Πανεπιστήμιο Πειραιώς

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ. «Πληροφοριακά Συστήματα και Υπηρεσίες»

Κατεύθυνση: «Μεγάλα Δεδομένα και Αναλυτική»

Μία προσέγγιση μεταφοράς γνώσης μέσω προ-εκπαιδευμένων μοντέλων σε συστήματα συστάσεων

Διπλωματική Εργασία

Αθανάσιου Ανδριόπουλου

ΜΕ2101

Επιβλέπουσα:

ΜΑΡΙΑ ΧΑΛΚΙΔΗ

Αναπληρώτρια Καθηγήτρια

Ιούλιος 2023

Contents

Ευχαριστίες

Το αντικείμενο με το οποίο ασχολήθηκα ήταν ένα από τα πιο ενδιαφέροντα θέματα , το οποίο μου προκάλεσε από την πρώτη στιγμή την ανάγκη να ψάξω και να μελετήσω για την ευρεία χρήση του σε πολλούς τομείς της πληροφορικής αλλά και της καθημερινότητας γενικότερα.

Για αυτούς και ακόμη παραπάνω λόγους θα ήθελα να ευχαριστήσω την Καθηγήτρια κυρία Μαρία Χαλκίδη η οποία μου έδωσε την δυνατότητα να ασχοληθώ με ένα τόσο ωραίο θέμα και μέσα από το οποίο έμαθα πάρα πολλά πράγματα τα οποία δεν ήξερα ούτε ότι υπήρχαν αλλά ούτε και την χρήση τους στην καθημερινότητα. Επίσης για την συνεχή βοήθεια της ώστε να καταφέρω να φτάσω στην ολοκλήρωση αυτής της διπλωματικής εργασίας μέσα από διευκρινήσεις και την γενική επικοινωνία που είχαμε στο διάστημα της διπλωματικής εργασίας.

Ακόμη θα ήθελα να πω και ένα ακόμη ευχαριστώ στα υπόλοιπα μέλη της επιτροπής της διπλωματικής εργασίας τους Καθηγητές κ. Χρήστο Δουλκερίδη και κ. Μιχαήλ Φιλιππάκη όχι μόνο για την συμμετοχή τους στην εξέταση και συνεπώς στην ολοκλήρωση της εργασίας αλλά και για την αλληλεπίδραση που είχα και μαζί τους σε αυτό το μεταπτυχιακό, για τις διαφορές γνώσης που έλαβα μέσα από τα μαθήματα τους και την διάθεση τους πάντα για βοήθεια και επίλυση αποριών.

Τέλος θα ήθελα να πω και ένα μεγάλο ευχαριστώ στην οικογένεια μου αλλά ειδικότερα στην μητέρα μου όπου με στήριξε και δεν έπαψε να πιστεύει σε εμένα καθ' όλη τη διάρκεια των σπουδών μου.

Περίληψη

Το αντικείμενο αυτής της διπλωματικής εργασίας είναι η καλύτερη κατανόηση της έννοιας μεταφορά γνώσης όχι μόνο σαν μια μέθοδος μηχανικής μάθησης όπου η εφαρμογή της γνώσης που αποκτάται από ένα μοντέλο που χρησιμοποιείται σε μια εργασία μπορεί να επαναχρησιμοποιηθεί ως θεμέλιο για μια άλλη εργασία αλλά και στην εφαρμογή αυτής της γνώσης σε πεδία ενός συστήματος συστάσεων.

Πιο συγκεκριμένα μπορεί ένα σύστημα συστάσεων να διαθέτει πληροφορίες για ένα χρήστη και να μπορεί να του προσφέρει διαφορετικές επιλογές και παράλληλα με την βοήθεια της μεταφοράς γνώσης (transfer learning) να μπορεί να του προτείνει και για άλλα πεδία τα οποία ουδέποτε το σύστημα δεν είχε πληροφορίες για τις προτιμήσεις του χρήστη. Για παράδειγμα θα μπορούσε να δημιουργηθεί ένα συστήματα συστάσεων (Recommended Systems) το οποίο θα πρότεινε στον χρήστη ανάλογα με το είδος των ταινιών που του αρέσουν και τις βαθμολογίες που έχει δώσει και σε άλλες παρεμφερείς ταινίες μία νέα ταινία.

