Connecting the dots: Reader ratings, bibliographic data, and machine-learning algorithms for monograph selection

Jingshan Xiao, Associate Director, Library Digital Services
University of Houston Clear Lake

Wenli Gao, Data Services Librarian
University of Houston

Big Data & Recommender Systems

“Small” data

Culmination of data

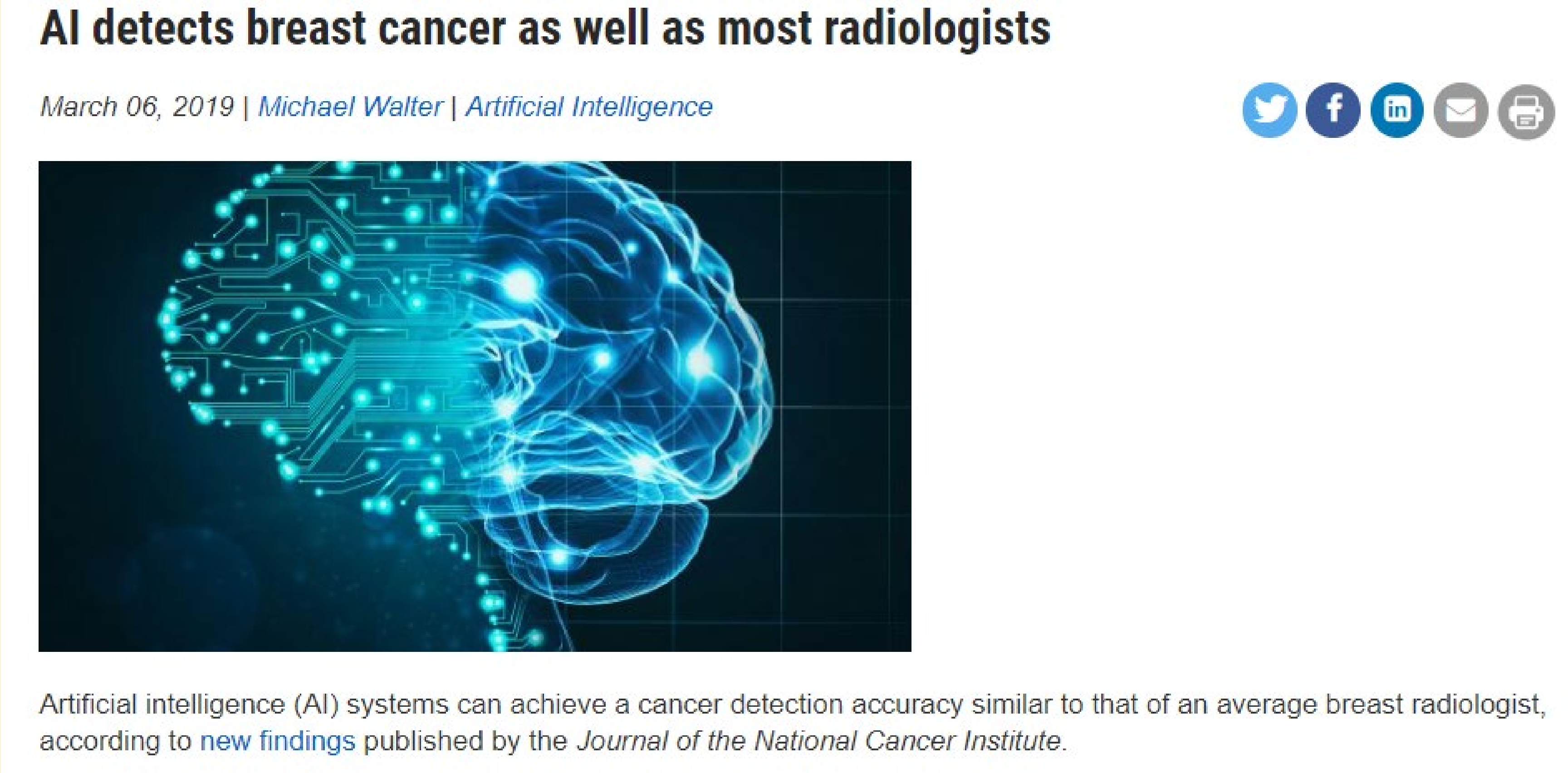
Big data

Ys1
Zs1
alfa
n_dv
delta_alfa
n_sk
delta_t
Yo
Zo
Yg
Zg
EG
tau
EF
FG
beta
ksi
Ye
Ze

Library big data

- Catalogue data
- Process/transactional data

Machine learning



AI detects breast cancer as well as most radiologists

March 06, 2019 | *Michael Walter* | *Artificial Intelligence*

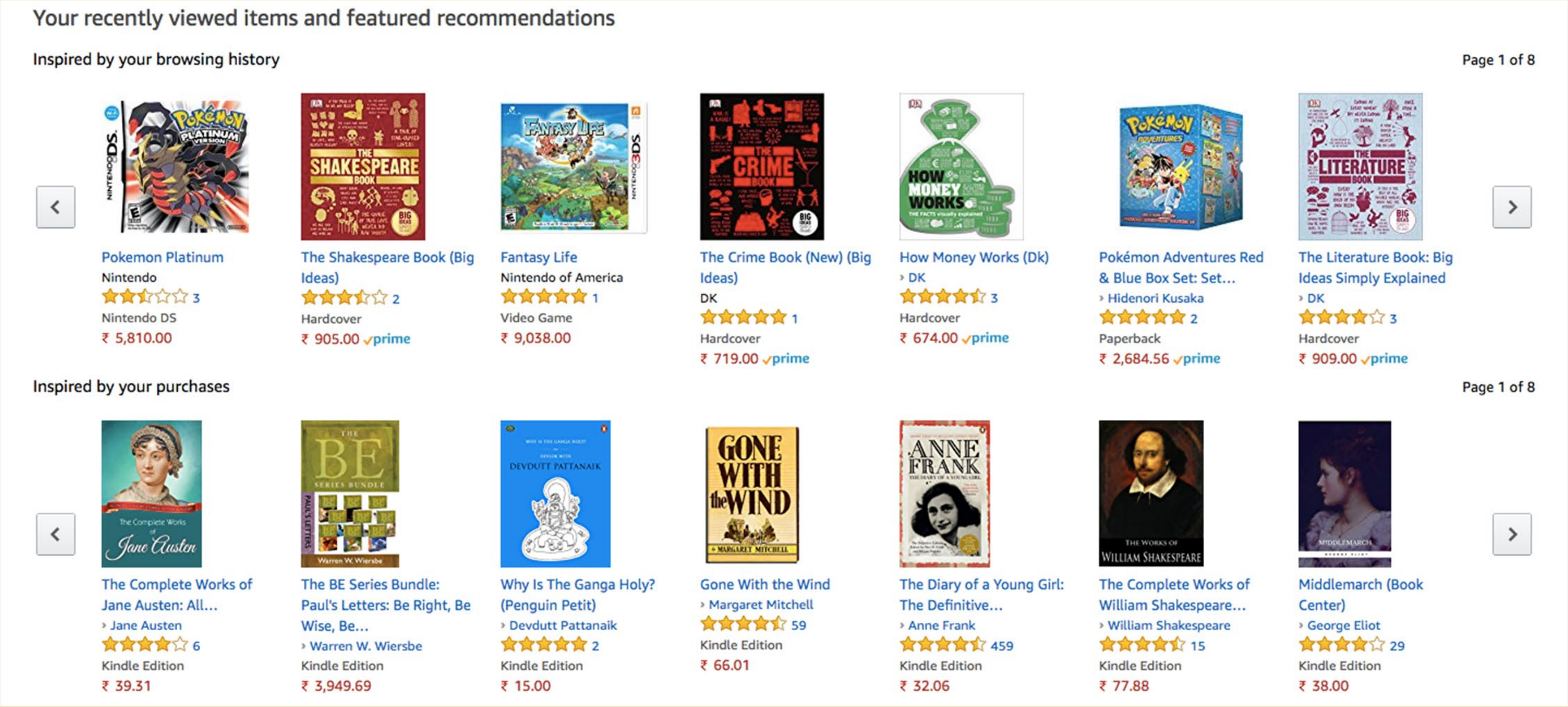
Artificial intelligence (AI) systems can achieve a cancer detection accuracy similar to that of an average breast radiologist, according to new findings published by the *Journal of the National Cancer Institute*.

