Πανεπιστήμιο Δυτικής Αττικής
Τμήμα Μηχανικών Πληροφορικής και Υπολογιστών

ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΗΣ ΚΛΙΜΑΚΑΣ

ICE-359

ΓΑΓΓΑΣ ΙΩΑΝΝΗΣ
ΑΜ: 19390038 (ΠΑΔΑ)
ΠΡΩΙΟΣ ΠΑΝΤΕΛΕΗΜΩΝ
ΑΜ: 18390023 (ΠΑΔΑ)
ΜΑΝΤΖΟΥΚΑΣ ΑΓΓΕΛΟΣ ΒΑΣΙΛΕΙΟΣ
ΑΜ: 19390128 (ΠΑΔΑ)

Εργασία Εξαμήνου

Διαχείριση Δεδομένων Μεγάλης Κλίμακας

Εισαγωγή

Στα όρια του μαθήματος, κατασκευάσαμε ένα book recommendation system (σύστημα σύστασης βιβλίων). Το dataset που χρησιμοποιήθηκε για την παρούσα εργασία, υπάρχει διαθέσιμο στο kaggle. Το dataset, αποτελείται από 3 csv αρχεία, όπου το books.csv περιέχει πληροφορίες των βιβλίων, το users.csv για τους χρήστες και το ratings.csv παρέχει πληροφορίες για την αξιολόγηση ορισμένων χρηστών προς κάποια βιβλία με σχέση μεταξύ βιβλίων και χρηστών πολλά-προς-πολλά.

Ορισμός προβλήματος και κίνητρο

Το πρόβλημα το οποίο υπήρξε, είναι η εύστοχη σύσταση σε κάποιον χρήστη και όχι η τυχαία. Οι συστάσεις βασίζονται χρησιμοποιώντας, την ηλικία, τα βιβλία που έχει αξιολογήσει, ακόμα και την γεωγραφική τοποθεσία του σε σύγκριση με άλλους χρήστες. Τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν από τους ίδιους τους χρήστες εώς και πλατφόρμες ανάγνωσης/πώλησης βιβλίων.

Περιγραφή του συνόλου δεδομένων

Το dataset το οποίο προαναφέραμε, περιέχει 3 csv αρχεία. Το Books, το οποιο περιέχει:
το ISBN, τον τίτλο, τον συγγραφέα, το έτος έκδοσης και τον εκδοτικό οίκο.

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher
1	2005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
2	60973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
3	374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux
4	393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company
7	671870432	PLEADING GUILTY	Scott Turow	1993	Audioworks
...	...	...	...	...	...
271355	440400988	There's a Bat in Bunk Five	Paula Danziger	1988	Random House Childrens Pub (Mm)
271356	525447644	From One to One Hundred	Teri Sloat	1991	Dutton Books
271357	006008667X	Lily Dale : The True Story of the Town that Ta...	Christine Wicker	2004	HarperSanFrancisco
271358	192126040	Republic (World's Classics)	Plato	1996	Oxford University Press
271359	767409752	A Guided Tour of Rene Descartes' Meditations o...	Christopher Biffle	2000	McGraw-Hill Humanities/Social Sciences/Languages

Εικόνα 1: Δομή Books.csv

Το user, το οποιο περιέχει:
το ID του χρήστη, την γεωγραφική του τοποθεσία και την ηλικία του.

	User-ID	Location	Age
0	1	nyc, new york, usa	32
1	2	stockton, california, usa	18
2	3	moscow, yukon territory, russia	32
3	4	porto, v.n.gaia, portugal	17
4	5	farnborough, hants, united kingdom	32
...	...	...	...
278853	278854	portland, oregon, usa	32
278854	278855	tacoma, washington, united kingdom	50
278855	278856	brampton, ontario, canada	32
278856	278857	knoxville, tennessee, usa	32
278857	278858	dublin, n/a, ireland	32

