ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

BOOKNEST

Βαρβάρης Πολυχρόνης
ΑΜ: 185152
polihronisv@gmail.com

ΘΕΣΣΑΛΟΝΙΚΗ, 10/1/25

Εργασία υποβαλλόμενη για το Μάθημα « Ανάκτηση Πληροφοριών – Μηχανές Αναζήτησης »

Περίληψη

Η εργασία επικεντρώνεται στην ανάπτυξη μιας τοπικής desktop-based μηχανής αναζήτησης που αξιοποιεί τεχνικές επεξεργασίας γλώσσας και αλγορίθμους αναζήτησης, για την ανάλυση και ανάκτηση πληροφοριών-δεδομένων. Ο στόχος της είναι η αποτελεσματική εύρεση δεδομένων και η ακριβής παρουσίαση των αποτελεσμάτων σύμφωνα με τα κριτήρια του χρήστη. Η εργασία απαιτεί την επιλογή μιας πηγής δεδομένων που πληροί συγκεκριμένες προδιαγραφές, όπως να περιέχει τουλάχιστον 500 εγγραφές, η ύπαρξη πεδίων κειμένου και αριθμητικών τιμών και η δυνατότητα αναζήτησης σε διαφορετικά πεδία.

Λέξεις Κλειδιά:

1. **TF-IDF (Term Frequency-Inverse Document Frequency):** Λογισμικό που χρησιμοποιείται για να αξιολογηθεί η σημασία μιας λέξης σε ένα έγγραφο, λαμβάνοντας υπόψη τη συχνότητά της και το πόσο κοινή είναι σε ολόκληρο το σύνολο δεδομένων.
2. **Tokenization: Δ**ιαδικασία διαχωρισμού κειμένου σε μικρότερες μονάδες, όπως λέξεις ή προτάσεις, που ονομάζονται tokens.
3. **Stop-Words Removal: Δ**ιαδικασία αφαίρεσης λέξεων που είναι αρκετά συχνές και δεν προσφέρουν ουσιαστική πληροφορία (π.χ., "και", "του", "αυτό").
4. **Stemming:** Μέθοδος για την ανάκτηση της "ρίζας" μιας λέξης, αφαιρώντας καταλήξεις (π.χ., "τρέχω" -> "τρέχ").
5. **Lemmatization: Δ**ιαδικασία που επιστρέφει τη "βασική μορφή" μιας λέξης (π.χ., "τρέχει" -> "τρέχω").
6. **Συνάφεια (Relevance):** Ένα μέτρο αρίθμισης για το πόσο καλά ταιριάζει ένα έγγραφο με το ερώτημα αναζήτησης. Υπολογίζεται με μετρικές όπως το cosine similarity.

7. **Γραφικό Περιβάλλον Χρήστη (GUI) (Graphical User Interface):** Ένα οπτικό περιβάλλον που επιτρέπει στους χρήστες να αλληλεπιδρούν με το σύστημα μέσω κουμπιών, πεδίων εισαγωγής και πινάκων.

Οι παραπάνω πληροφορίες είναι ορισμοί απο τα επίσημα Documentations των τεχνολογειών NLP

1 Εισαγωγή

Το *BookNest* είναι μια desktop-based εφαρμογή που αναπτύχθηκε με στόχο τη δημιουργία μιας μηχανής αναζήτησης για βιβλιοθήκες. Το όνομα της εφαρμογής αντικατοπτρίζει τον σκοπό της ως μια "φωλιά" βιβλίων, όπου οι πληροφορίες οργανώνονται και καθίστανται εύκολα προσβάσιμες. Σχεδιασμένο για να εξυπηρετεί τις ανάγκες αναζήτησης και διαχείρισης βιβλιογραφικών δεδομένων, το σύστημα επιτρέπει την αναζήτηση σε πολλαπλά πεδία, όπως ο τίτλος, ο συγγραφέας και η περιγραφή των βιβλίων, με την χρήση σύγχρονων τεχνολογειών όπως οι μετρικές tf-idf, η επεξεργασία κειμένου και η αξιολόγηση της σχετικότητας μεταξύ των βιβλείων. Παράλληλα, προσφέρει δυνατότητες ταξινόμησης, στατιστικά γραφήματα και στατιστικά στοιχεία για τη βάση δεδομένων.

