Book Recommendation System

Group 02

Team Member

Lê Hoàng Khang

- Write Proposal
- Search and collect data
- Content-based filtering algorithm

Nguyễn Thanh Hùng

- Writing Presentations
- Data preprocessing
- Collaborative filtering algorithm

Nguyễn Duy Thái

- Write Report
- Search and collect data
- Content-based filtering algorithm

Lê Phước Yên

- Write Milestone
- EDA
- Collaborative filtering algorithm

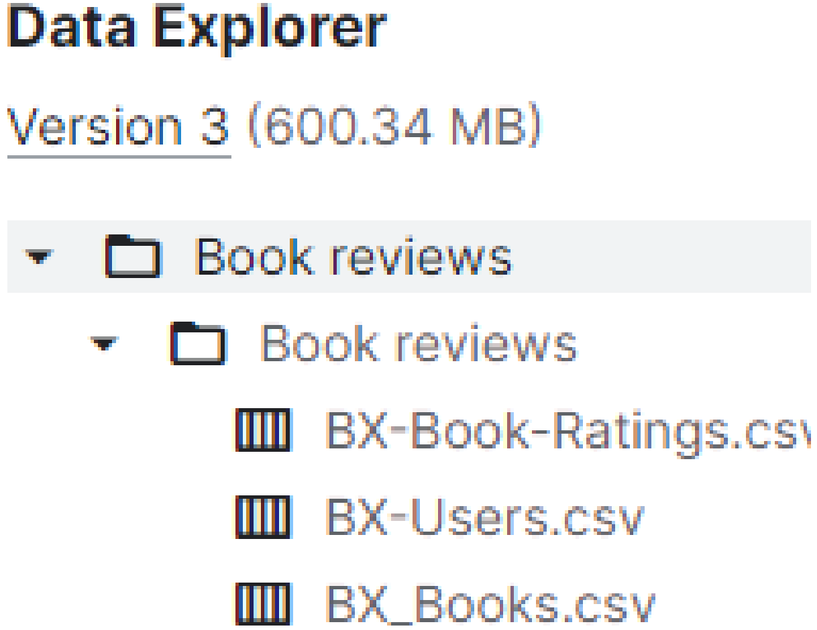
Table Of Content

REASON FOR CHOOSING TOPIC

ABOUT DATASET

- Link: *https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset*
- The data set consists of 3 main files containing about 278,858 users, providing more than 1 million ratings on about 270,379 books:

- BX-Book-Ratings: includes the user's ID, the code of each book, and the user's rating for that book.
- BX-Users: includes user ID, address and age.
- BX-Books: includes each book's code, book title, author, year of publication, publisher, summary and link to the book's photo.



EDA

Join 3 tables

```
1  df = ratings.join(users, on='User-ID')
2  df = df.join(books, on='ISBN')
```