Recommender systems

CUSTOMERS WHO BOUGHT THIS ITEM:
ALSO BOUGHT:
piccolo

Use Cases

Commercial use



Your recently viewed items and featured recommendations

Inspired by your browsing history

Page 1 of 8

Pokemon Platinum
Nintendo
3
Nintendo DS
₹ 5,810.00

The Shakespeare Book (Big Ideas)
2
Hardcover
₹ 905.00 ✓prime

Fantasy Life
Nintendo of America
1
Video Game
₹ 9,038.00

The Crime Book (New) (Big Ideas)
DK
1
Hardcover
₹ 719.00 ✓prime

How Money Works (Dk)
› DK
3
Hardcover
₹ 674.00 ✓prime

Pokémon Adventures Red & Blue Box Set: Set...
› Hidenori Kusaka
2
Paperback
₹ 2,684.56 ✓prime

The Literature Book: Big Ideas Simply Explained
› DK
3
Hardcover
₹ 909.00 ✓prime

Inspired by your purchases

Page 1 of 8

The Complete Works of Jane Austen: All...
› Jane Austen
6
Kindle Edition
₹ 39.31

The BE Series Bundle: Paul's Letters: Be Right, Be Wise, Be...
› Warren W. Wiersbe
Kindle Edition
₹ 3,949.69

Why Is The Ganga Holy? (Penguin Petit)
› Devdutt Pattanaik
2
Kindle Edition
₹ 15.00

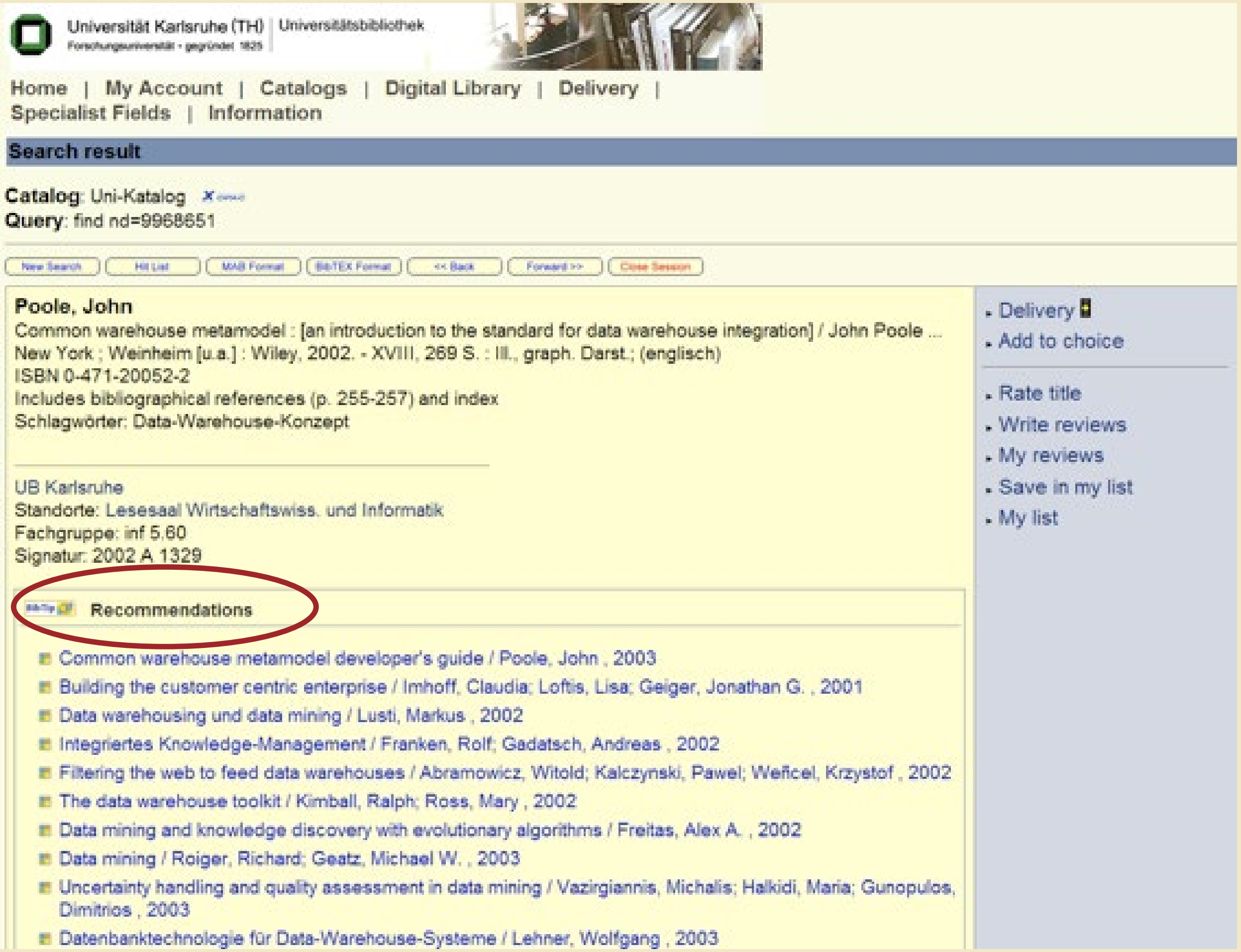
Gone With the Wind
› Margaret Mitchell
59
Kindle Edition
₹ 66.01

The Diary of a Young Girl: The Definitive...
› Anne Frank
459
Kindle Edition
₹ 32.06

The Complete Works of William Shakespeare...
› William Shakespeare
15
Kindle Edition
₹ 77.88

Middlemarch (Book Center)
› George Eliot
29
Kindle Edition
₹ 38.00

Library use-Catalogue



Search result

Catalog: Uni-Katalog

Query: find nd=9968651

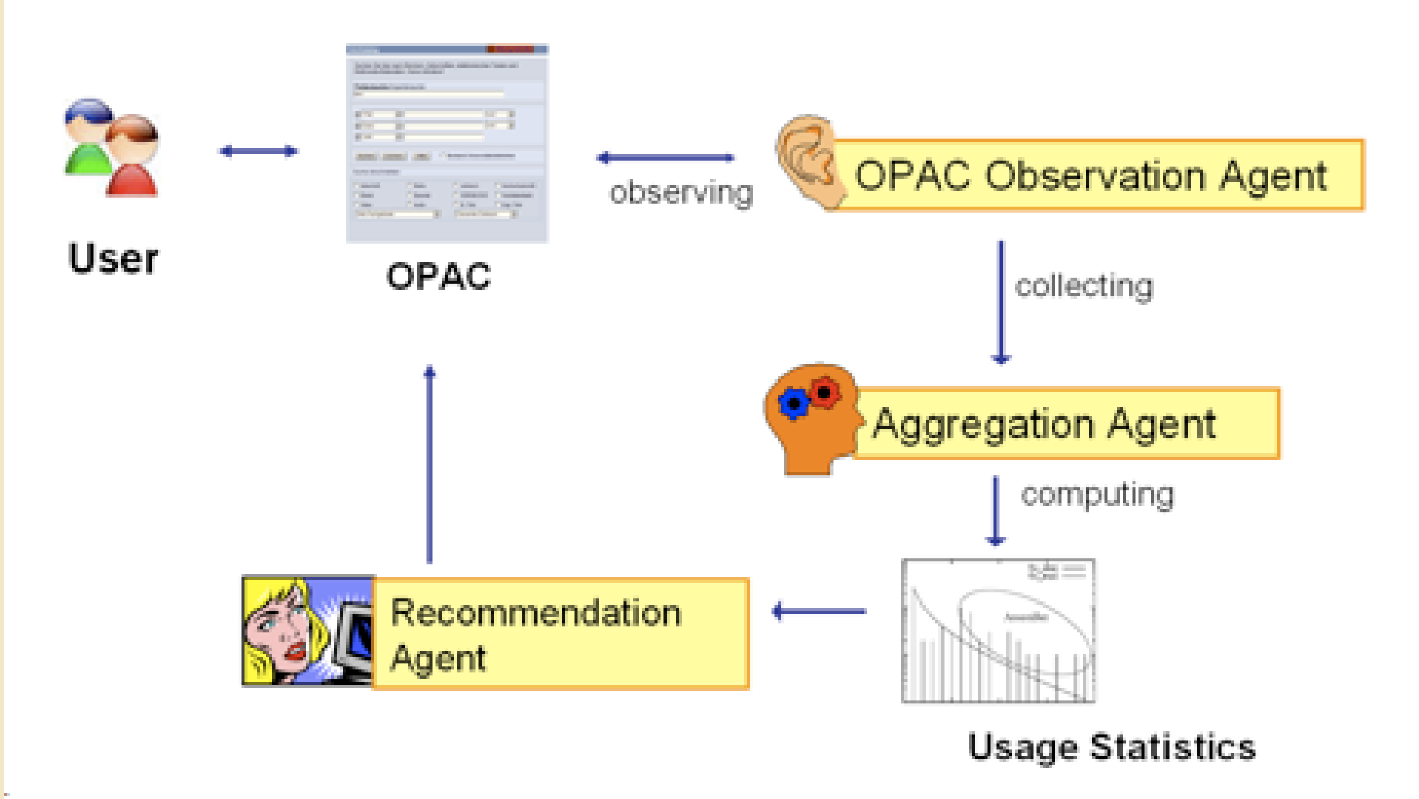
Poole, John
Common warehouse metamodel : [an introduction to the standard for data warehouse integration] / John Poole ...
New York ; Weinheim [u.a.] : Wiley, 2002. - XVIII, 269 S. : Ill., graph. Darst.; (englisch)
ISBN 0-471-20052-2
Includes bibliographical references (p. 255-257) and index
Schlagwörter: Data-Warehouse-Konzept