2 Περιγραφή του Dataset

Το dataset που χρησιμοποιείται στην εργασία είναι το "Goodreads Books 100k", το οποίο διατίθεται μέσω της πλατφόρμας Kaggle και είναι προσβάσιμο στη διεύθυνση Goodreads Books 100k. Αποτελείται από δεδομένα που αφορούν 100.000 βιβλία και περιλαμβάνει τα παρακάτω πεδία :

- **Author**: Το όνομα του συγγραφέα ή των συγγραφέων κάθε βιβλίου.
- **BookFormat**: Το είδος του βιβλίου, π.χ. έντυπο, e-book κ.λπ.
- **Desc**: Μια σύντομη περιγραφή του βιβλίου.
- **Genre**: Τα είδη στα οποία ανήκει το βιβλίο, π.χ. φαντασίας, ιστορικό, επιστημονικό.
- **Img**: Σύνδεσμος για την εικόνα του εξώφυλλου του βιβλίου.
- **ISBN**: Ο διεθνής αναγνωριστικός αριθμός του βιβλίου.
- **ISBN13**: Ο διευρυμένος διεθνής αναγνωριστικός αριθμός του βιβλίου.
- **Link**: Η αντίστοιχη σελίδα του βιβλίου στον ιστότοπο Goodreads.
- **Pages**: Ο συνολικός αριθμός σελίδων του βιβλίου.
- **Rating**: Η μέση βαθμολογία του βιβλίου, όπως έχει προκύψει από τους χρήστες του Goodreads.
- **Reviews**: Ο αριθμός των κριτικών που έχουν καταχωριστεί για το βιβλίο.

- **Title**: Ο τίτλος του βιβλίου.
- **TotalRatings**: Ο συνολικός αριθμός αξιολογήσεων που έχει λάβει το βιβλίο.

Το συγκεκριμένο dataset επιλέχθηκε για την εργασία, καθώς πληρεί όλα τα απαιτούμενα κριτήρια που τέθηκαν. Περιέχει περισσότερες από 100.000 εγγραφές, γεγονός που διασφαλίζει μια μεγάλη ποικιλία δεδομένων για ανάλυση και δοκιμές. Επιπλέον, περιλαμβάνει τόσο πεδία κειμένου, όπως η περιγραφή (**Desc**) και τα είδη (**Genre**), όσο και αριθμητικά και αναγνωριστικά δεδομένα, όπως ο αριθμός σελίδων (**Pages**) και οι κωδικοί ISBN (**ISBN**, **ISBN13**). Η περιγραφή (**Desc**) και οι κατηγορίες (**Genre**) παρέχουν ιδανική βάση για τη χρήση τεχνικών επεξεργασίας φυσικής γλώσσας (NLP), όπως το tokenization, η ανάλυση tf-idf, και η απομάκρυνση stop-words. Επιπλέον, η ύπαρξη αναγνωριστικών και αριθμητικών πεδίων (**Pages**, **Rating**, **TotalRatings**) επιτρέπει την προσθήκη λειτουργιών ταξινόμησης και φίλτρων. Τέλος, η αναγνωρισιμότητα του ιστότοπου Goodreads βοηθούν την ανάπτυξη μιας ρεαλιστικής και λειτουργικής εφαρμογής αναζήτησης.

3 Εργαλεία και Βιβλιοθήκες

Για την ανάπτυξη της εργασίας "BookNest" χρησιμοποιήθηκε η γλώσσα προγραμματισμού **Python**, η οποία επιλέχθηκε λόγω της ευρείας υποστήριξης βιβλιοθηκών και εργαλείων που προσφέρει για την ανάλυση δεδομένων, την επεξεργασία γλώσσας (NLP) και την ανάπτυξη γραφικών περιβαλλόντων χρήστη (GUI). Η ανάπτυξη του κώδικα πραγματοποιήθηκε στο **Visual Studio Code**, ένα από τα πιο δημοφιλή εργαλεία επεξεργασίας κώδικα, το οποίο προσφέρει υποστήριξη για Python μέσω επεκτάσεων και debugging εργαλείων.