✓ 0.0s

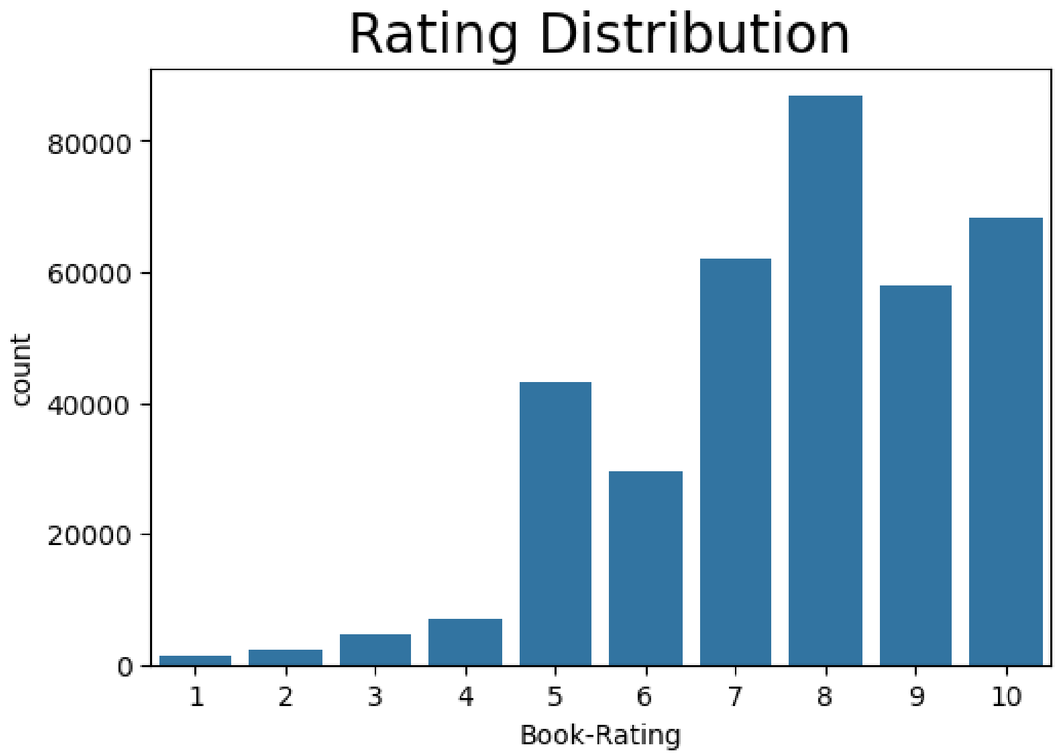
```
1  df.count()
```

✓ 4.7s

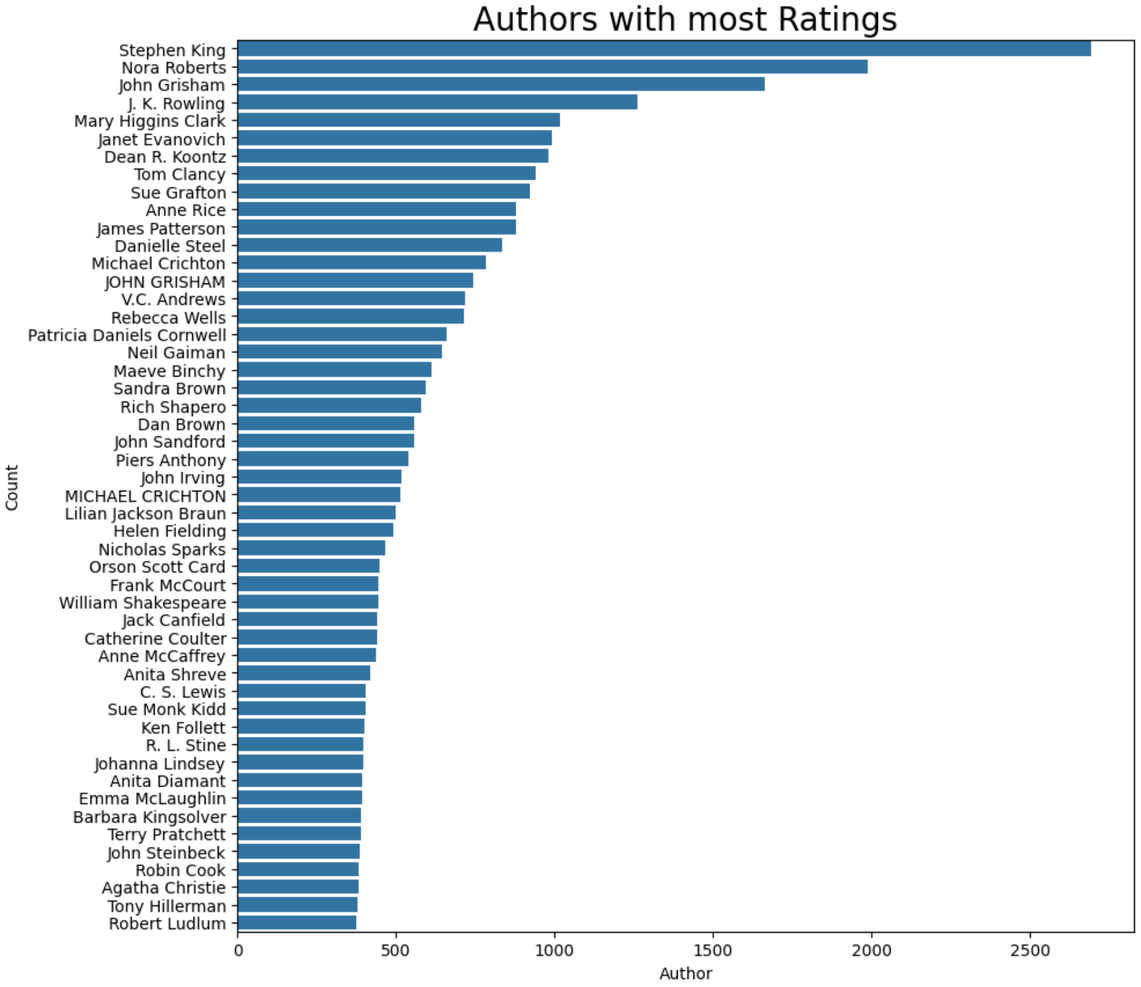
```
632238
```

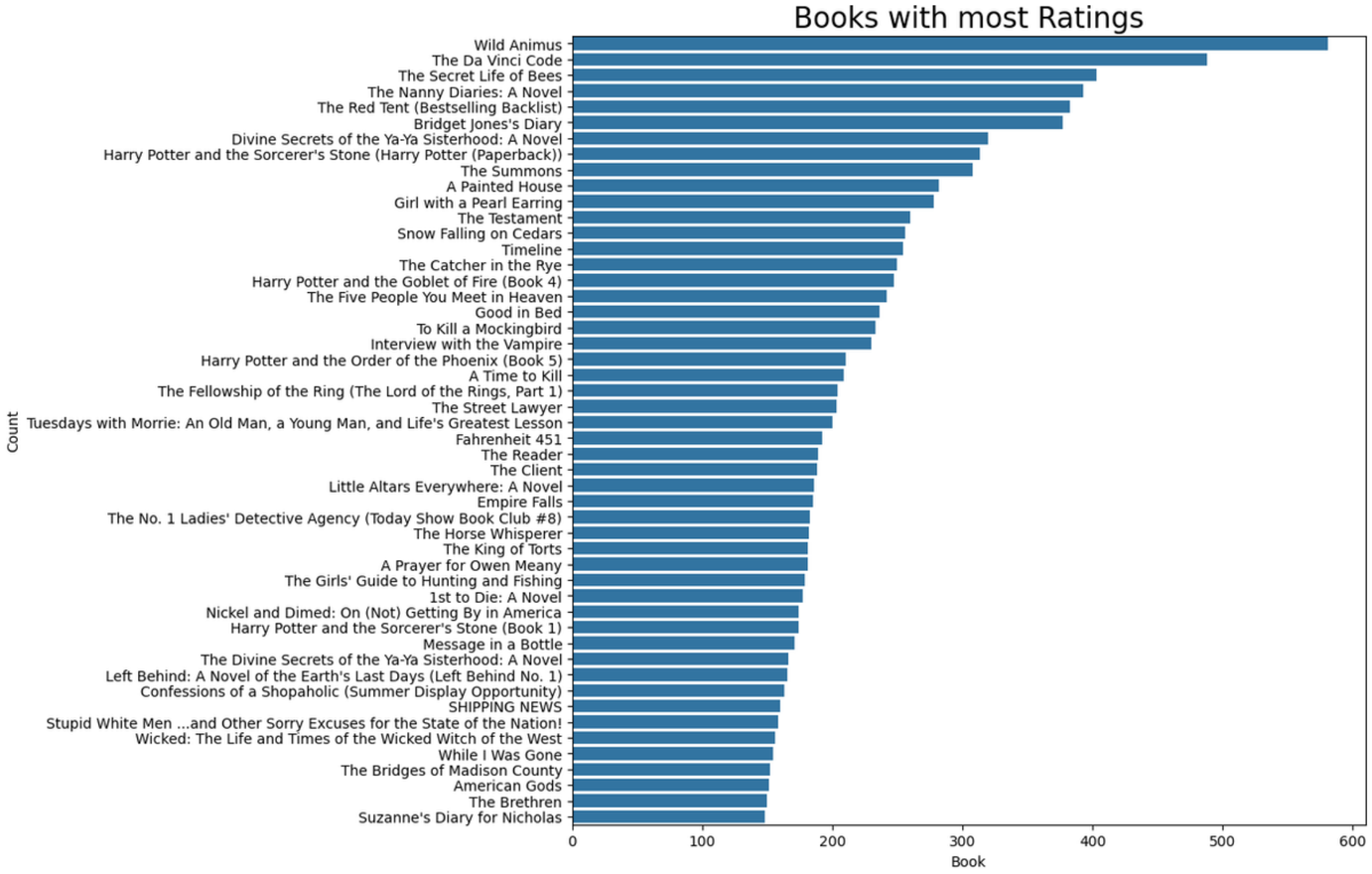
Convert data types

```
df = df.withColumn("Age", df["Age"].cast(IntegerType()))
df = df.withColumn("Year-Of-Publication",
                   df["Year-Of-Publication"].cast(IntegerType()))
df = df.withColumn("Book-Rating", df["Book-Rating"].cast(IntegerType()))
df = df.withColumn("User-ID", df["User-ID"].cast(IntegerType()))
```