Εικόνα 2: Δομή Users.csv

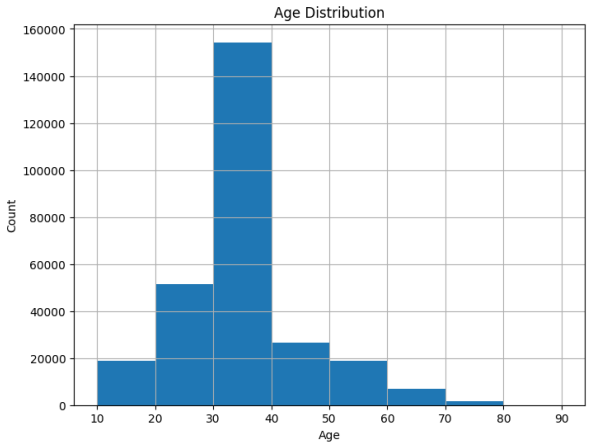
Τέλος, το ratings παρέχει πληροφορίες για την αξιολόγηση χρηστών σε βιβλία χρησιμοποιώντας το ISBN και το ID των χρηστών αξιολογώντας τα από 0 έως 10.

	User-ID	ISBN	Book-Rating
0	276725	034545104X	0
1	276726	0155061224	5
2	276727	0446520802	0
3	276729	052165615X	3
4	276729	0521795028	6
...	...	...	...
1149775	276704	1563526298	9
1149776	276706	0679447156	0
1149777	276709	0515107662	10
1149778	276721	0590442449	10
1149779	276723	05162443314	8

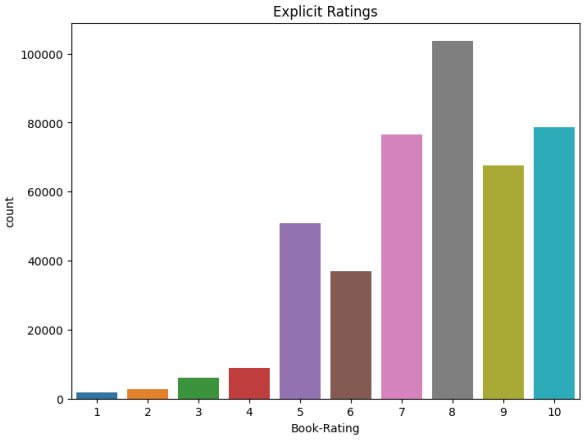
Εικόνα 3: Δομή Ratings.csv

Στο dataset έπρεπε να γίνει καθαρισμός των δεδομένων διότι υπήρχαν διάφορα προβλήματα.