Έπειτα όμως αφού έχει την πληροφορία για αυτά τα πεδία και τον τρόπο με τον οποίο θα μπορούσε αν διαχειριστεί το πεδίο των βαθμολογήσεων για τις ταινίες το μοντέλο μπορεί μέσα από την μεταφορά γνώσης (transfer learning) να χρησιμοποιηθεί και για να προτείνει βιβλία στον χρήστη εφόσον έχουν και αυτά το ίδιο κοινό σημείο που είναι η βαθμολόγηση τους. Έτσι δεν θα δώσει στην τύχη το σύστημα επιλογές στον χρήστη αλλά θα έχει ένα σημείο αναφοράς τις ταινίες ώστε να προσεγγίσει όσο το δυνατόν πιο εύκολα τις προτιμήσεις του χρήστη και για τα βιβλία.

Γενικά υπάρχουν πολλά μοντέλα με τα οποία μπορεί να εκπαιδευτεί το σύστημα ώστε να προσεγγίσει με τον καλύτερο δυνατό τρόπο τα βιβλία για τον χρήστη.

Σε αυτή την εργασία σκοπός είναι να πάρουμε δεδομένα από μια βάση δεδομένων που αφορά ταινίες που προτιμούν διάφοροι χρήστες ,και να μπορέσουμε μέσα από την επεξεργασία των διαφόρων πεδίων τους να προτείνουμε σε χρήστες βιβλία που θα τους αρέσουν.

Οπότε η είσοδος στον αλγόριθμό μας θα είναι κείμενο από τα dataset που διαθέτουν ταινίες και πιο συγκεκριμένα το ml-latest-small και το αρχείο rating με τις βαθμολογίες των ταινιών και δύο csv με βιβλία ενός με τους τίτλους και ενός με τις βαθμολογίες των βιβλίων από τους χρήστες. Πιο αναλυτικά το σύνολο δεδομένων ml-latest-small αποτελείται από αξιολογήσεις ταινιών, ενώ το σύνολο δεδομένων goodbooks-10k αποτελείται από αξιολογήσεις βιβλίων.

Οι δύο εργασίες που θέλουμε να επιτύχουμε είναι:

- Μελέτη τεχνικών μεταφορά γνώσης
- Σχεδιασμός και υλοποίηση μίας προσέγγισης μεταφοράς γνώσης σε συστήματα συστάσεων.

Ο στόχος αφού εκτελέσουμε και τα δύο κομμάτια της εργασίας είναι να υπάρχει ένα μοντέλο που θα έχει δημιουργηθεί για τις ταινίες και όπου με τις διάφορες τεχνικές μεταφοράς να χρησιμοποιήσουμε αυτό το μοντέλο για να βαθμολογήσουμε και στην τελική να προτείνουμε στον χρήστη βιβλία τα οποία δεν έχει βαθμολογήσει.

Πιο αναλυτικά οι ήδη γνωστές αναπαραστάσεις και τα βάρη του προ-εκπαιδευμένου μοντέλου επαναχρησιμοποιούνται ως το σημείο εκκίνησης για τη λεπτομερή ρύθμιση στο νέο σύνολο δεδομένων. Αυτή η διαδικασία επιτρέπει στο μοντέλο να μάθει καλύτερες αναπαραστάσεις των προτιμήσεων των χρηστών και να βελτιώσει την ακρίβεια των προτάσεων.

1. Εισαγωγή

1.1 Σκοπός της Διπλωματικής Εργασίας

Η μάθηση με μεταφορά γνώσης στοχεύει στη βελτίωση της απόδοσης των μαθητών-στόχων σε τομείς πεδία με τη μεταφορά της γνώσης που περιέχεται σε διαφορετικούς αλλά σχετικούς τομείς προέλευσης. Με αυτόν τον τρόπο, η εξάρτηση από έναν μεγάλο αριθμό δεδομένων τομέα πεδίου μπορεί να μειωθεί για την κατασκευή άλλου τομέα. Λόγω των ευρειών προοπτικών εφαρμογής, η μάθηση με μεταφορά γνώσης έχει γίνει ένας δημοφιλής και πολλά υποσχόμενος τομέας στη μηχανική μάθηση.

Σε αυτή την διπλωματική προσπαθήσαμε να δούμε αν με την χρήση προ εκπαιδευμένων μοντέλων μπορούμε να αναπαραστήσουμε την πληροφορία σε ένα πεδίο εφαρμογής, να ορίσουμε τα κοινά στοιχεία σε δύο διαφορετικά πεδία και να μεταφέρουμε γνώση από το ένα πεδίο σε ένα άλλο ώστε να βελτιωθεί η διαδικασία των συστάσεων .