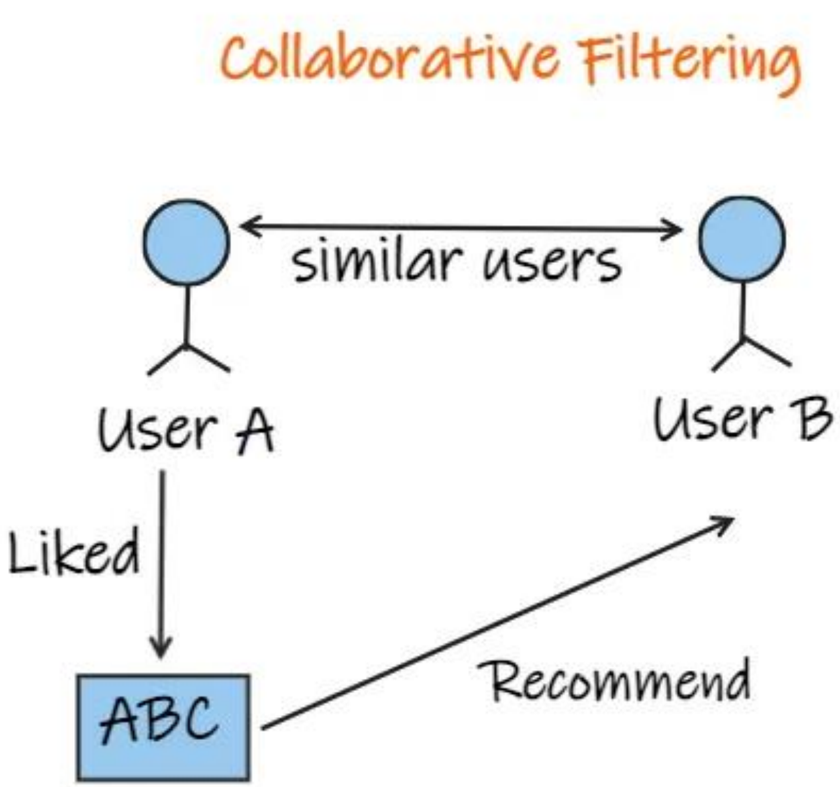


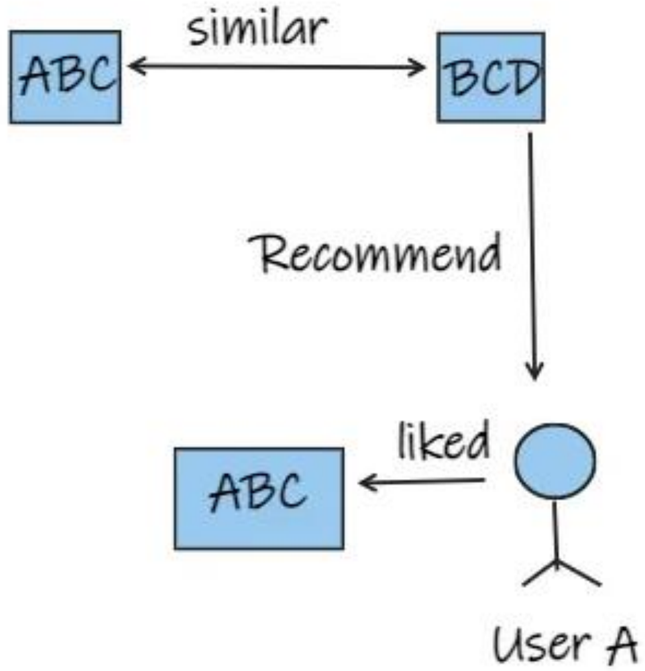
EDA


Authors with most Ratings
Stephen King
Nora Roberts
John Grisham
J. K. Rowling
Mary Higgins Clark
Janet Evanovich
Dean R. Koontz
Tom Clancy
Sue Grafton
Anne Rice
James Patterson
Danielle Steel
Michael Crichton
JOHN GRISHAM
V.C. Andrews
Rebecca Wells
Patricia Daniels Cornwell
Neil Gaiman
Maeve Binchy
Sandra Brown
Rich Shapero
Dan Brown
John Sandford
Piers Anthony
John Irving
MICHAEL CRICHTON
Lilian Jackson Braun
Helen Fielding
Nicholas Sparks
Orson Scott Card
Frank McCourt
William Shakespeare
Jack Canfield
Catherine Coulter
Anne McCaffrey
Anita Shreve
C. S. Lewis
Sue Monk Kidd
Ken Follett
R. L. Stine
Johanna Lindsey
Anita Diamant
Emma McLaughlin
Barbara Kingsolver
Terry Pratchett
John Steinbeck
Robin Cook
Agatha Christie
Tony Hillerman
Robert Ludlum
Count
0
500
1000
1500
2000
2500
Author


Books with most Ratings
Wild Animus
The Da Vinci Code
The Secret Life of Bees
The Nanny Diaries: A Novel
The Red Tent (Bestselling Backlist)
Bridget Jones's Diary
Divine Secrets of the Ya-Ya Sisterhood: A Novel
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
The Summons
A Painted House
Girl with a Pearl Earring
The Testament
Snow Falling on Cedars
Timeline
The Catcher in the Rye
Harry Potter and the Goblet of Fire (Book 4)
The Five People You Meet in Heaven
Good in Bed
To Kill a Mockingbird
Interview with the Vampire
Harry Potter and the Order of the Phoenix (Book 5)
A Time to Kill
The Fellowship of the Ring (The Lord of the Rings, Part 1)
The Street Lawyer
Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson
Fahrenheit 451
The Reader
The Client
Little Altars Everywhere: A Novel
Empire Falls
The No. 1 Ladies' Detective Agency (Today Show Book Club #8)
The Horse Whisperer
The King of Torts
A Prayer for Owen Meany
The Girls' Guide to Hunting and Fishing
1st to Die: A Novel
Nickel and Dimed: On (Not) Getting By in America
Harry Potter and the Sorcerer's Stone (Book 1)
Message in a Bottle
The Divine Secrets of the Ya-Ya Sisterhood: A Novel
Left Behind: A Novel of the Earth's Last Days (Left Behind No. 1)
Confessions of a Shopaholic (Summer Display Opportunity)
SHIPPING NEWS
Stupid White Men ...and Other Sorry Excuses for the State of the Nation!
Wicked: The Life and Times of the Wicked Witch of the West
While I Was Gone
The Bridges of Madison County
American Gods
The Brethren
Suzanne's Diary for Nicholas
Count
0
100
200
300
400
500
600
Book

