Δημιουργία συστήματος συστάσεων με τεχνικές συνεργατικού φιλτραρίσματος

Εξαμηνιαία Εργασία μαθήματος “Ανάλυση και Σχεδιασμός Πληροφοριακών Συστημάτων”

Μπενέτου Σμαραγδή
Στράτη Θεώνη Μαρία
Τσαρμποπούλου Μαργαρίτα Ελένη

Περιεχόμενα

- Συστήματα Συστάσεων-Συνεργατικό Φιλτράρισμα
- Εξοικείωση με το σύνολο δεδομένων
- Προετοιμασία δεδομένων
- Καθορισμός training και test set
- Μείωση διαστάσεων με SVD σε Apache Spark
- Εκτίμηση και σφάλμα εκτίμησης στο test set

Συστήματα Συστάσεων-Συνεργατικό Φιλτράρισμα

Τα συστήματα συστάσεων προβλέπουν τις προτιμήσεις των χρηστών αναλύοντας την προηγούμενη συμπεριφορά. Οι τύποι συστημάτων συστάσεων περιλαμβάνουν συνεργατικό φιλτράρισμα, βασισμένο σε περιεχόμενο, υβριδικά μοντέλα, βασισμένα στη γνώση και με επίγνωση περιεχομένου. Το συνεργατικό φιλτράρισμα μπορεί να χρησιμοποιήσει μια προσέγγιση βάσει μοντέλου ή γειτονιάς/μνήμης. Το φιλτράρισμα βάσει περιεχομένου λαμβάνει υπόψη τα χαρακτηριστικά των στοιχείων, το συνεργατικό τις αξιολογήσεις χρηστών, ενώ το υβριδικό φιλτράρισμα συνδυάζει αξιολογήσεις και χαρακτηριστικά.

Προσέγγιση Γειτονιάς/Μνήμης: Similarity Metrics

Cosine similarity:

$$sim_{\cos}(i,j) = \frac{i \cdot j}{\|i\|^2 \| j\|^2}$$

Adjusted Cosine similarity:

$$sim_{adj\cos}(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R_u})(R_{u,j} - \bar{R_u})}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R_u})^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R_u})^2}}$$

Pearson correlation:

$$sim_{adj\cos}(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R_i})(R_{u,j} - \bar{R_j})}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R_i})^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R_j})^2}}$$

Euclidean Distance:

$$sim_{adj\cos}(i,j) = \frac{\sum_{u \in U}(k_{u,i} - \bar{k_i})(k_{u,j} - \bar{k_j})}{\sqrt{\sum_{u \in U}(k_{u,i} - \bar{k_i})^2}\sqrt{\sum_{u \in U}(k_{u,j} - \bar{k_j})^2}}$$

Spearman Correlation:

$$sim_{adj\cos}(i,j) = \frac{\sum_{u \in U}(k_{u,i} - \bar{k_i})(k_{u,j} - \bar{k_j})}{\sqrt{\sum_{u \in U}(k_{u,i} - \bar{k_i})^2}\sqrt{\sum_{u \in U}(k_{u,j} - \bar{k_j})^2}}$$

Mean Squared Distance:

$$sim_{adj\cos}(i,j) = \frac{\sum_{u \in U}(k_{u,i} - \bar{k_i})(k_{u,j} - \bar{k_j})}{\sqrt{\sum_{u \in U}(k_{u,i} - \bar{k_i})^2}\sqrt{\sum_{u \in U}(k_{u,j} - \bar{k_j})^2}}$$

Προσέγγιση Γειτονιάς/Μνήμης: Prediction Approaches

Weighted Average:

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u(i)} sim(i,j) r_{uj}}{\sum_{j \in N_u(i)} |sim(i,j)|}$$

Mean centering:

$$\hat{r}_{ui} = \bar{r_i} + \frac{\sum_{j \in N_u(i)} sim(i,j)(r_{uj} - \bar{r_j})}{\sum_{j \in N_u(i)} |sim(i,j)|}$$

