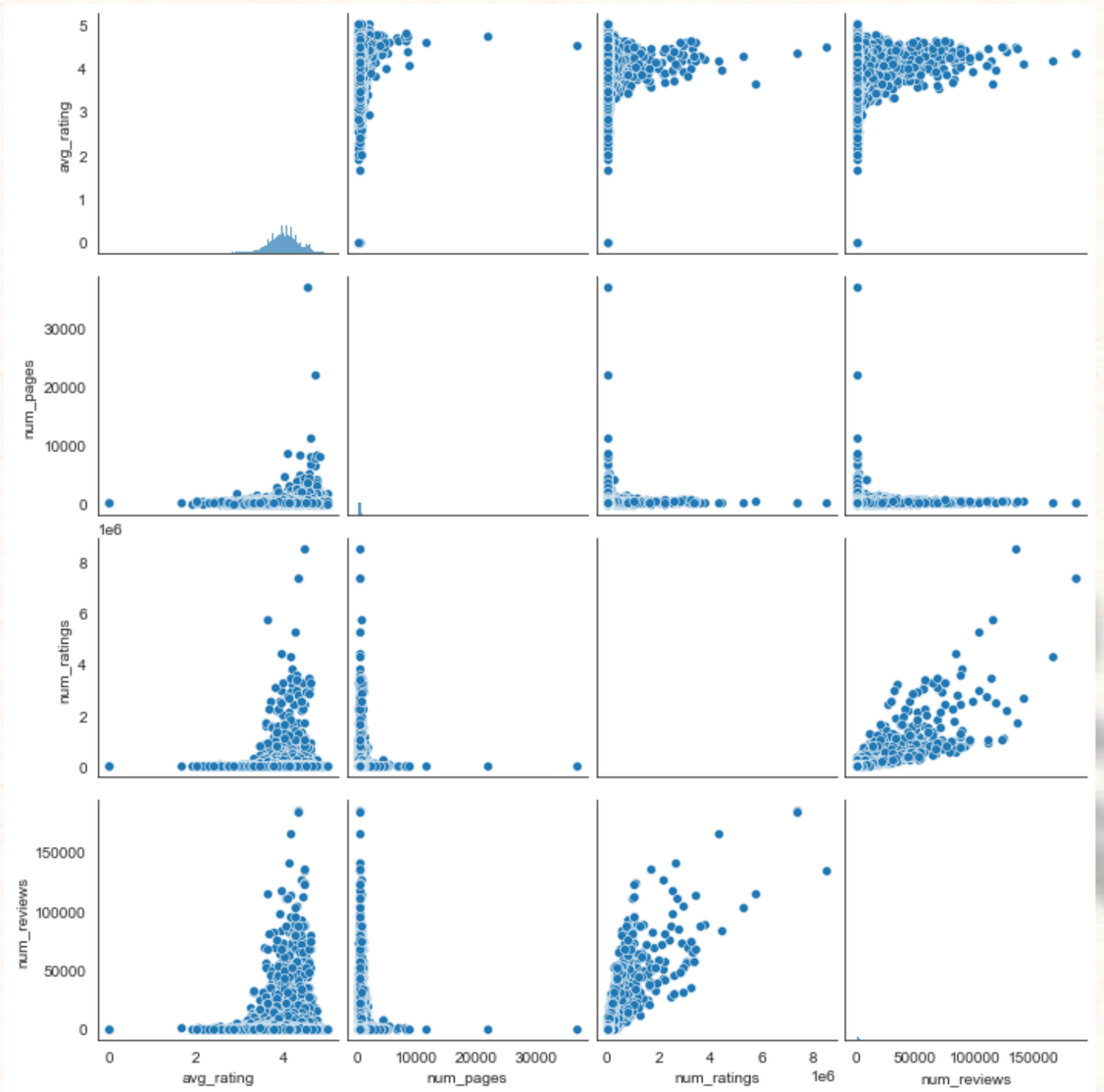
Part 2: Data Exploration

Bonus

Looking for correlation in the quantitative data



Nothing of significance stands out. Except for the correlation of the number of ratings to the number of reviews. But this is expected since a person that review the book would most probably give it a rating too.