UB Karlsruhe
Standorte: Lesesaal Wirtschaftswiss. und Informatik
Fachgruppe: inf 5.60
Signatur: 2002 A 1329

Recommendations

- Common warehouse metamodel developer's guide / Poole, John , 2003
- Building the customer centric enterprise / Imhoff, Claudia; Loftis, Lisa; Geiger, Jonathan G. , 2001
- Data warehousing und data mining / Lusti, Markus , 2002
- Integriertes Knowledge-Management / Franken, Rolf; Gadatsch, Andreas , 2002
- Filtering the web to feed data warehouses / Abramowicz, Witold; Kalczynski, Pawel; Weñcel, Krzystof , 2002
- The data warehouse toolkit / Kimball, Ralph; Ross, Mary , 2002
- Data mining and knowledge discovery with evolutionary algorithms / Freitas, Alex A. , 2002
- Data mining / Roiger, Richard; Geatz, Michael W. , 2003
- Uncertainty handling and quality assessment in data mining / Vazirgiannis, Michalis; Halkidi, Maria; Gunopulos, Dimitrios , 2003
- Datenbanktechnologie für Data-Warehouse-Systeme / Lehner, Wolfgang , 2003

Recommender workflow from BibTip



User

OPAC

observing

OPAC Observation Agent

collecting

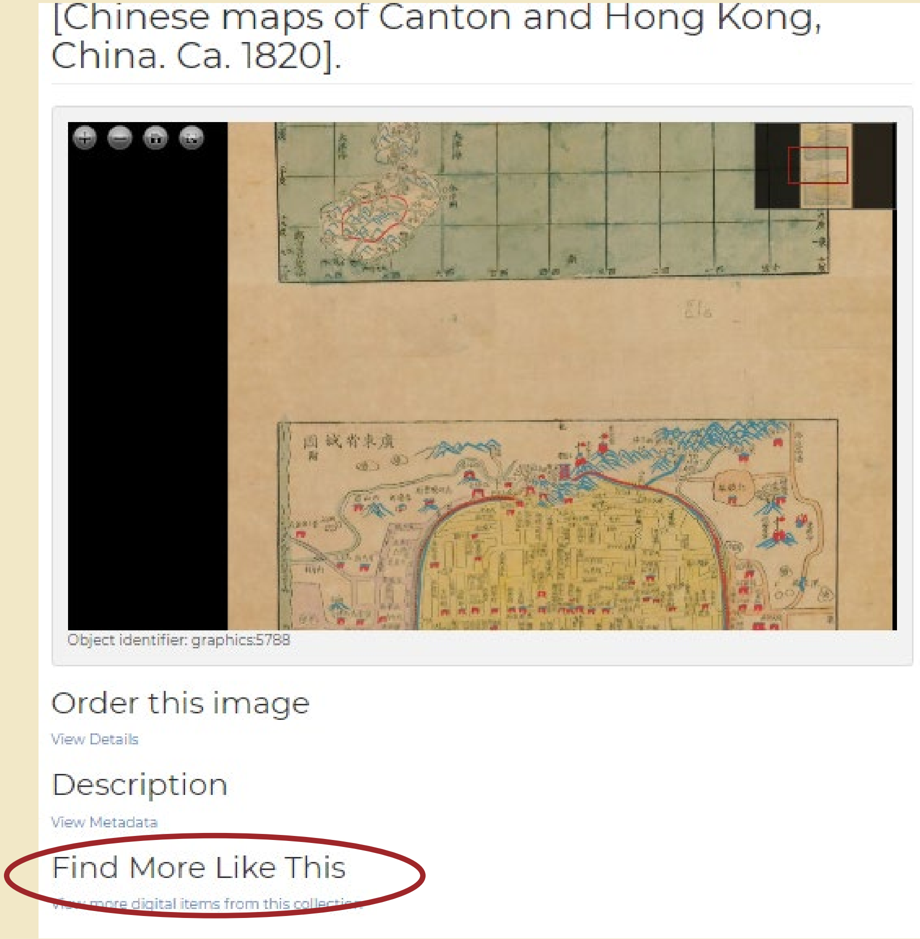
Aggregation Agent

computing

Usage Statistics

Recommendation Agent

Library use–Special collections



[Chinese maps of Canton and Hong Kong, China. Ca. 1820].

Object identifier: graphics:5788

Order this image

View Details

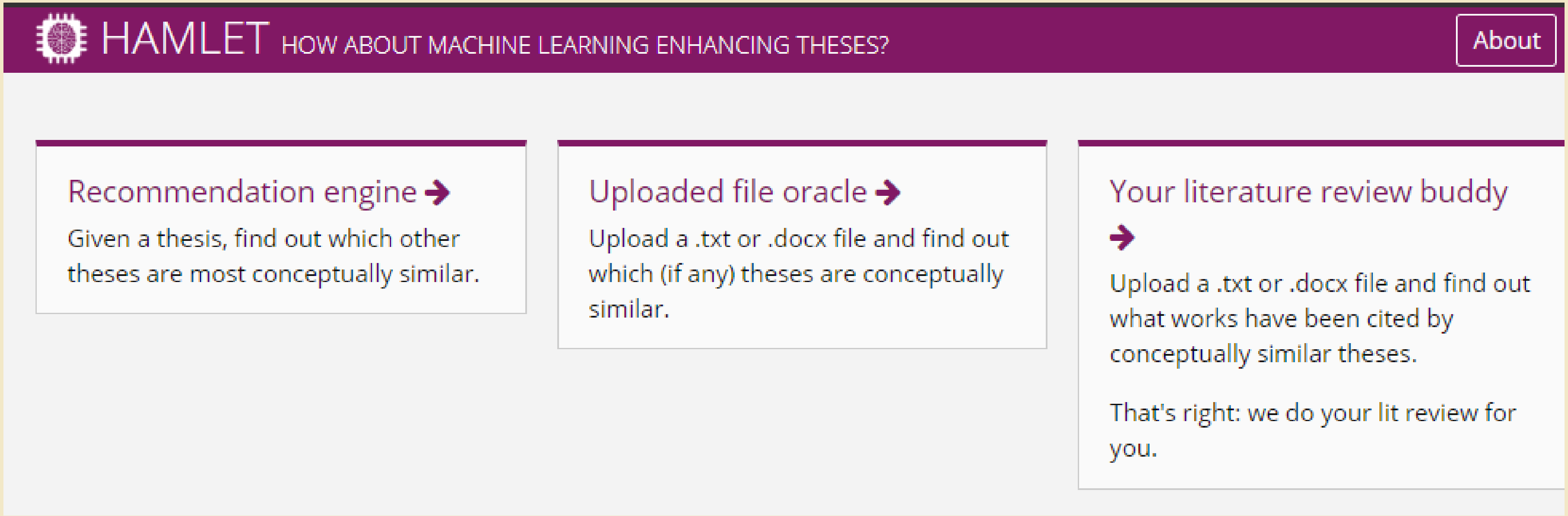
Description

View Metadata

Find More Like This

View more digital items from this collection

Library use–Theses collection



HAMLET HOW ABOUT MACHINE LEARNING ENHANCING THESES?

Recommendation engine ➔

Given a thesis, find out which other theses are most conceptually similar.