Προσέγγιση Γειτονιάς/Μνήμης: Performance Metrics

Mean absolute error:

$$Error_{mae} = \frac{\sum_{i=1}^{N} |r_i - \hat{r_i}|}{N}$$

Root mean squared error:

$$Error_{rmse} = \sqrt{\frac{\sum_{i=1}^{N} (r_i - \hat{r_i})^2}{N}}$$

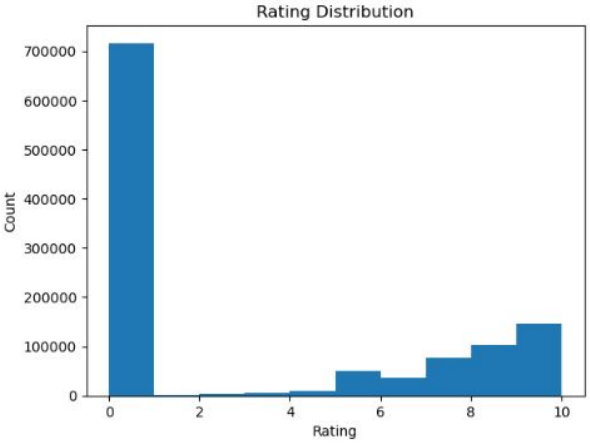
Προσέγγιση με βάση το μοντέλο

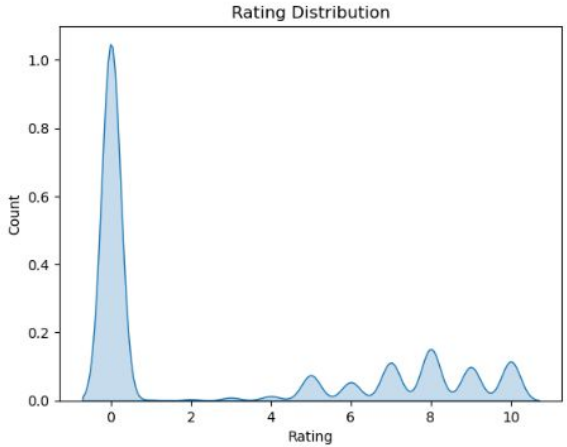
Μείωση διαστάσεων ενός πίνακα A σε
k διαστάσεις με SVD:

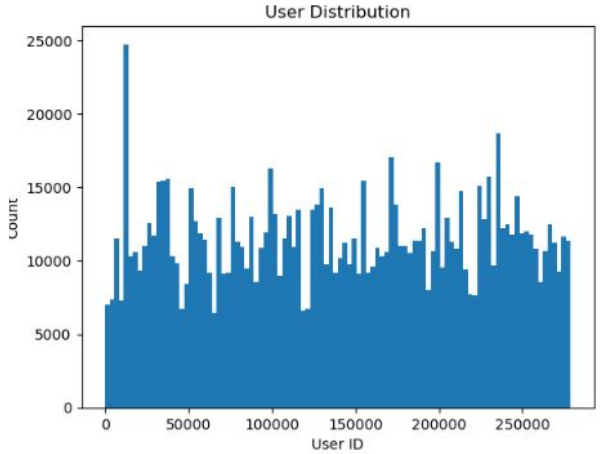
$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

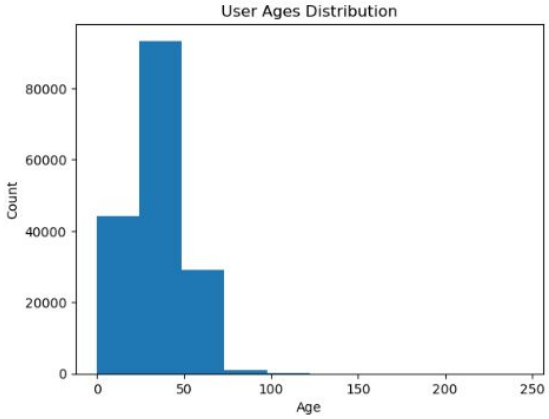
Εξοικείωση με το σύνολο δεδομένων

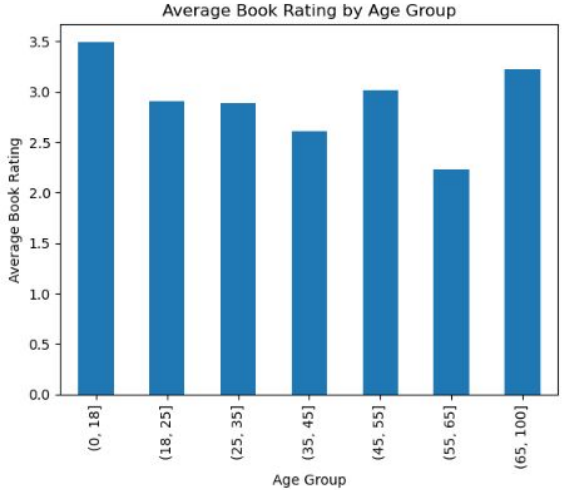
- Διαβάζουμε το dataset που περιέχει πληροφορίες για τους χρήστες, τα βιβλία και τις κριτικές
- Προχωράμε σε οπτικοποιήσεις του συνόλου δεδομένων με διαγράμματα