Part 3: Recommender System

Cold Start

Without past users ratings and information how do we recommend similar books to a reader base on just the title?

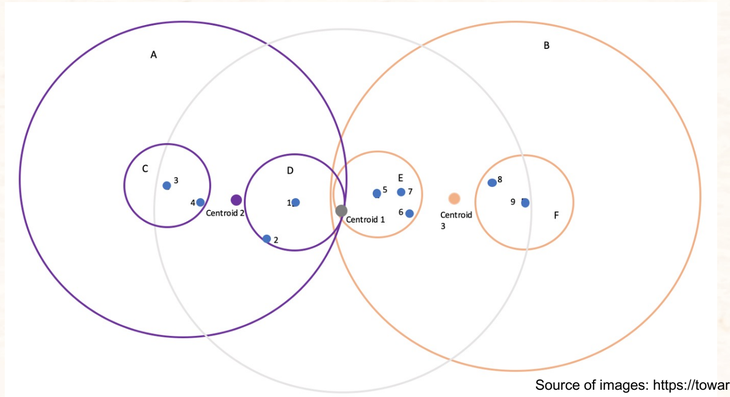
MinMaxScaler( )

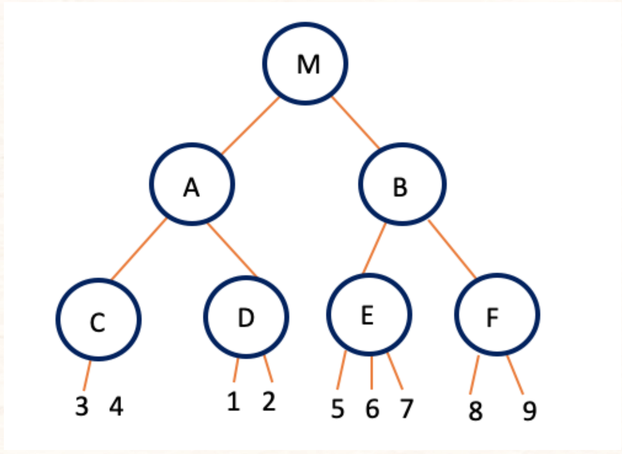
This scales all data in the range of 0-1. Normalizing the data to prevent bias during fitting.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Ball Tree Algorithm

Using Nearest Neighbour to generate a tree like structure by calculating the distance between the different points.





Source of images: https://towardsdatascience.com/tree-algorithms-explained-ball-tree-algorithm-vs-kd-tree-vs-brute-force-9746debcd940

Method 1a: Ball Tree with different Features

Using the features:
1. Avg Rating
2. Language
3. Number of Reviews
4. Number of Pages

```
# Testing the recommender system
model_1 = pickle.load(open("./models/ball_tree_1","rb"))
recommendations = scf.ball_tree_recommender("Harry Potter And The Deathly Hallows",df=df,id_list=model_1[1])
```

```
Book Recommendations:
1 Six Of Crows
  Author: Leigh Bardugo

2 The Lightning Thief
  Author: Rick Riordan

3 The Girl With The Dragon Tattoo
  Author: Stieg Larsson

4 Wonder
  Author: R J Palacio

5 The Giver
  Author: Lois Lowry
```

Method 1b: Ball Tree with Genres alone

```
# Testing the 2nd model
model_2 = pickle.load(open("./models/ball_tree_2","rb"))
recommendations_2 = scf.ball_tree_recommender("Harry Potter And The Deathly Hallows",
```

```
Book Recommendations:
1 Harry Potter Ja Surma Vägised
  Author: J K Rowling

2 Harry Potter Series Box Set
  Author: J K Rowling

3 Harry Potter And The Goblet Of Fire
  Author: J K Rowling

4 Harry Potter And The Order Of The Phoenix
  Author: J K Rowling

5 Harry Potter And The Half-Blood Prince
  Author: J K Rowling
```

Part 3: Recommender System

Cold Start

Without past users ratings and information how do we recommend similar books to a reader base on just the title?