Για το σχεδιασμό του γραφικού περιβάλλοντος χρήστη (GUI), χρησιμοποιήθηκε το **Figma**, μια δημοφιλής πλατφόρμα σχεδιασμού που διευκόλυνε τη δημιουργία ενός εύχρηστου και λειτουργικού περιβάλλοντος για την εφαρμογή.

Κατά τη διάρκεια της ανάπτυξης της εφαρμογής, χρησιμοποιήθηκαν οι εξής βιβλιοθήκες Python:

1. **pandas**: Για τη διαχείριση και ανάλυση δεδομένων του dataset.

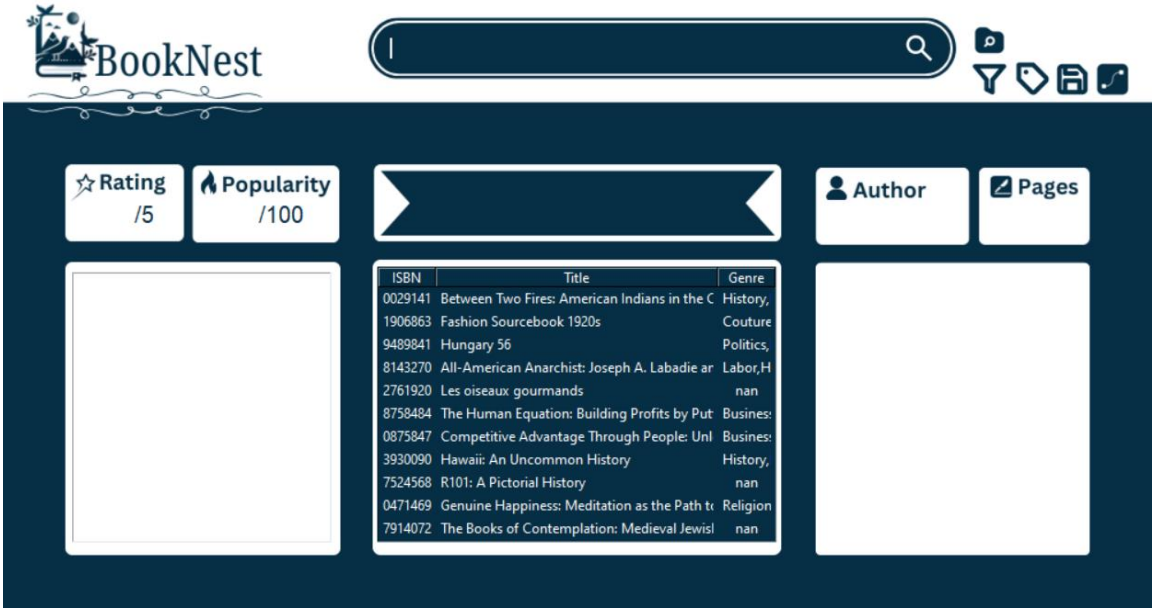
2. **sklearn.feature_extraction.text.TfidfVectorizer**: Για τον υπολογισμό των τιμών **TF-IDF** στις περιγραφές των βιβλίων.
3. **nltk**: Για την επεξεργασία φυσικής γλώσσας, όπως το tokenization, το stemming, το lemmatization, και η απομάκρυνση stop-words.
4. **sklearn.metrics.pairwise.cosine_similarity**: Για τον υπολογισμό της ομοιότητας μεταξύ του ερωτήματος αναζήτησης και των περιγραφών των βιβλίων.
5. **matplotlib**: Για τη δημιουργία γραφικών και την οπτικοποίηση δεδομένων.
6. **wordcloud**: Για τη δημιουργία σύννεφων λέξεων από τις περιγραφές των βιβλίων.
7. **tkinter**: Για την ανάπτυξη του γραφικού περιβάλλοντος χρήστη (GUI).
8. **Pillow (PIL)**: Για τη διαχείριση εικόνων, όπως η φόρτωση και η απεικόνιση εξωφύλλων των βιβλίων.
9. **requests**: Για την απόκτηση δεδομένων εικόνας από URLs.
10. **collections.Counter**: Για τη μέτρηση των εμφανίσεων λέξεων στις περιγραφές.
11. **datetime**: Για την αποθήκευση χρονοσήμανσης σε λειτουργίες αποθήκευσης δεδομένων.
12. **numpy**: Για λειτουργίες αριθμητικής ανάλυσης.