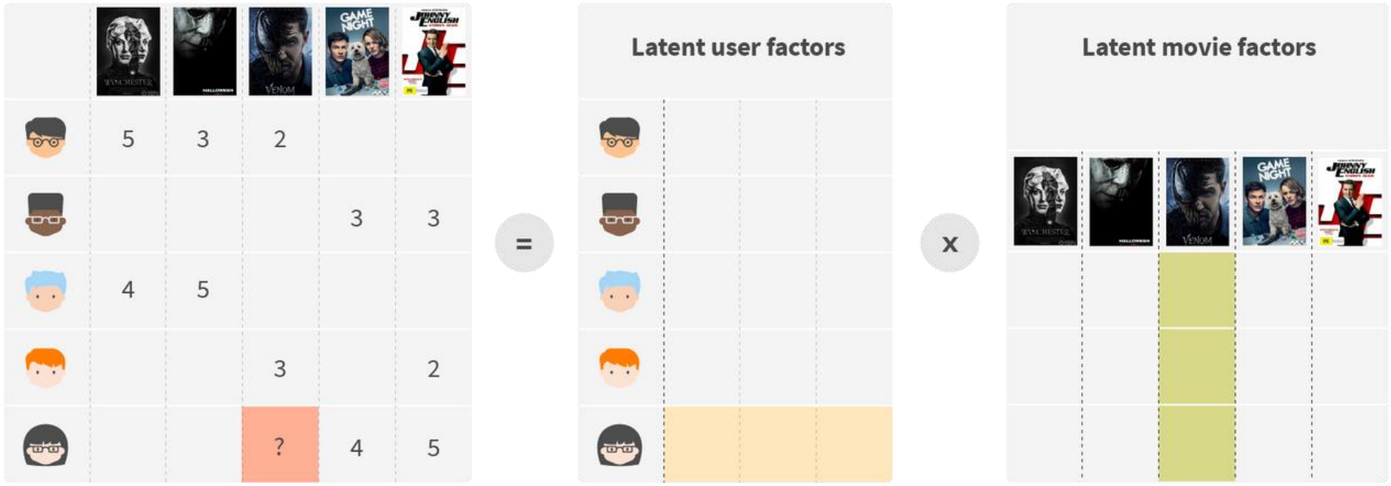
MODEL DEVELOPMENT





MODEL DEVELOPMENT

COLLABORATIVE FILTERING USING ALS



$$P_{nxm} = U_{nxk} \times V_{kxm}$$

Parameter	Description	Default value
rank	Số lượng latent factors	10
regParam	Regularization parameter	1
maxIters	Số lần lặp tối đa	10

MODEL DEVELOPMENT

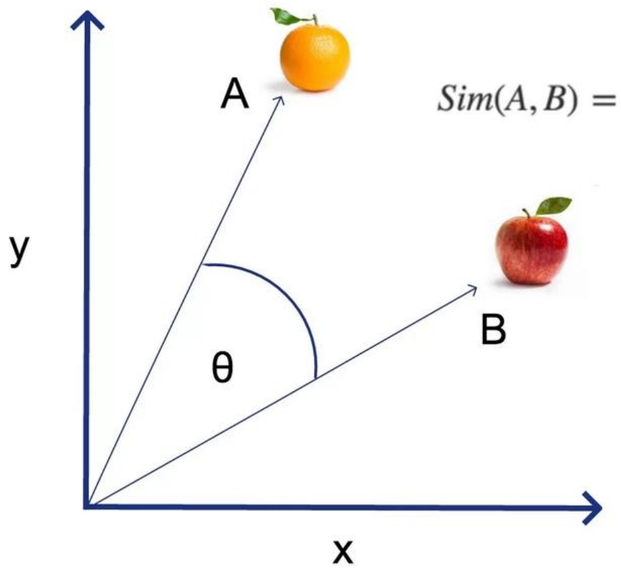
CONTENT-BASED USING TF-IDF AND COSIN SIMILARITY

$$Term\ Frequency\ TF(t,d) = \frac{count\ of\ t\ in\ d}{number\ of\ words\ in\ d}$$

$$Inverse\ Document\ Frequency = log_2\left(\frac{number\ of\ documents\ N}{number\ of\ documents\ containing\ the\ term\ t}\right)$$

$$tf - idf(d,t) = tf(t,d) * idf(t)$$

Cosine Similarity



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Experiment and Evaluate

Collaborative Filtering

```
[42]
1  for_one_user = predictions.filter(col("User-ID") == 98391).select(
2      "User-ID", "Book-Title", "Image-URL-M", "Book-Rating", "prediction"
3  )
4  for_one_user.show(5, False)
```