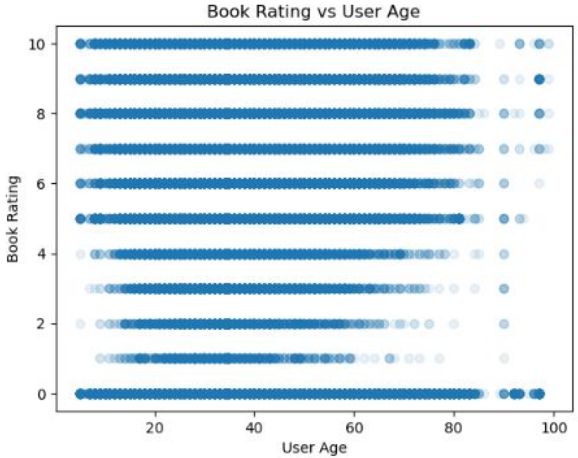

Rating Distribution
Count
Rating
0
100000
200000
300000
400000
500000
600000
700000
0
2
4
6
8
10

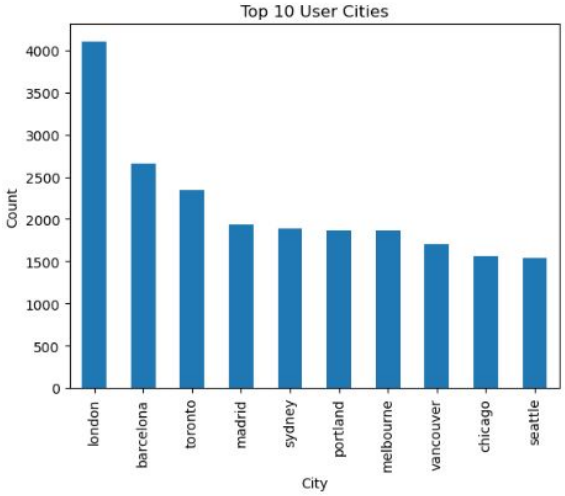

Rating Distribution
Count
Rating
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10

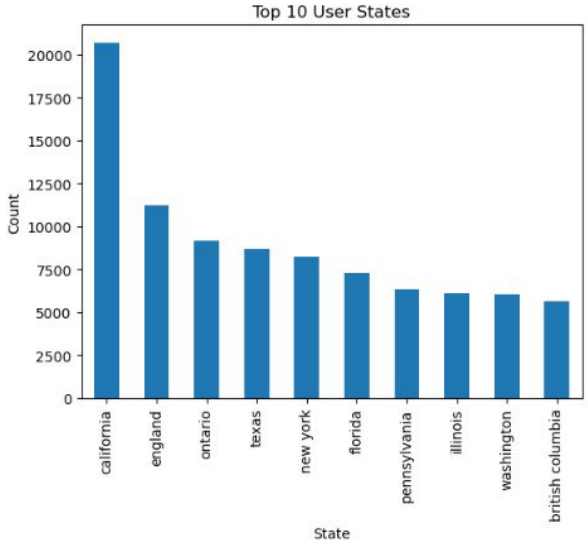

User Distribution
Count
User ID
0
5000
10000
15000
20000
25000
0
50000
100000
150000
200000
250000