4 Ανάλυση Application

Η εφαρμογή αποτελείται από τέσσερα κύρια αρχεία:

1. **main.py**: Η κεντρική είσοδος της εφαρμογής.
2. **functions.py**: Περιέχει βοηθητικές μεθόδους για τη λειτουργία του συστήματος.
3. **gui.py**: Υλοποιεί τη γραφική διεπαφή χρήστη (GUI) και συνδέει τη μηχανή αναζήτησης με τον χρήστη.
4. **search_engine.py**: Φιλοξενεί την κύρια μηχανή αναζήτησης, υλοποιώντας τεχνικές όπως TF-IDF και Tokenization.

Η διεπαφή χρήστη περιλαμβάνει 6 κουμπιά και παρέχει τέσσερις διαφορετικούς τρόπους αναζήτησης. Ο πίνακας της εφαρμογής απεικονίζει μόνο τα πεδία **ISBN**, **Title**, και **Genre**. Αυτή η επιλογή έγινε για την προστασία δεδομένων και την καλύτερη οπτικοποίηση. Επιπλέον, έχουν αφαιρεθεί βιβλία χωρίς τίτλο ή ISBN, καθώς αυτά θεωρούνται απαραίτητα πεδία.



Αρχηκό στάδιο του Application

Κάθε φορά που ο χρήστης επιλέγει ένα βιβλίο στον πίνακα, ενημερώνονται αυτόματα τα πεδία **Rating**, **Author**, και **Pages**, καθώς και το **Popularity** (x/100), που υπολογίζεται με τη φόρμουλα: **weights = {'rating': 0.3, 'reviews': 0.2, 'totalratings': 0.4, 'pages': 0.1}**. Επιπλέον, εμφανίζεται η εικόνα του εξωφύλλου του βιβλίου που επιλέχτηκε. Σε περίπτωση διπλού κλικ σε ένα βιβλίο, ανοίγει νέο παράθυρο που περιλαμβάνει όλες τις λεπτομέρειες, όπως **author**, **bookformat**, **desc**, **genre**, **img**, **isbn**, **isbn13**, **link**, **pages**, **rating**, **reviews**, **title**, και **totalratings**.

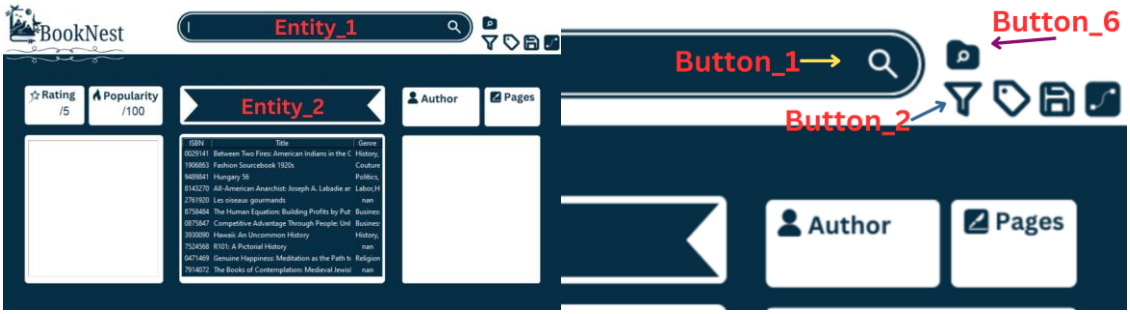
Simple click και double click

Οι διαθέσιμοι τρόποι αναζήτησης περιλαμβάνουν:

1. **Αναζήτηση στον πίνακα**: Ο χρήστης εισάγει δεδομένα στο **entity_1** και πατά το **button_1**. Η αναζήτηση περιορίζεται στα πεδία του πίνακα (**Title**, **Genre**, **ISBN**) και εμφανίζει μόνο τα αντίστοιχα αποτελέσματα.
2. **Απλή αναζήτηση περιγραφής**: Ο χρήστης εισάγει δεδομένα στο **entity_2** και πατά το **button_2**. Η αναζήτηση γίνεται στο πεδίο **Description** της βάσης δεδομένων και εμφανίζονται τα βιβλία που περιέχουν τις λέξεις της εισαγωγής. Το **Description** εμφανίζεται με υπογραμμισμένες τις λέξεις-κλειδιά που ταιριάζουν.
3. **Προηγμένη αναζήτηση**: Χρησιμοποιώντας το **button_6** και το **entity_2**, πραγματοποιείται αναζήτηση με τεχνικές όπως **TF-IDF**, **Tokenization**, **Stop-**

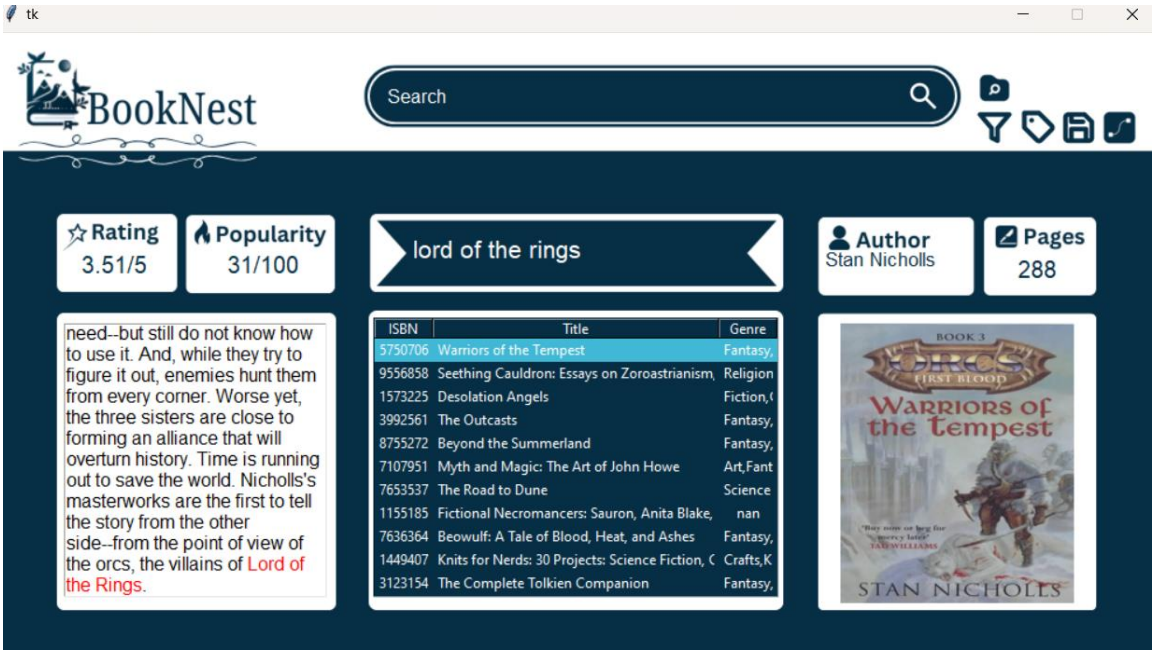
Words Removal, **Stemming**, και **Lemmatization**. Η αναζήτηση καλύπτει τα πεδία **Description**, **Title**, **ISBN**, και **Genre** και επιστρέφει τα 30 πιο συναφή βιβλία με την εισαγωγή. Τα αποτελέσματα εμφανίζονται στον πίνακα, ενώ στο terminal εκτυπώνονται οι συνάφειες ως ποσοστά.