Uploaded file oracle ➔

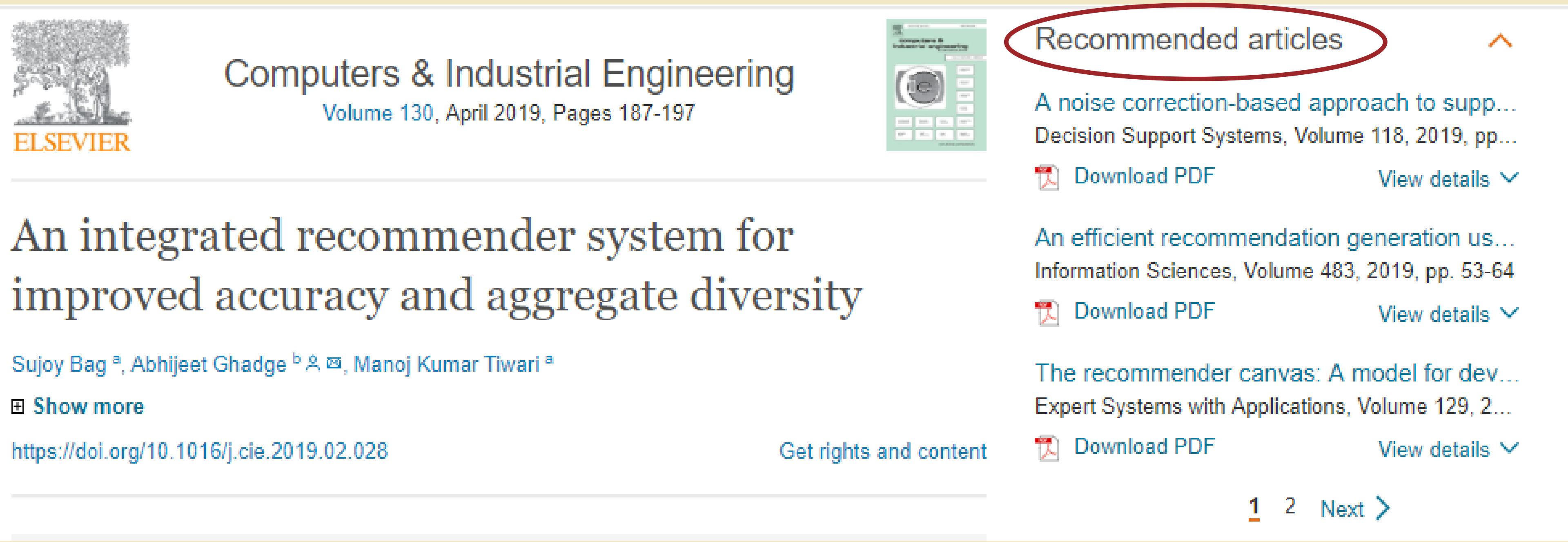
Upload a .txt or .docx file and find out which (if any) theses are conceptually similar.

Your literature review buddy ➔

Upload a .txt or .docx file and find out what works have been cited by conceptually similar theses.

That's right: we do your lit review for you.

Library use- Article recommendation



ELSEVIER

Computers & Industrial Engineering
Volume 130, April 2019, Pages 187-197

An integrated recommender system for improved accuracy and aggregate diversity

Sujoy Bag [a], Abhijeet Ghadge [b], Manoj Kumar Tiwari [a]

⊞ Show more

https://doi.org/10.1016/j.cie.2019.02.028

Get rights and content

Recommended articles

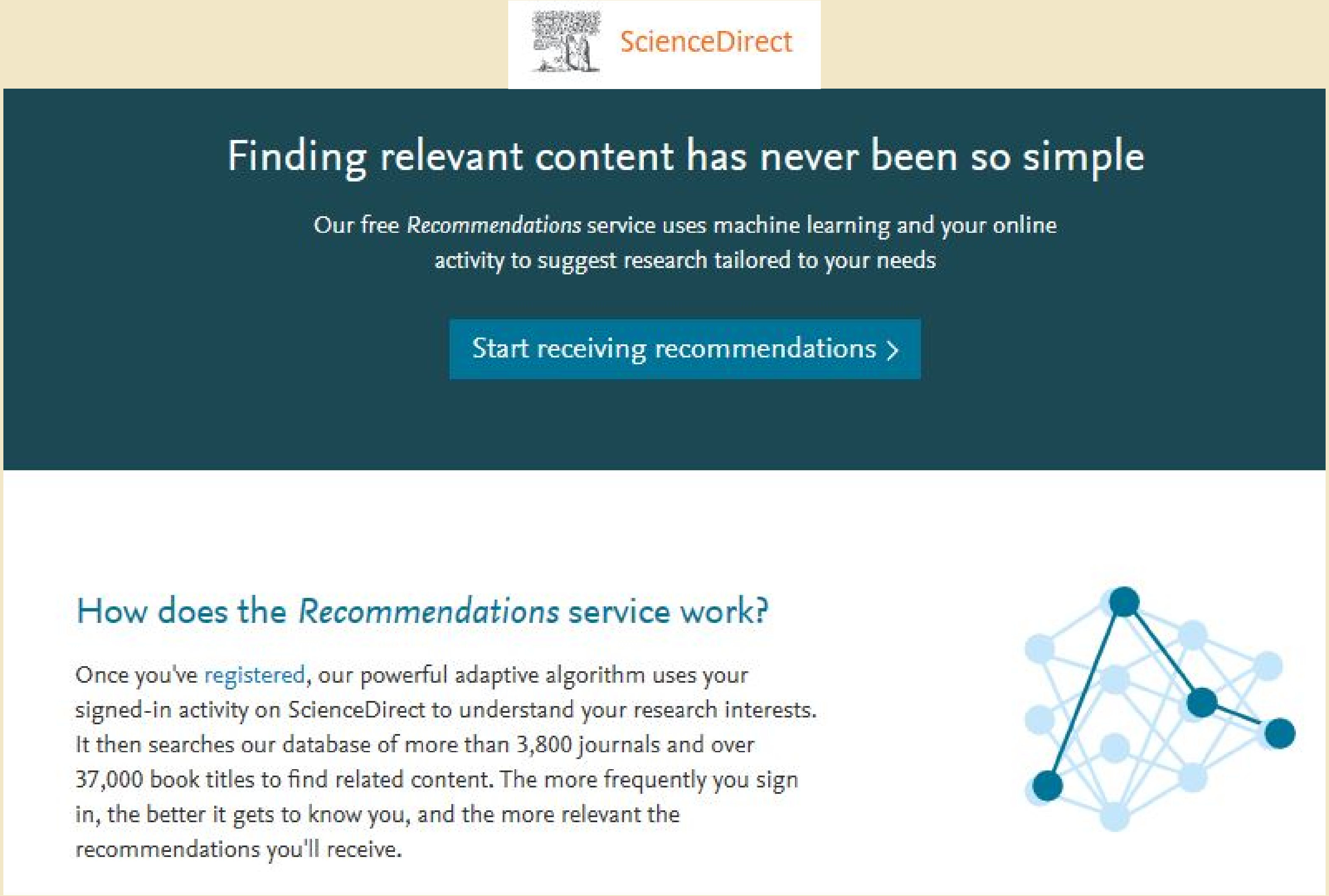
A noise correction-based approach to supp...
Decision Support Systems, Volume 118, 2019, pp...
Download PDF
View details

An efficient recommendation generation us...
Information Sciences, Volume 483, 2019, pp. 53-64
Download PDF
View details

The recommender canvas: A model for dev...
Expert Systems with Applications, Volume 129, 2...
Download PDF
View details