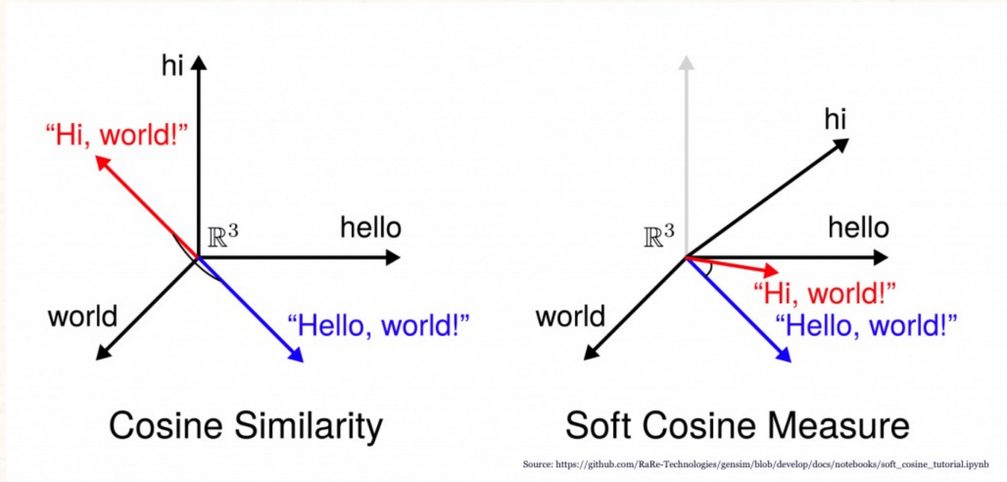
SentenceTransformer( )

“SentenceTransformers is a Python framework for state-of-the-art sentence, text and image embeddings. The initial work is described in our paper Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. You can use this framework to compute sentence / text embeddings for more than 100 languages… the code is tuned to provide the highest possible speed”

- Direct from the documentation
- source link

Cosine Similarity

Finding similar titles regardless of size, by measuring the cosine angle between vectors in a multi dimensional space.



Results using Cosine Similarity

```
# Testing cs model
cs_recommendations = scf.cosine_similarity_recommender("Harry Potter and the Deathly Hallows",
```

```
Book Recommendations:
1 Harry Potter And The Goblet Of Fire
  Author: J k rowling

2 Harry Potter And The Half-Blood Prince
  Author: J k rowling

3 Harry Potter Series Box Set
  Author: J k rowling

4 Harry Potter And The Order Of The Phoenix
  Author: J k rowling

5 Harry Potter Ja Surma Vägised
  Author: J k rowling
```

Combined Recommender System

	Recommendations	Author
1	Six Of Crows	leigh bardugo
2	The Lightning Thief	rick riordan
3	The Girl With The Dragon Tattoo	stieg larsson
4	Wonder	r j palacio
5	The Giver	lois lowry
6	Harry Potter Ja Surma Vägised	j k rowling
7	Harry Potter Series Box Set	j k rowling
8	Harry Potter And The Goblet Of Fire	j k rowling
9	Harry Potter And The Order Of The Phoenix	j k rowling
10	Harry Potter And The Half-Blood Prince	j k rowling

Conclusion and Reflection

Final Thoughts

Conclusion

I've created 3 recommender systems to recommend books in a cold start scenario using content base filtering. And then combining them together to get a unique list of books that is similar by way of distance calculated using cosine similarity and a ball tree classifier. It would seem that the best feature to use would be the genre column in this case.

There is another way, using the multi-armed bandit method, where random books are recommended to the user to get the user feedback. And if the user rates a book positively, the recommender would then generate a new list of recommendations. However, due to time constraint, I shall not be exploring this method.

One recurring issue I constantly face was the lack of memory/ram to run and test different models and ideas. This restricted my ability to run more test and visualize some graphs. One way to work around this would be to work in a cloud environment. But once again due to time constraint and financial reasons I did not take that step. But it would be a good way to further enhance the recommender system.

Reflections

1. Importance of allocating enough time
2. Web crawler
3. Importance of proper data collection
4. Working with Big Data
5. Model Deployment

Thank You