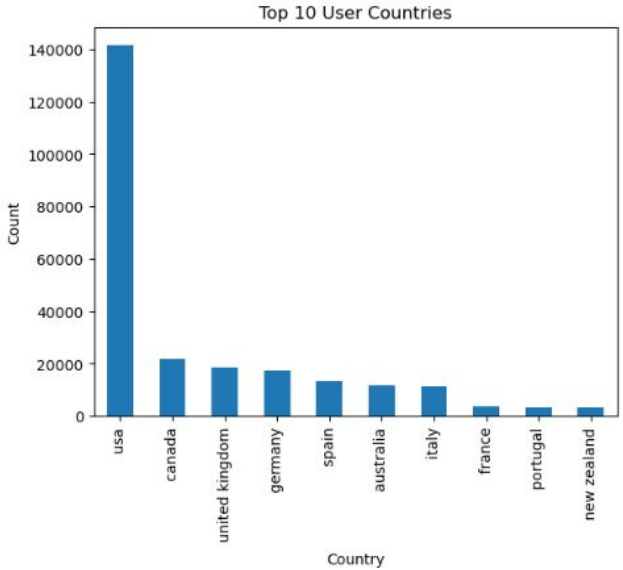

User Ages Distribution
Count
Age
0
20000
40000
60000
80000
0
50
100
150
200
250

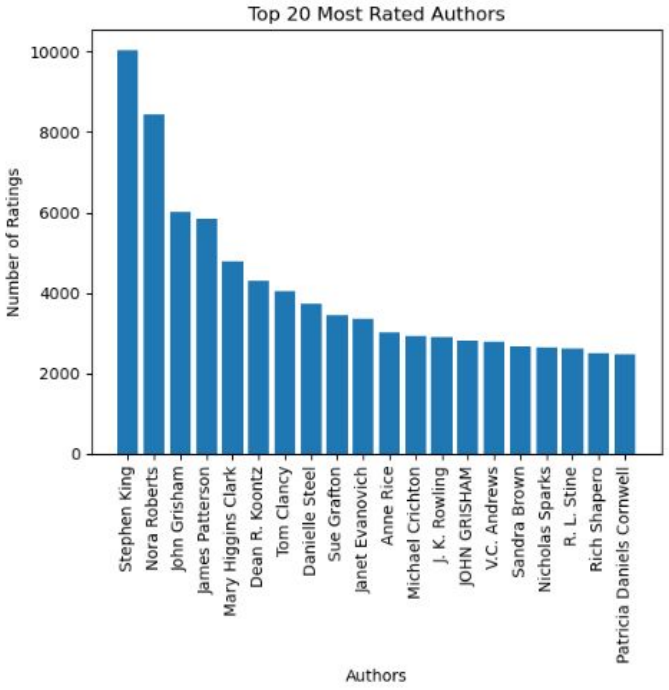

Average Book Rating by Age Group
Average Book Rating
Age Group
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
(0, 18]
(18, 25]
(25, 35]
(35, 45]
(45, 55]
(55, 65]
(65, 100]

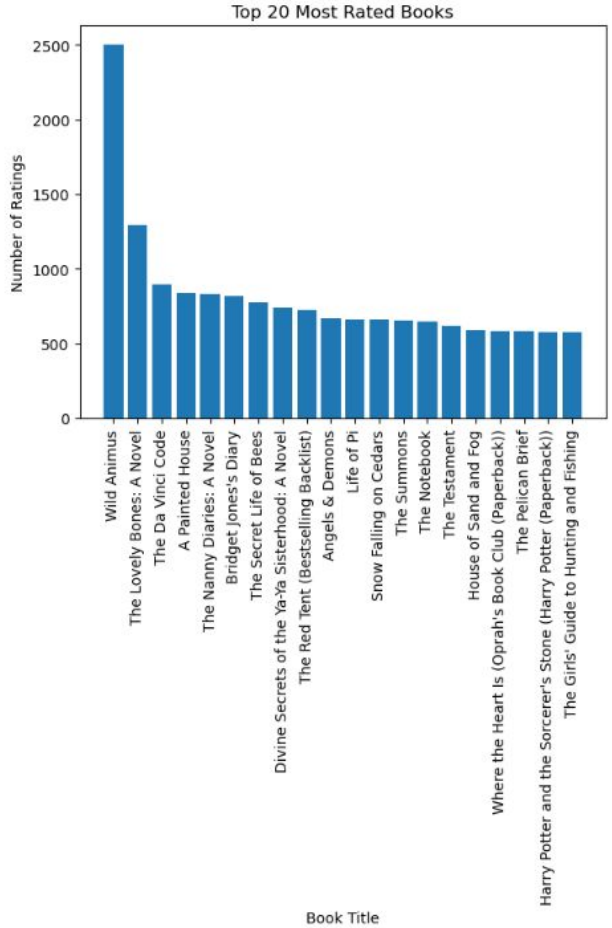

Book Rating vs User Age
Book Rating
User Age
0
2
4
6
8
10
20
40
60
80
100


Top 10 User Cities
Count
4000
3500
3000
2500
2000
1500
1000
500
0
london
barcelona
toronto
madrid
sydney
portland
melbourne
vancouver
chicago
seattle
City