1 2 Next

Library use- Personalized recommendation service



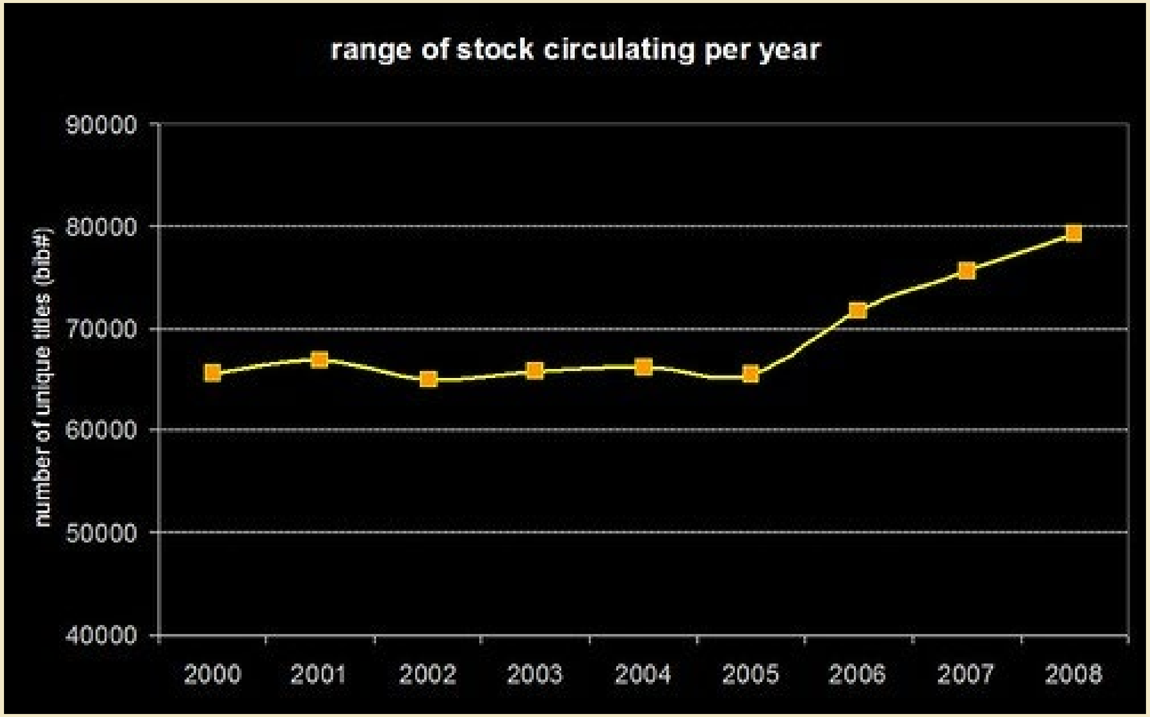
Finding relevant content has never been so simple

Our free *Recommendations* service uses machine learning and your online activity to suggest research tailored to your needs

How does the *Recommendations* service work?

Once you've registered, our powerful adaptive algorithm uses your signed-in activity on ScienceDirect to understand your research interests. It then searches our database of more than 3,800 journals and over 37,000 book titles to find related content. The more frequently you sign in, the better it gets to know you, and the more relevant the recommendations you'll receive.

Benefit of recommender systems



Effect of adding “people who borrowed this, also borrowed…” suggestions at the end of 2005 and adding personalized “we think you might be interested in…” suggestions in 2006 in a UK library

Recommendation techniques

- Content based
- Collaborative filtering-based
- Knowledge-based
- Hybrid
- Computational intelligence-based
- Social network-based
- Context awareness
- Group

Our focus

Collaborative filtering-based

- make choices based on the opinions of other people who share similar interests

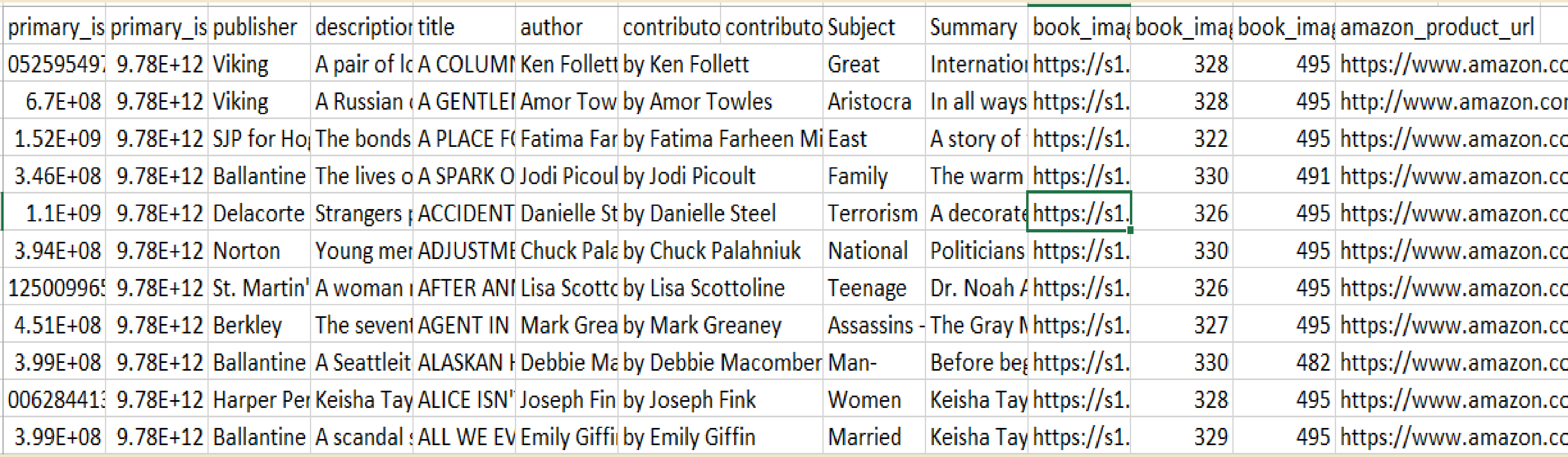
Content based

- Pair specific users to library items based on the metadata of the item and what is known about the user

Our project

Data Sources

- The New York Times
  - Hardcover Fiction Best Sellers (2018)
  - Weekly best sellers, 15 books each week



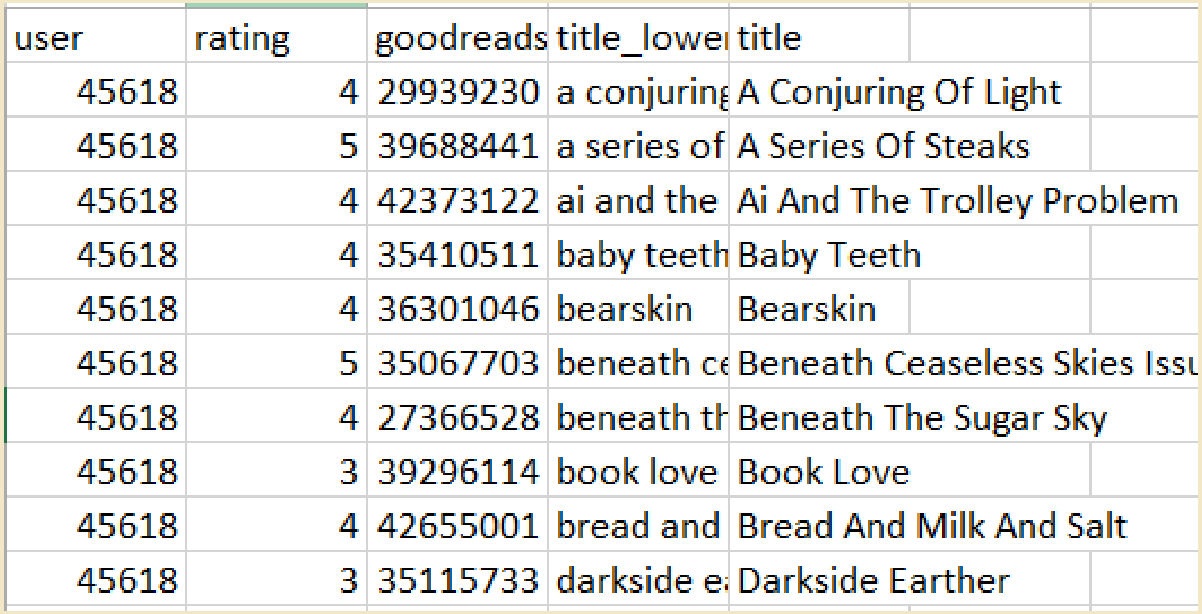
primary_is	primary_is	publisher	descriptio	title	author	contributo	contributo	Subject	Summary	book_ima	book_ima	book_ima	amazon_product_url
05259549	9.78E+12	Viking	A pair of l	A COLUM	Ken Follett	by Ken Follett		Great	Internatio	https://s1.	328	495	https://www.amazon.c
6.7E+08	9.78E+12	Viking	A Russian	A GENTLE	Amor Tow	by Amor Towles		Aristocra	In all ways	https://s1.	328	495	http://www.amazon.co
1.52E+09	9.78E+12	SJP for Ho	The bonds	A PLACE F	Fatima Far	by Fatima Farheen Mi		East	A story of	https://s1.	322	495	https://www.amazon.c
3.46E+08	9.78E+12	Ballantine	The lives o	A SPARK O	Jodi Picoul	by Jodi Picoult		Family	The warm	https://s1.	330	491	https://www.amazon.c
1.1E+09	9.78E+12	Delacorte	Strangers	ACCIDENT	Danielle St	by Danielle Steel		Terrorism	A decorat	https://s1.	326	495	https://www.amazon.c
3.94E+08	9.78E+12	Norton	Young me	ADJUSTM	Chuck Pal	by Chuck Palahniuk		National	Politicians	https://s1.	330	495	https://www.amazon.c
12500996	9.78E+12	St. Martin'	A woman	AFTER AN	Lisa Scott	by Lisa Scottoline		Teenage	Dr. Noah	https://s1.	326	495	https://www.amazon.c
4.51E+08	9.78E+12	Berkley	The seven	AGENT IN	Mark Grea	by Mark Greaney		Assassins -	The Gray	https://s1.	327	495	https://www.amazon.c
3.99E+08	9.78E+12	Ballantine	A Seattleit	ALASKAN	Debbie Ma	by Debbie Macomber		Man-	Before be	https://s1.	330	482	https://www.amazon.c
00628441	9.78E+12	Harper Pe	Keisha Tay	ALICE ISN	Joseph Fin	by Joseph Fink		Women	Keisha Tay	https://s1.	328	495	https://www.amazon.c
3.99E+08	9.78E+12	Ballantine	A scandal	ALL WE EV	Emily Giffi	by Emily Giffin		Married	Keisha Tay	https://s1.	329	495	https://www.amazon.c