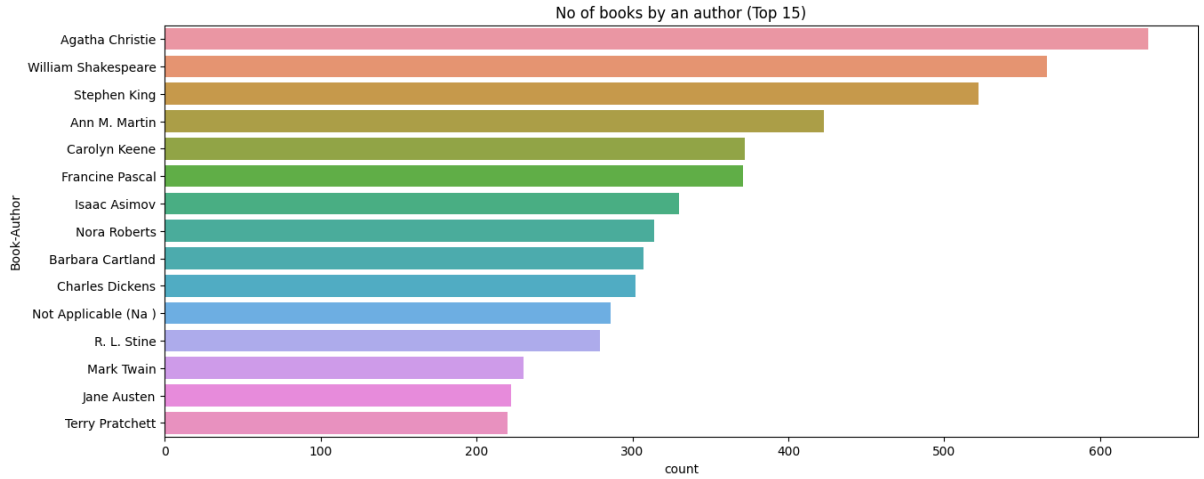
Μέθοδοι ανάλυσης των δεδομένων



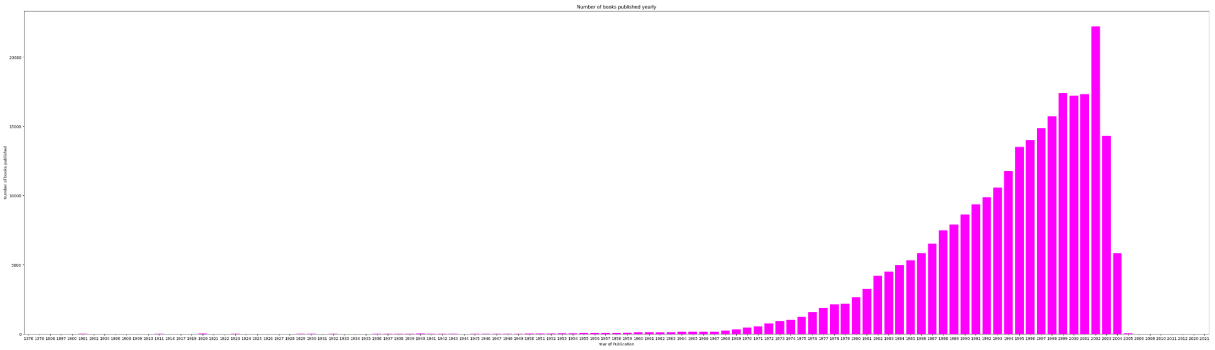
Ηλικιακή διάταξη των χρηστών



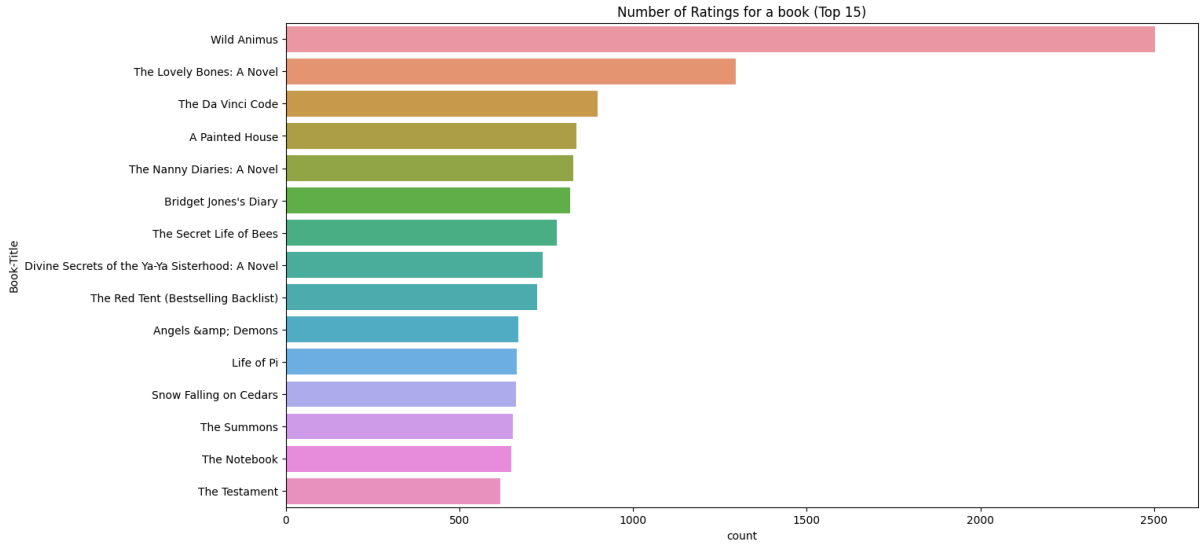
Διάταξη βαθμολογιών



Αριθμός βιβλίων ανά συγγραφέα



Αριθμός βιβλίων ανά έτος έκδοσης



Αριθμός κριτικών ανά βιβλίο

Πειραματικά αποτελέσματα

Αποτελέσματα για είσοδο: “The Da Vinci Code”

```
Books by same Author:

Digital Fortress : A Thriller
Illuminati.
Deception Point
Angels &amp;amp; Demons
El Codigo Da Vinci / The Da Vinci Code


Books by same Publisher:

Bleachers
The Curious Incident of the Dog in the Night-Time : A Novel
A Painted House
Tuesdays with Morrie: An Old Man, a Young Man, and Life's Greatest Lesson
This man Jesus;: An essay toward a New Testament Christology
```