Top 10 User States
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
california
england
ontario
texas
new york
florida
pennsylvania
illinois
washington
british columbia
State

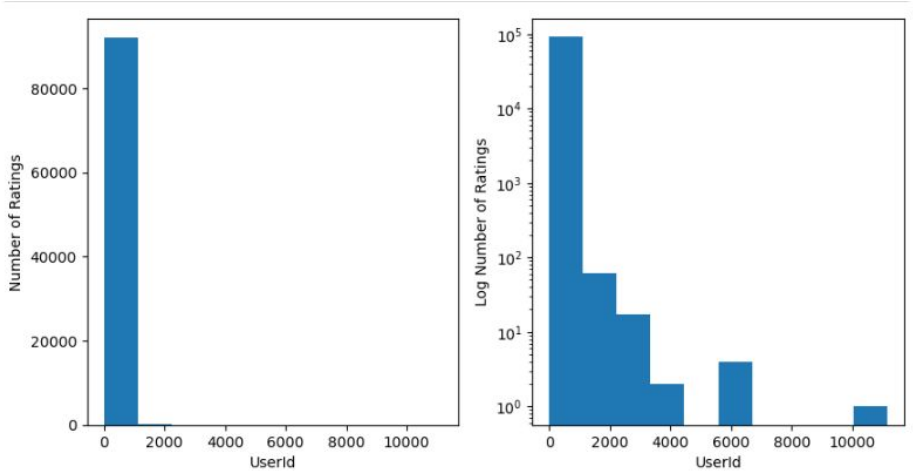

Top 10 User Countries
Count
140000
120000
100000
80000
60000
40000
20000
0
usa
canada
united kingdom
germany
spain
australia
italy
france
portugal
new zealand
Country

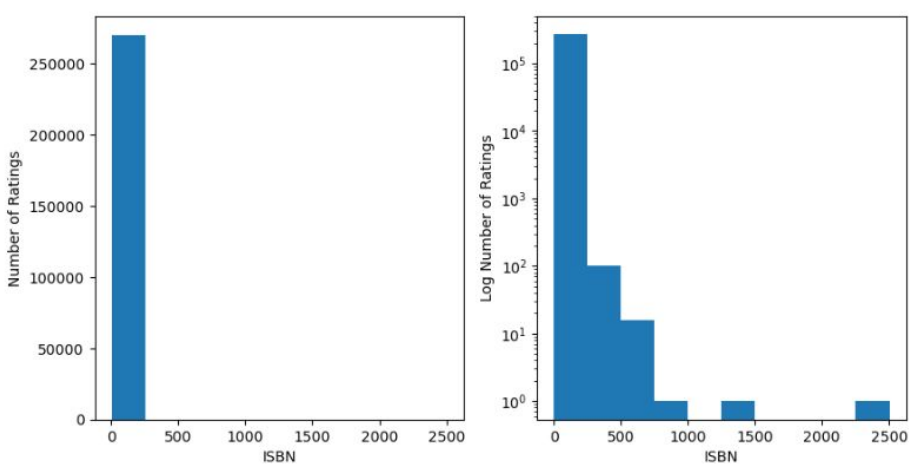

Top 20 Most Rated Authors
Number of Ratings
0
2000
4000
6000
8000
10000
Stephen King
Nora Roberts
John Grisham
James Patterson
Mary Higgins Clark
Dean R. Koontz
Tom Clancy
Danielle Steel
Sue Grafton
Janet Evanovich
Anne Rice
Michael Crichton
J. K. Rowling
JOHN GRISHAM
V.C. Andrews
Sandra Brown
Nicholas Sparks
R. L. Stine
Rich Shapero
Patricia Daniels Cornwell
Authors


Top 20 Most Rated Books
Number of Ratings
0
500
1000
1500
2000
2500
Wild Animus
The Lovely Bones: A Novel
The Da Vinci Code
A Painted House
The Nanny Diaries: A Novel
Bridget Jones's Diary
The Secret Life of Bees
Divine Secrets of the Ya-Ya Sisterhood: A Novel
The Red Tent (Bestselling Backlist)
Angels & Demons
Life of Pi
Snow Falling on Cedars
The Summons
The Notebook
The Testament
House of Sand and Fog
Where the Heart Is (Oprah's Book Club (Paperback))
The Pelican Brief
Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))
The Girls' Guide to Hunting and Fishing
Book Title