Data Sources

- Goodreads

  -Ratings from the most popular 99 reviewers

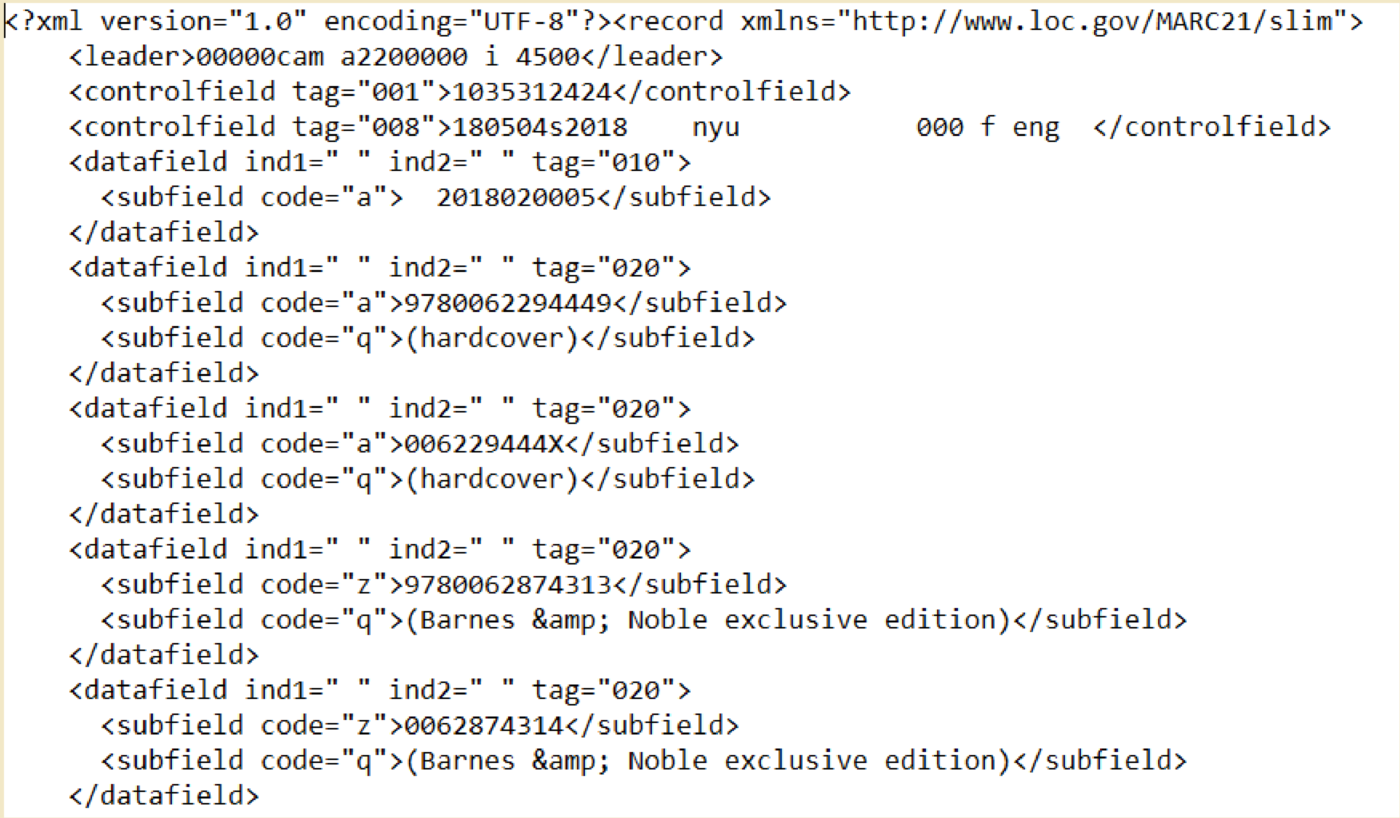
  - Their reviews from 2018



user	rating	goodreads	title_lowe	title
45618	4	29939230	a conjuring	A Conjuring Of Light
45618	5	39688441	a series of	A Series Of Steaks
45618	4	42373122	ai and the	Ai And The Trolley Problem
45618	4	35410511	baby teeth	Baby Teeth
45618	4	36301046	bearskin	Bearskin
45618	5	35067703	beneath ce	Beneath Ceaseless Skies Issu
45618	4	27366528	beneath th	Beneath The Sugar Sky
45618	3	39296114	book love	Book Love
45618	4	42655001	bread and	Bread And Milk And Salt
45618	3	35115733	darkside e	Darkside Earther