4. **Αναζήτηση βάσει tags**: Ο χρήστης μπορεί να φιλτράρει τα βιβλία με βάση συγκεκριμένες ετικέτες ή κατηγορίες.



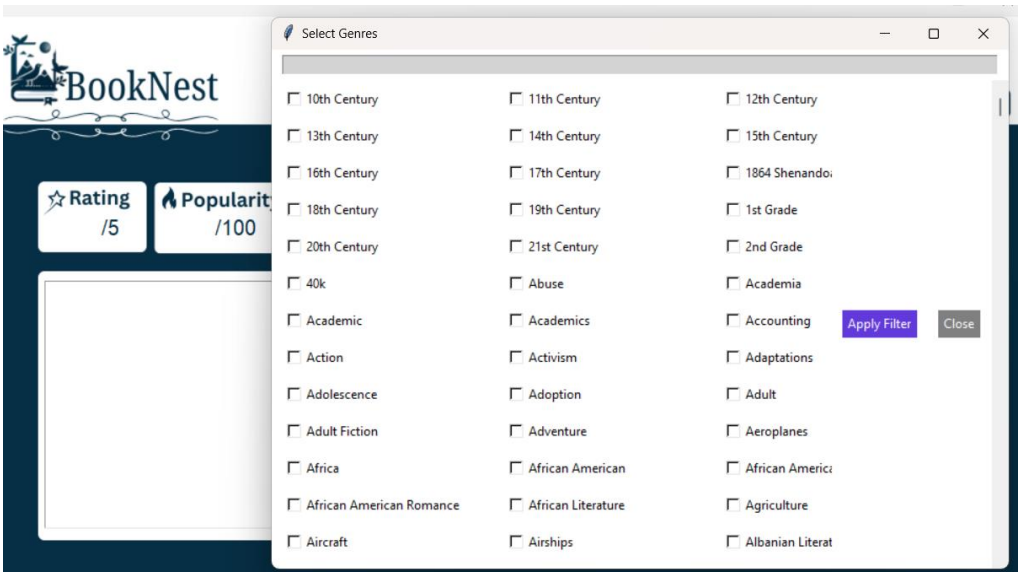

```
PS C:\Users\polih> python C:\Users\polih\OneDrive\Desktop\project\main.py
File loaded successfully.
File loaded successfully.
Building index...
Index built successfully.
Total Documents: 85518
Top Terms (Weighted Scores and Percentages):
book: 967.22 (Weighted Score), 0.14%
life: 826.04 (Weighted Score), 0.12%
new: 808.69 (Weighted Score), 0.12%
world: 700.68 (Weighted Score), 0.10%
story: 645.92 (Weighted Score), 0.10%
time: 577.28 (Weighted Score), 0.09%
love: 504.35 (Weighted Score), 0.08%
year: 504.34 (Weighted Score), 0.08%
work: 499.55 (Weighted Score), 0.07%
history: 485.55 (Weighted Score), 0.07%
Top Relevant Books:
Title: Gravitation, Volume 09, Relevance: 56.54%
Title: Murakami: Ego, Relevance: 49.99%
Title: Haruki Murakami and the Music of Words, Relevance: 39.82%
Title: Godzilla: The Half Century War, Relevance: 15.85%
Title: 999 Frogs Wake Up, Relevance: 15.82%
Title: A Wild Sheep Chase / Dance Dance Dance, Relevance: 14.19%
Title: Coin Locker Babies, Relevance: 12.41%
Title: Forgetting the Art World, Relevance: 12.25%
Title: Granta 124: Travel, Relevance: 11.16%
Title: 100 Artists' Manifestos: From the Futurists to the Stuckists, Relevance: 10.86%
Title: The Islanders, Relevance: 10.19%
Title: The Weird: A Compendium of Strange and Dark Stories, Relevance: 10.14%
Title: The Big New Yorker Book of Cats, Relevance: 10.13%
Title: March Was Made of Yarn, Relevance: 9.91%
Title: Start Here: Read Your Way Into 25 Amazing Authors, Relevance: 7.53%
Title: Lignes, Relevance: 7.14%
Title: History and Repetition, Relevance: 6.86%
Title: Collecting Contemporary, Relevance: 5.32%
```