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την ανάπτυξη αυτής της διπλωματικής εργασίας είναι το ml-latest-small και το goodbooks καθώς και τα δύο φέρουν κοινά στοιχεία τα οποία μπορούμε να τα αξιοποιήσουμε από το ένα dataset ώστε να μεταφέρουμε την γνώση στο άλλο dataset όπως προαναφέρθηκε. Στην συγκεκριμένη εργασία χρησιμοποιήθηκε όπως αναφέρεται και παρακάτω το πεδίο ratings και από τα δύο dataset.

Για να επιτευχθεί αυτός ο στόχος η μεταφορά γνώσης (transfer learning) σε αυτόν το πρόβλημα περιλαμβάνει τη χρήση του μοντέλου συλλογικού φιλτραρίσματος (Collaborative filtering) ώστε η γνώση που αποκτήθηκε από το ml-latest-small σύνολο δεδομένων να μπορέσει να χρησιμοποιηθεί στο goodbooks σύνολο δεδομένων.

Στο θεωρητικό κομμάτι ωστόσο δεν αναλύονται μόνο το μοντέλο συλλογικού φιλτραρίσματος αλλά και τα μοντέλα Content-Based Filtering και Hybrid Recommendation Systems όπου το ένα αφορά στο περιεχόμενο και το άλλο στον συνδυασμό του Collaborative filtering με το Content-Base Filtering.

Επίσης όλα τα δεδομένα από όλα τα dataset αναλύονται και επεξεργάζονται ώστε να μπορέσουν να δώσουν το καλύτερο δυνατό αποτέλεσμα με τις βιβλιοθήκες της pandas αλλά και με άλλες βιβλιοθήκες τις python οι οποίες θα αναλυθούν στα επόμενα κεφάλαια.

Τελικός σκοπός αυτής της διπλωματικής είναι με χρήση συγκεκριμένων τεχνικών αξιολογήσεις να δούμε τις τιμές τις οποίες προβλέπει ο κώδικας μας αν οι προβλέψεις του μοντέλου είναι πιο κοντά στις πραγματικές αξιολογήσεις και αν τελικά έγινε πιο ακριβές με την μεταφορά της γνώσης το μοντέλο μας.

1.2 Διάρθρωση της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία θα αναπτυχθεί σε τέσσερα συνολικά κεφάλαια. Και η δομή της εργασίας είναι η εξής:

- Στο Κεφάλαιο 2 παρατίθεται η θεωρητική γνώση για την μεταφορά γνώσης αλλά και για τα συστήματα συστάσεων.
- Στο Κεφάλαιο 3 τα εργαλεία που χρησιμοποιήθηκαν για την επίλυση του προβλήματος καθώς και τα δεδομένα και οι μέθοδοι που θα ακολουθηθούν.
- Στο Κεφάλαιο 4 υπάρχουν διαγράμματα και αποτελέσματα από την χρήση αυτών των δεδομένων και τελικά οι τελικές προτάσεις των συστημάτων των συστάσεων για τον χρήστη.

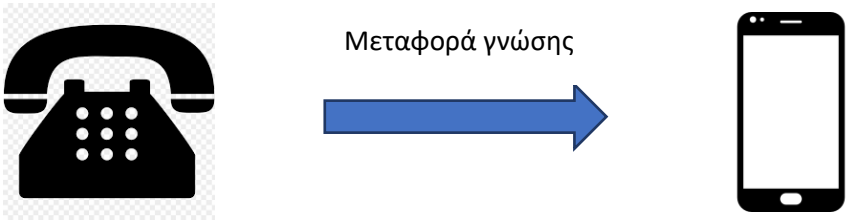
Ακόμη περιέχει και μετά τα 4 κύρια κεφάλαια ειδικές ενότητες οι οποίες είναι για τα:

- Συμπεράσματα όπου γίνεται μια τελική αναφορά στην εργασία και έπειτα χρήσεις της και εξέλιξης της.
- Πίνακας ανάλυσης συμβόλων όπου επεξηγούνται όλα τα σύμβολα που υπάρχουν μέσα στην διπλωματική εργασία.
- Ένας πίνακας από όλες τις εικόνες που έχουν δημιουργηθεί για να βοηθήσουν στην κατανόηση των διαφορετικών κομματιών που αναλύονται μέσα στην διπλωματική εργασία.
- Βιβλιογραφία όπου είναι επίσης ένας πίνακας με τις βιβλιογραφικές αναφορές για την διπλωματική εργασία.