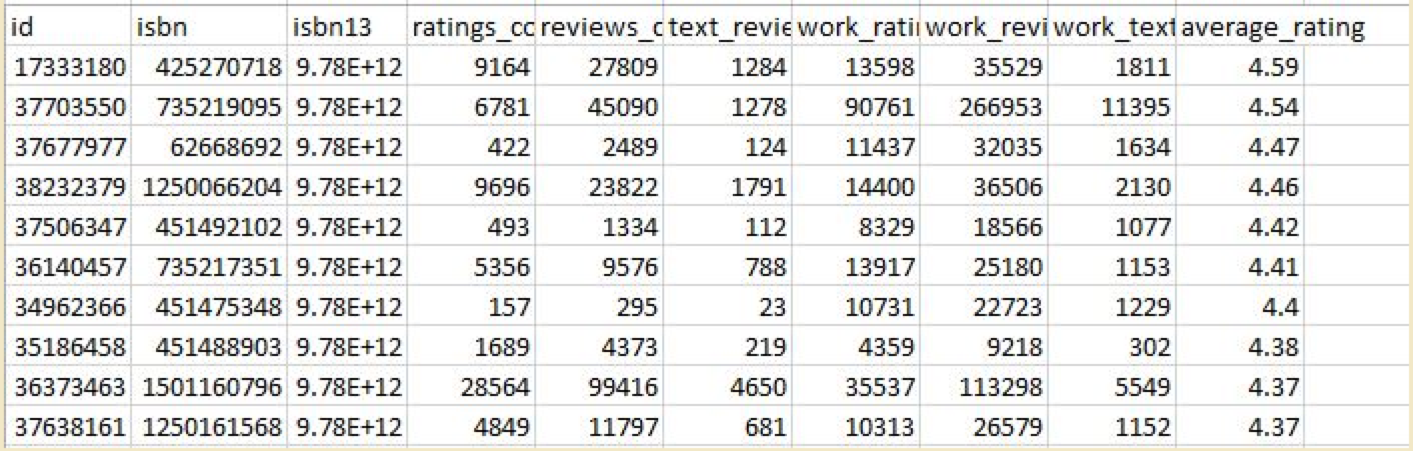
Data Sources

- Worldcat

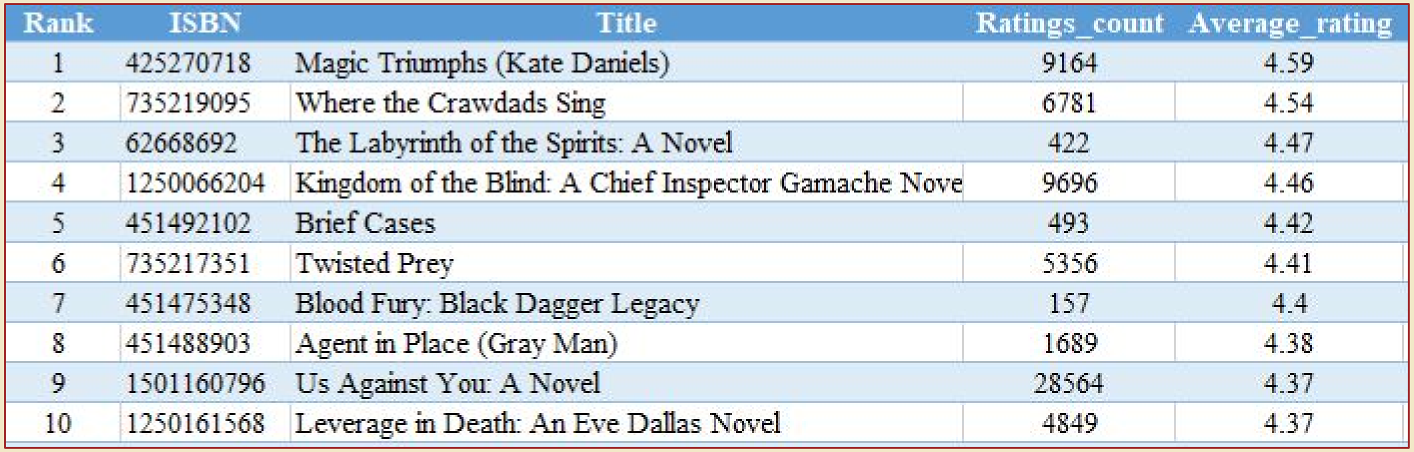


```
<?xml version="1.0" encoding="UTF-8"?><record xmlns="http://www.loc.gov/MARC21/slim">
    <leader>00000cam a2200000 i 4500</leader>
    <controlfield tag="001">1035312424</controlfield>
    <controlfield tag="008">180504s2018    nyu           000 f eng  </controlfield>
    <datafield ind1=" " ind2=" " tag="010">
      <subfield code="a">  2018020005</subfield>
    </datafield>
    <datafield ind1=" " ind2=" " tag="020">
      <subfield code="a">9780062294449</subfield>
      <subfield code="q">(hardcover)</subfield>
    </datafield>
    <datafield ind1=" " ind2=" " tag="020">
      <subfield code="a">006229444X</subfield>
      <subfield code="q">(hardcover)</subfield>
    </datafield>
    <datafield ind1=" " ind2=" " tag="020">
      <subfield code="z">9780062874313</subfield>
      <subfield code="q">(Barnes &amp; Noble exclusive edition)</subfield>
    </datafield>
    <datafield ind1=" " ind2=" " tag="020">
      <subfield code="z">0062874314</subfield>
      <subfield code="q">(Barnes &amp; Noble exclusive edition)</subfield>
    </datafield>
```

Top 10 Highly Rated NYT Bestsellers



id	isbn	isbn13	ratings_cc	reviews_c	text_revie	work_ratii	work_revi	work_text	average_rating
17333180	425270718	9.78E+12	9164	27809	1284	13598	35529	1811	4.59
37703550	735219095	9.78E+12	6781	45090	1278	90761	266953	11395	4.54
37677977	62668692	9.78E+12	422	2489	124	11437	32035	1634	4.47
38232379	1250066204	9.78E+12	9696	23822	1791	14400	36506	2130	4.46
37506347	451492102	9.78E+12	493	1334	112	8329	18566	1077	4.42
36140457	735217351	9.78E+12	5356	9576	788	13917	25180	1153	4.41
34962366	451475348	9.78E+12	157	295	23	10731	22723	1229	4.4
35186458	451488903	9.78E+12	1689	4373	219	4359	9218	302	4.38
36373463	1501160796	9.78E+12	28564	99416	4650	35537	113298	5549	4.37
37638161	1250161568	9.78E+12	4849	11797	681	10313	26579	1152	4.37



Rank	ISBN	Title	Ratings_count	Average_rating
1	425270718	Magic Triumphs (Kate Daniels)	9164	4.59
2	735219095	Where the Crawdads Sing	6781	4.54
3	62668692	The Labyrinth of the Spirits: A Novel	422	4.47
4	1250066204	Kingdom of the Blind: A Chief Inspector Gamache Nove	9696	4.46
5	451492102	Brief Cases	493	4.42
6	735217351	Twisted Prey	5356	4.41
7	451475348	Blood Fury: Black Dagger Legacy	157	4.4
8	451488903	Agent in Place (Gray Man)	1689	4.38
9	1501160796	Us Against You: A Novel	28564	4.37
10	1250161568	Leverage in Death: An Eve Dallas Novel	4849	4.37

Programming Language: Python

Libraries used

- Pandas--high-performance, easy-to-use data structures and data analysis tools
- Numpy--support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions
- Sklearn--machine learning library, tools for natural language processing

Methods

- Bayesian Estimate Algorithm

Weighted Rating (WR) =

$$\left(\frac{v}{v+m} \times R\right) + \left(\frac{m}{v+m} \times C\right)$$

Where:

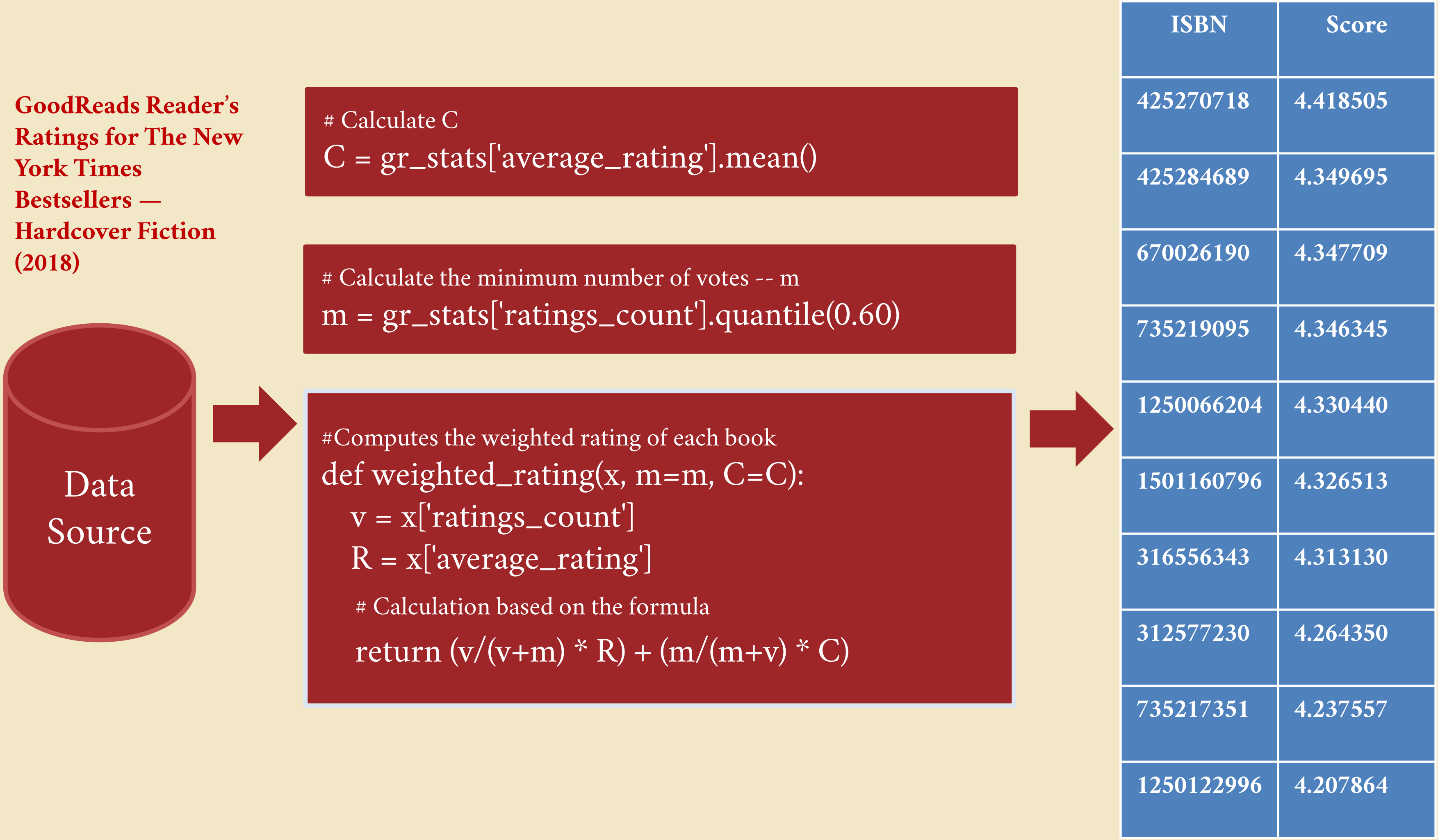
R = average for the books (mean) = (rating)

v = number of votes for the books = (votes)

m = minimum votes required to be listed

C = the mean vote across the whole dataset

Bayesian Estimate Recommendation



GoodReads Reader's Ratings for The New York Times Bestsellers — Hardcover Fiction (2018)