Εικόνα απο τα αποτελέσματα με Προηγμένη αναζήτηση.

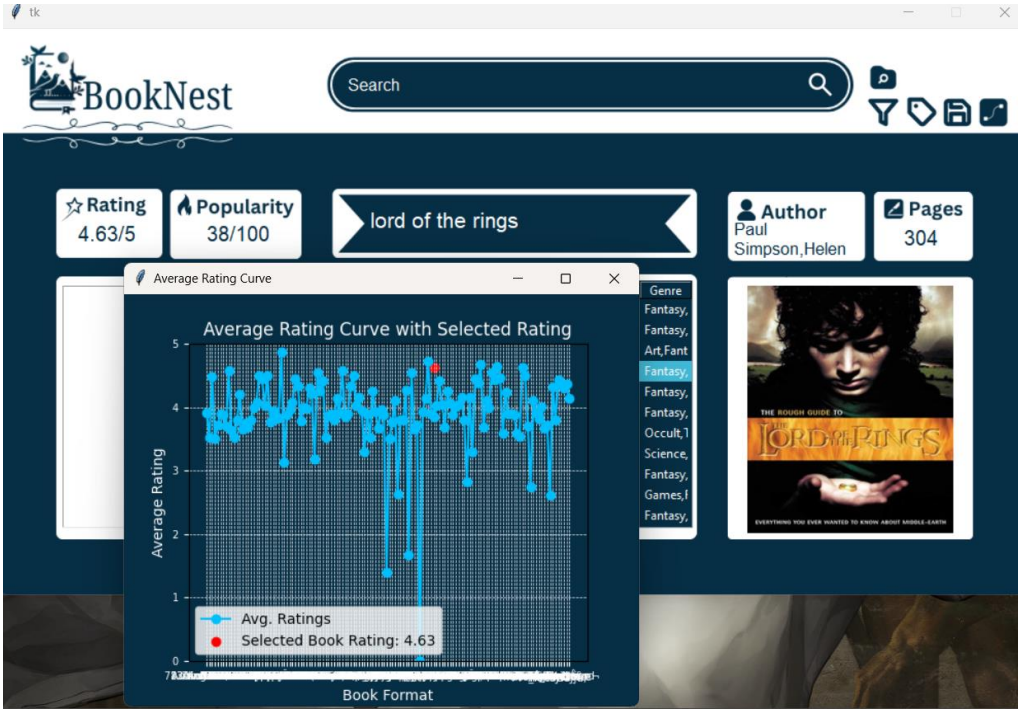


Εικόνα απο τα αποτελέσματα με Απλή αναζήτηση περιγραφής.