Προετοιμασία δεδομένων

- Επιλέγουμε τις πιο σημαντικές κατηγορίες του συνόλου, δηλαδή τα User-IDs, ISBNs, Book Ratings
- Σχεδιάζουμε ιστογράμματα για την οπτικοποίηση της κατανομής των αξιολογήσεων σε χρήστες και βιβλία
- Ο μέσος αριθμός αξιολογήσεων ανά χρήστη βρέθηκε να είναι 11,20, ενώ ο διάμεσος αριθμός αξιολογήσεων ανά χρήστη ήταν 1,00. Ομοίως, ο μέσος αριθμός αξιολογήσεων ανά αντικείμενο ήταν 3,82 και ο μέσος αριθμός αξιολογήσεων ανά αντικείμενο ήταν 1,00.

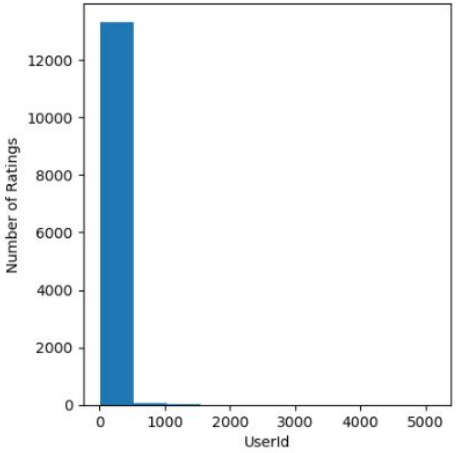


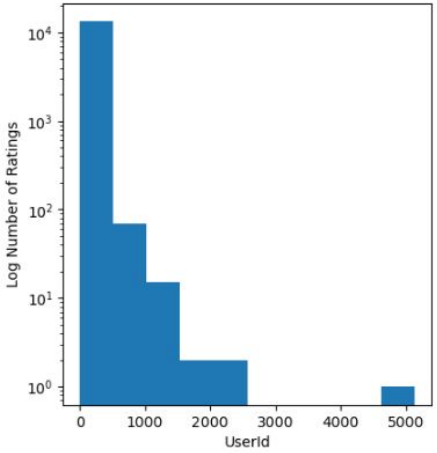


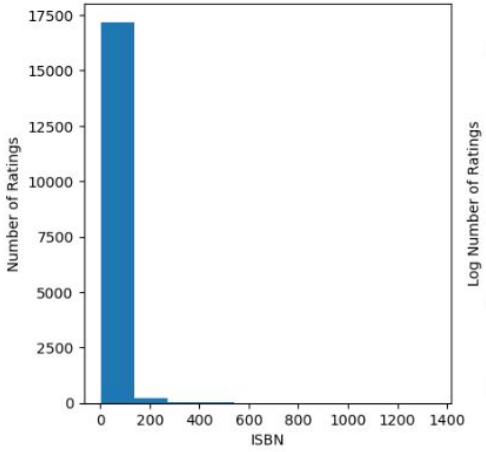
- Παρατηρούμε ότι ένα μεγάλο μέρος των δεδομένων μπορεί να αποτελείται από χρήστες και αντικείμενα με λίγες μόνο αξιολογήσεις. Επομένως, ορίζουμε έναν ελάχιστο αριθμό αξιολογήσεων ως κατώτατο όριο, ώστε να χρησιμοποιούνται μόνο τα πιο αξιόπιστα δεδομένα.

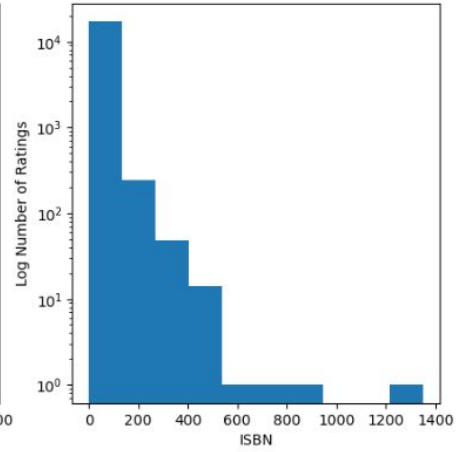
- Έτσι, επιλέγουμε να αφαιρέσουμε από το σύνολο δεδομένων τα βιβλία με λιγότερες από 10 αξιολογήσεις και οι χρήστες που έχουν αξιολογήσει λιγότερο από 5 φορές.
- Δημιουργούμε δύο pivot tables με τα preprocessed data των αξιολογήσεων βιβλίων και χρηστών και συμπληρώνουμε 0 στις τιμές που λείπουν.
- Υπολογίζουμε τις κανονικοποιημένες τιμές, αφαιρώντας τη μέση βαθμολογία από κάθε βαθμολογία στους πίνακες pivot, ώστε να εξαλείψουμε το bias και να βελτιωθεί η ποιότητα του συνόλου δεδομένων.