Data Source

```
# Calculate C
C = gr_stats['average_rating'].mean()
```

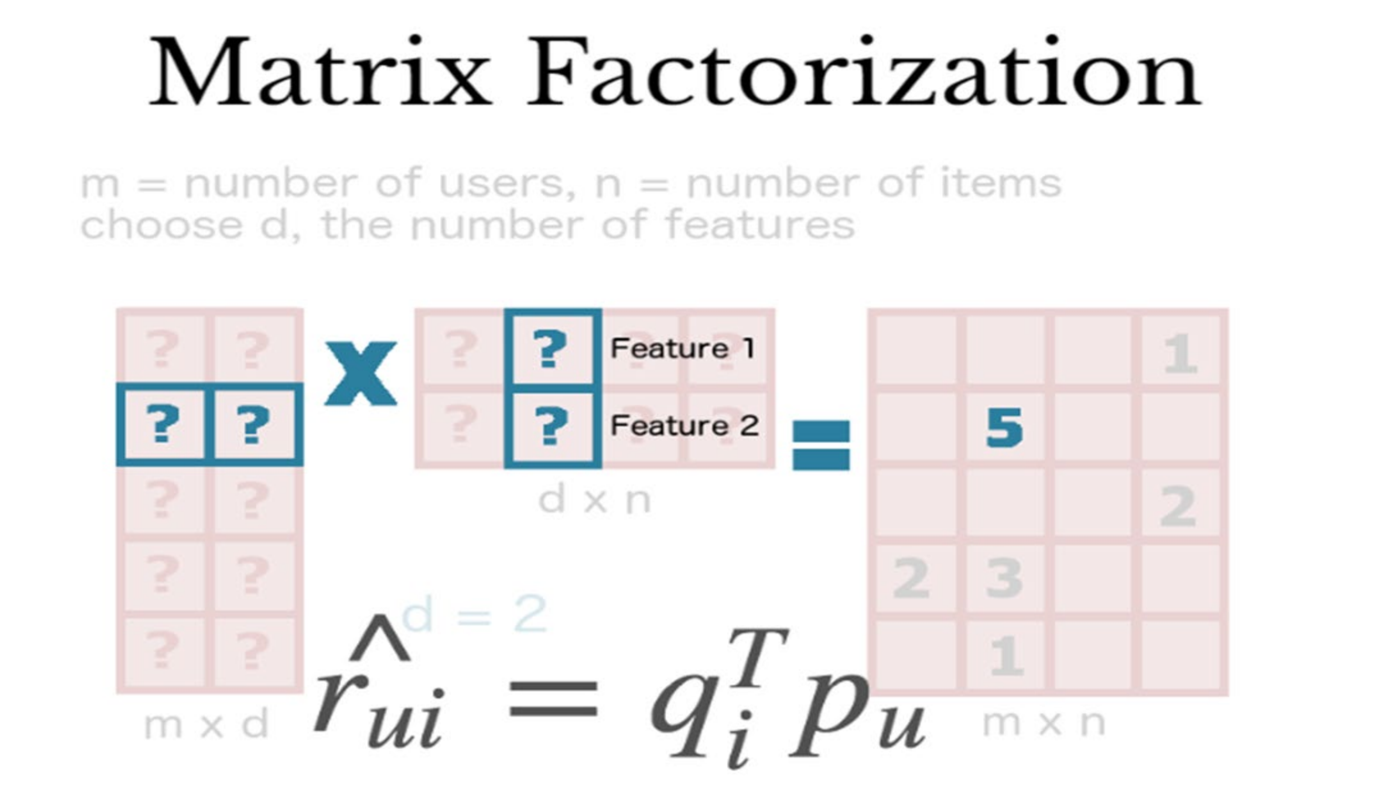
```
# Calculate the minimum number of votes -- m
m = gr_stats['ratings_count'].quantile(0.60)
```

```
#Computes the weighted rating of each book
def weighted_rating(x, m=m, C=C):
    v = x['ratings_count']
    R = x['average_rating']
     # Calculation based on the formula
     return (v/(v+m) * R) + (m/(m+v) * C)
```

ISBN	Score
425270718	4.418505
425284689	4.349695
670026190	4.347709
735219095	4.346345
1250066204	4.330440
1501160796	4.326513
316556343	4.313130
312577230	4.264350
735217351	4.237557
1250122996	4.207864

Methods

- Collaborative Filtering - Matrix Factorization Algorithm



$$\hat{r}_{ui} = q_i^T p_u$$

Collaborative Filtering - Matrix Factorization Recommendation

```
from sklearn.decomposition import TruncatedSVD
```

```
# Transpose book titles and userID
X=df.values.T
# Reduce dimension
SVD = TruncatedSVD(n_components=15, random_state=None)
matrix = SVD.fit_transform(X)
# Caculate the Pearson r correlation coefficient for every book
corr=np.corrcoef(X) alc
```

```
Recommend similar book to Evidence Of The Affair
['Clockwork Angel', 'Clockwork Prince', 'Clockwork Princess', 'Leah On The Offbeat', 'My Favorite Half Night Stand', 'One Day In December', 'The Lighthouse Keeper S Daughter']
```

Methods

- Content-Based – Cosine Similarity Algorithm

$$cosine(x, y) = \frac{x \,.\, y^T}{||x||\,.\,||y||}$$

Content-Based – Book Metadata Similarity Recommendation

```
from sklearn.metrics.pairwise import cosine_similarity
from sklearn.feature_extraction.text import CountVectorizer
```

```
# Join author and subject fields
def create_metadata(x):
    return ' '.join(x['author']) + ' '+ ' '.join(x['subject'].split('\n'))
                                                    …
# Compute the cosine similarity score
cosine_sim = cosine_similarity(count_matrix, count_matrix)
```

Recommend similar book to A PLACE FOR US

HEADS YOU WIN
TWISTED PREY
CHERRY
NEED TO KNOW
THE PRESIDENT IS MISSING

Methods

- Content-Based – TF-IDF Algorithm (Term Frequency — Inverse Document Frequency)

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i})$$

Where

- wi, j is the weight of word i in document j
- dfi is the number of documents that contain the term i
- N is the total number of documents

Content-Based – Book Summary TF-IDF Similarity Recommendation

```
from sklearn.feature_extraction.text  import  TfidfVectorizer
from sklearn.metrics.pairwise  import  linear_kernel
```

```
# Construct the required TF-IDF matrix
tfidf_matrix = tfidf.fit_transform(df['summary'])
# Compute the cosine similarity matrix
cosine_sim = linear_kernel(tfidf_matrix, tfidf_matrix)
```

Recommend similar book to ALL WE EVER WANTED

- WHERE THE CRAWDADS SING
- THE CLOCKMAKER'S DAUGHTER
- UNBOUND
- PIECES OF HER
- ALTERNATE SIDE
- THE IMMORTALISTS

Limitations

- Availability of data activity for supporting academic
- Lack of academic data integrity

Harvard Library APIs & Datasets

The Harvard Library provides open access to metadata through bibliographic datasets and APIs.

https://library.harvard.edu/services-tools/harvard-library-apis-datasets

- Privacy issues
- Unclear algorithms

Future uses

- Circulation data
- Interlibrary loan data
- Data from libraries and generating cross-library recommendations

Open Questions

What do you expect to see in libraries as machine learning systems integrate more in our daily operations?

Open Questions

What structural and personnel changes can we expect to see in libraries as machine learning systems become "good enough" for work done by us now?

Questions?

Jingshan Xiao
xiao@uhcl.edu
Wenli Gao
wgao5@uh.edu