Books with Same Author & Publisher

```
Input Book:
The Da Vinci Code

RECOMMENDATIONS:

Middlesex: A Novel
Angels &amp;amp; Demons
The Lovely Bones: A Novel
Nights in Rodanthe
The Secret Life of Bees
```

Collaborative Filtering

```
Input Book:

The Da Vinci Code

Recommended Books:

Middlesex: A Novel
The Brethren
Angels &amp;amp; Demons
Chicken Soup for the Soul (Chicken Soup for the Soul)
The Lovely Bones: A Novel
```

Hybrid Approach (Content & Collaborative)

Recommended Books:

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0671027360	Angels & Demons	Dan Brown	2001	Pocket Star
1	0345452577	The Conspiracy Club	JONATHAN KELLERMAN	2003	Ballantine Books
2	0446531316	Street Dreams	Faye Kellerman	2003	Warner Books
3	042516828X	Sudden Mischief	Robert B. Parker	1999	Berkley Publishing Group
4	042518904X	Widow's Walk	Robert B. Parker	2003	Berkley Publishing Group

Correlation Based

```
Recommended Books:

The Brethren
Chicken Soup for the Soul (Chicken Soup for the Soul)
Pigs in Heaven
A Is for Alibi (Kinsey Millhone Mysteries (Paperback))
Rising Sun
```

Content Based

Αποτελέσματα για είσοδο: “The Little Prince”

```
Books by same Author:

The Little Prince (Wordsworth Collection)
Terre Des Hommes
Il Piccolo Principe Prince Italn
Le Petit Prince
El principito (Spanish)
Der Kleine Prinz Prince German Hardy Boys


Books by same Publisher:

The Crimson Petal and the White
To the Lighthouse
Life of Pi
Between the Acts (Harvest Book)
The Name of the Rose: including Postscript to the Name of the Rose
East of the Mountains (Vintage Contemporaries (Paperback))
```

Books with Same Author & Publisher

```
Input Book:
The Little Prince

RECOMMENDATIONS:

When the Wind Blows
Life of Pi
Hideaway
House of Sand and Fog
While I Was Gone
Kiss the Girls
```

Collaborative Filtering

```
Input Book:

The Little Prince

Recommended Books:

When the Wind Blows
The Prince of Tides
Life of Pi
Little Altars Everywhere: A Novel
Hideaway
Daddy's Little Girl
```

Hybrid Approach (Content & Collaborative)

Recommended Books:

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher
0	0553277081	River of Swans (Spanish Bit Saga, No 10)	Don Coldsmith	1989	Bantam Books
1	0395933463	I Married a Communist	Philip Roth	1998	Houghton Mifflin Co
2	0310230055	Revelation Unveiled	Tim Lahaye	1999	Zondervan Publishing Company
3	0451524233	Tarzan of the Apes (Tarzan)	Edgar Rice Burroughs	1990	Signet Classics
4	0913668680	The Moosewood Cookbook: Recipes from Moosewood Restaurant, Ithaca, New York	Mollie Katzen	1977	Ten Speed Press
5	0786861495	Kink: An Autobiography	Dave Davies	1997	Hyperion Books

Correlation Based

```
Recommended Books:

The Prince of Tides
Little Altars Everywhere: A Novel
Daddy's Little Girl
Balzac and the Little Chinese Seamstress : A Novel
Don't Sweat the Small Stuff and It's All Small Stuff : Simple Ways to Keep the Little Things from Taking Over Your Life (Don't Sweat the Small Stuff Series)
From the Corner of His Eye
```