Python

```
+-------+-----------------------------------------------------+-----------------------------------------------------------------------+-----------+----------+
|User-ID|Book-Title                                           |Image-URL-M                                                            |Book-Rating|prediction|
+-------+-----------------------------------------------------+-----------------------------------------------------------------------+-----------+----------+
|98391  |Flirting With Trouble (Harlequin American Romance Series)|http://images.amazon.com/images/P/037375020X.01.MZZZZZZZ.jpg|8          |3.8005092 |
|98391  |Anything Goes (Harlequin Blaze, 112)                 |http://images.amazon.com/images/P/037379116X.01.MZZZZZZZ.jpg|9          |7.2175493 |
|98391  |The Rake: Lessons in Love                            |http://images.amazon.com/images/P/038082082X.01.MZZZZZZZ.jpg|8          |7.8791027 |
|98391  |Key of Light (Key Trilogy (Paperback))               |http://images.amazon.com/images/P/051513628X.01.MZZZZZZZ.jpg|10         |7.6027946 |
|98391  |The Reluctant Suitor                                 |http://images.amazon.com/images/P/0060185708.01.MZZZZZZZ.jpg|9          |8.014934  |
+-------+-----------------------------------------------------+-----------------------------------------------------------------------+-----------+----------+
only showing top 5 rows
```

RMSE value 2.21967872716262

Experiment and Evaluate

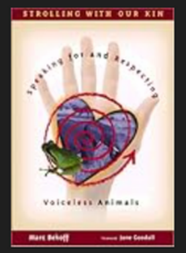
Content-Based Filtering

Books similar to keyword - wild animal

ISBN	similarity_score	Book-Title	Book-Author	Year-Of-Publication	Publisher
059031971X	0.6469376166083098	The Gentle Jungle	Toni R. Helfer	1981	Scholastic Pa
1564585506	0.595001952867062	Big Pictures: Wild Animals	Mary Ling	1994	Dorling Kinde
0440159415	0.5845439562587951	My Wild World	Joan Embery	1981	Bantam Books
1881699021	0.5478580272597805	Strolling with Our Kin: Speaking for and Respecting Voiceless Animals	Marc Bekoff	2000	Amer Anti-Viv
0486296547	0.5397765730533401	Riddles, Riddles, Riddles (Dover Game & Puzzle Activity Books)	Darwin A. Hindman	1997	Dover Publica

only showing top 5 rows

Strolling with Our Kin: Speaking for and Respecting Voiceless Animals

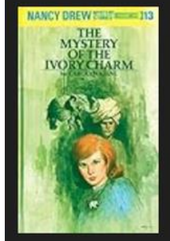


A look at animal protection and compassion, using animals as fur and food, animal pain and suffering, and dissection and vivisection.

Big Pictures: Wild Animals

Each photograph of a wild mammal or bird is accompanied by a question and answer offering information about the animal or its behavior

The Mystery of the Ivory Charm (Nancy Drew Mystery Stories, No 13)



Nancy Drew determines whether an ivory elephant charm really protects its wearer from harm when she investigates the involvement of a member of the Bengleton Wild-Animal Show in a mysterious scheme.

Conclusion and development direction

Conclusion

- Through the process of implementing the project, the group learned more knowledge about recommendation models using PySpark for book suggestions.
- Overall, the topic has completed the basic level, providing suggestions based on user preferences and entered keywords.

Development direction

- Perform suggestions with large amounts of data
- Combine interfaces to be more intuitive
- Process automation

Thanks