Τα νέα ιστογράμματα που προκύπτουν είναι τα παρακάτω:









Καθορισμός training και test set

- Η αξιολόγηση της απόδοσης του συστήματος συστάσεων γίνεται με το διαχωρισμό των δεδομένων σε training και test έτσι ώστε να σχεδιαστεί το μοντέλο με βάση το σύνολο εκπαίδευσης και να εξεταστεί η εγκυρότητά του στο σύνολο test.
- Λόγω της μεγάλης πολυπλοκότητας υπολογισμού των πινάκων SVD χωρίζουμε τυχαία το dataset, έτσι ώστε το 90% να είναι training και το 10% test αντί να χρησιμοποιήσουμε την τεχνική K-Fold Cross Validation.

Υπολογισμός SVD

- Στόχος είναι να μειώσουμε τις διαστάσεις του pivot table που προκύπτει από το σύνολο εκπαίδευσης δηλαδή τον πίνακα που η κάθε γραμμή αντιστοιχεί σε διαφορετικό χρήστη και η κάθε στήλη σε διαφορετικό βιβλίο με στοιχεία την αξιολόγηση του δεδομένου χρήστη για το δεδομένο βιβλίο.
- Κανονικοποίηση του πίνακα με αφαίρεση από κάθε γραμμή την μέση τιμή αξιολόγησης του χρήστη.
- Ο υπολογισμός SVD έγινε και με την βιβλιοθήκη scipy.linalg και με pyspark

Περιβάλλον Apache Spark και χρήση Cluster

- Εγκαταστάθηκε Spark σε Yarn με σύστημα αρχείων HDFS στο cluster που είχαμε πρόσβαση από το okeanos knossos.
- Δύο machines(master και slave) με 8GB RAM και 4 CPUs το καθένα.
- Το σύνολο δεδομένων φορτώθηκε στο hdfs σε μορφή csv
- Μέσω pyspark και της συνάρτησης computeSVD υπολογίστηκε ο SVD πάνω στο pivot table

Εκτίμηση και σφάλμα εκτίμησης

- Στο τεστ σετ κάθε δείγμα δίνει το βιβλίο και τον χρήστη για τον οποίο θέλουμε να εκτιμήσουμε την αξιολόγηση.
- Στον πίνακα U, που παράγεται από τον SVD και αφού αφαιρεθούν οι στήλες ανάλογα με τις απαιτούμενες διαστάσεις, κάθε γραμμή αντιπροσωπεύει έναν χρήστη . Επομένως, για δεδομένο χρήστη μπορεί να υπολογιστεί η ομοιότητά του με άλλους(μόνο όσους έχουν αξιολογήσει το δεδομένο βιβλίο) είτε με cosine similarity είτε με εσωτερικό γινόμενο.
- Με την μετρική weighted average οι αξιολογήσεις των υπόλοιπων χρηστών αυτού του βιβλίου με βάρη την ομοιότητα τους με τον δεδομένο χρήστη παράγουν την εκτίμηση μας.
- Το σφάλμα εκτίμησης στο τεστ σύνολο παράγεται με την μετρική Root Mean Square Error.

Αποτελέσματα

Παρατηρούμε για παράδειγμα ότι για το δείγμα του τεστ σετ:

```
]: n = 2

predict(X_test[n][0], X_test[n][1], U30, 30)

3.789473684565029

+ Code + Markdown

]: y_test[n]

5
```

```
]: n = 2

]: predict(X_test[n][0], X_test[n][1], U20, 20)

3.789473684565029

y_test[n]

5
```

```
[182]: n = 2

predict(X_test[n][0], X_test[n][1], U10, 10)

[185… 5.286856875405329

+ Code + Markdown

[184]: y_test[n]

[184… 5
```