Content Based

Αποτίμηση αποτελεσμάτων

Στα αποτελέσματα content based, γίνεται χρήση του TF-IDF στον τίτλο με αντίκτυπο να μην είναι τόσο συναφή μεταξύ τους. Θα ήταν καλύτερο να υπάρχει περίληψη για το εκάστοτε βιβλίο και να γίνεται συγκρίση με αυτό και ίσος χρήση n-gram. Ωστόσο, οι υπόλοιποι μέθοδοι, όπως το correlation based και το collaborative filtering έχουν ως αποτέλεσμα τα βιβλία τα οποία έχουν το μεγαλύτερο score, να είναι πράγματι βιβλία τα οποία προηγούμενοι χρήστες έχουν αξιολογήσει θετικά και ταυτόχρονα, έχουν αξιολογήσει θετικα και αυτό τον τίτλο βιβλίου που χρησιμοποιήσαμε κατα την περισυλλογή πειραμάτων και αποτελεσμάτων.

Συμπεράσματα

Τα συμπεράσματα τα οποία λάβαμε από την συγκεκριμένη έρευνα και ανάπτυξη του συστήματος συστάσεων βιβλίων είναι:

1. Οι προτιμήσεις των χρηστών: Μπορούμε να αναγνωρίσουμε βιβλία τα οποία είναι δημοφιλή μεταξύ τους ακόμα και αν παίζει ρόλο η παράμετρος της τοποθεσίας των χρηστών καθώς και ποια είδη βιβλίων είναι πιο δημοφιλή και ποιοι συγγραφείς έχουν μεγαλύτερη απήχηση.
2. Σχέσεις μεταξύ βιβλίων: Είναι φανερό ότι οι ορισμένα βιβλία μπορεί να έχουν τον ίδιο εκδότη ή ακόμη και τον ίδιο συγγραφέα και να έχουν λίγες ή ακόμα και καμία διαφορά μεταξύ τους.
3. Προσωποποίηση συστάσεων: Μέσα από το ratings.csv ήμασταν σε θέση να αναλύσουμε καθώς και να προσωποποίησουμε τις συστάσεις βιβλίων ή ακόμα και να καθοδηγηθούμε σύμφωνα με τον κάθε χρήστη.
4. Αποτελεσματικότητα συστήματος συστάσεων: Ο λόγος που χρησιμοποιήθηκαν διαφορετικές μέθοδοι συστάσεων (Collaborative filtering, Content Based, Correlation Based, Hybrid approach) ήταν για να δούμε και να συγκρίνουμε την τυχόν απόκλιση μεταξύ των αποτελεσμάτων κάθε μεθόδου. Η σύγκριση των αποτελεσμάτων μεταξύ κάθε μεθόδου ήταν αρκετά διαφορετική και πολλές φορές τα αποτελέσματα απέκλειναν ελαφρά μεταξύ τους. Ο λόγος για τον οποίο υπάρχει αυτή απόκλιση είναι λόγω των διαφορετικών αρχών και τεχνικών που χρησιμοποιούν.

Βιβλιογραφία

Book-recommendation. (2023, June 15). Kaggle. Retrieved July 1, 2023, from: https://www.kaggle.com/code/claricesatikoaoto/book-recomendation

Book Recommendation Dataset. (n.d.). Kaggle. Retrieved July 1, 2023, from: https://www.kaggle.com/datasets/arashnic/book-recommendation-dataset/code

Book_recommender_system. (2023, May 22). Kaggle. Retrieved July 1, 2023, from: https://www.kaggle.com/code/mdwaquarazam/book-recommender-system

Guide, S., & Agrawal, R. (2021, June 27). *Book Recommendation System | Build A Book Recommendation System*. Analytics Vidhya. Retrieved July 1, 2023, from: https://www.analyticsvidhya.com/blog/2021/06/build-book-recommendation-system-unsupervised-learning-project/

Popularity-based & Collaborative Filtering. (n.d.). Kaggle. Retrieved July 1, 2023, from: https://www.kaggle.com/code/saijyotitripathy/popularity-based-collaborative-filtering