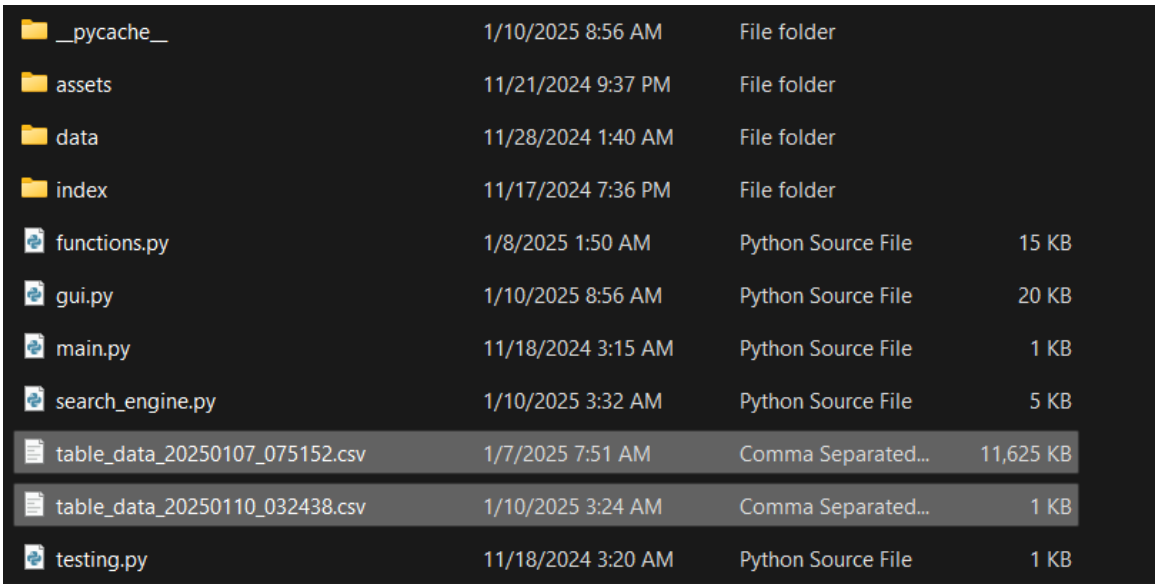
Επιπλέον λειτουργίες περιλαμβάνουν την αποθήκευση του πίνακα σε μορφή CSV, τη δημιουργία γραφημάτων με στατιστικά δεδομένα **Rating**, καθώς και την ταξινόμηση των βιβλίων στον πίνακα. Στο terminal, εκτυπώνονται επίσης πληροφορίες όπως ο συνολικός αριθμός εγγράφων της βάσης δεδομένων (**Total Documents**) και οι πιο συχνές λέξεις του ευρετηρίου.



Αναζητηση με επιλογη tag



Γράφημα Avg_rating



Εικόνα απο τα αποτελέσματα αποθήκευσης πινακα

Η δυνατότητα προβολής αναλυτικών στατιστικών και η δυναμική ενημέρωση των δεδομένων μέσω του πίνακα, ενισχύουν την αλληλεπίδραση με τον χρήστη και παρέχουν

μία ολοκληρωμένη εμπειρία χρήσης. Επιπλέον, οι επιλογές αποθήκευσης και οπτικοποίησης δεδομένων μέσω γραφημάτων προσφέρουν επιπλέον εργαλεία για τη διαχείριση της βιβλιοθήκης.

Το **BookNest** αναδεικνύει τη σημασία της αξιοποίησης τεχνολογιών αιχμής στη διαχείριση δεδομένων και θέτει τις βάσεις για περαιτέρω επεκτάσεις, όπως η ενσωμάτωση περισσότερων δεδομένων ή η βελτιστοποίηση των αλγορίθμων αναζήτησης. Με τη σωστή εφαρμογή του, το σύστημα μπορεί να αποτελέσει ένα ισχυρό εργαλείο για οργανισμούς, βιβλιοθήκες ή ακόμη και μεμονωμένους χρήστες που επιθυμούν να οργανώσουν και να εξερευνήσουν τη συλλογή βιβλίων τους.

Σημαντική βοήθεια κατά την ανάπτυξη της εργασίας ηταν το Claude 3.5 Sonnet extension στο VSC (δυστυχώς δεν δίνεται η δυνατότητα παροχής log αρχείου, αυτός ειναι και ο λόγος παροχής των ερωτημάτων σε μορφή txt file) και του Chatgpt (https://chatgpt.com/share/6780cc91-2b40-8002-b826-3fe928baab96) για την δημιουργία του ερευνικού χαρτιού , τα οποία χρησιμοποιήθηκαν για την επίλυση προβλημάτων, τη διόρθωση σφαλμάτων και τη βελτιστοποίηση του κώδικα.