2. Θεωρητικό Υπόβαθρο

2.1 Μεταφορά γνώσης

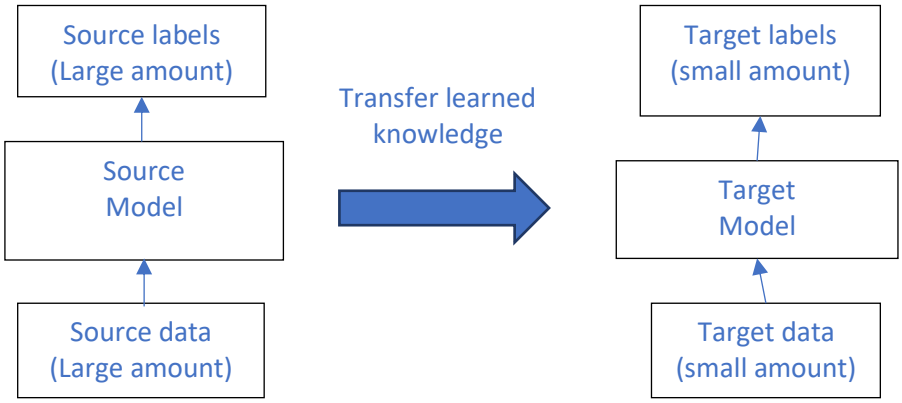
Η μεταφορά γνώσης είναι μια μέθοδος μηχανικής μάθησης όπου επαναχρησιμοποιούμε ένα προ εκπαιδευμένο μοντέλο ως σημείο εκκίνησης για ένα μοντέλο σε μια νέα εργασία. Η ιδέα σχετικά με τη μεταφορά μάθησης μπορεί αρχικά να προέρχεται από την ψυχολογία. Σύμφωνα με την γενική θεωρία της μεταφοράς γνώσης, όπως προτείνεται από τον ψυχολόγο C.H. Τζαντ, να ξέρεις πως να μεταφέρεις γνώση είναι το αποτέλεσμα της γενίκευσης της εμπειρίας και είναι δυνατό να πραγματοποιηθεί η μεταφορά από τη μια κατάσταση στην άλλη, όσο ένα άτομο γενικεύει την εμπειρία του. [1] Η έννοια της μεταφοράς γνώσης μπορεί αν γίνει και πιο κατανοητή από την παρακάτω εικόνα ().



Εικόνα 1: Μεταφορά γνώσης

Με πιο απλά λόγια ένα μοντέλο που έχει εκπαιδευτεί σε μια εργασία επανατοποθετείται σε μια δεύτερη, σχετική εργασία ως βελτιστοποίηση που επιτρέπει ταχεία πρόοδο κατά τη μοντελοποίηση της δεύτερης εργασίας. Εφαρμόζοντας τη μάθηση μεταφοράς σε μια νέα εργασία, μπορεί κανείς να επιτύχει σημαντικά υψηλότερες επιδόσεις από την εκπαίδευση με μικρό μόνο όγκο δεδομένων.

Η εκμάθηση με μεταφορά γνώσης είναι τόσο συνηθισμένη που είναι σπάνιο να εκπαιδευτεί ένα μοντέλο για εργασίες που σχετίζονται με την επεξεργασία εικόνας ή φυσικής γλώσσας από την αρχή. Αντίθετα, οι ερευνητές και οι επιστήμονες που ασχολούνται με αυτό το κομμάτι του machine learning προτιμούν να ξεκινήσουν από ένα προ εκπαιδευμένο μοντέλο που ξέρει ήδη πώς να ταξινομεί αντικείμενα και έχει μάθει γενικά χαρακτηριστικά όπως άκρες, σχήματα σε εικόνες.



Εικόνα 2:Μεταφορά γνώσης 2

Παρακάτω δίνεται ο ορισμός της μεταφοράς γνώσης.

Ορισμός: Ο ορισμός της μεταφοράς γνώσης δίνεται ως προς τους τομείς και τις εργασίες. Ένας τομέας D αποτελείται από ένα χώρο χαρακτηριστικών X και μία οριακή κατανομή πιθανοτήτων P(X) όπου $X=\{\chi_1,\chi_2\ldots\ldots,\chi_n\} \epsilon$ X. Δεδομένου ενός συγκεκριμένου τομέα D={ X,P(X)}, μία εργασία αποτελείται από δύο στοιχεία ένα χώρο ετικετών Y και μία αντικειμενική προγνωστική συνάρτηση f: X→y. Η λειτουργία f χρησιμοποιείται για την πρόβλεψη της αντίστοιχης ετικέτας f(x) για μία νέα περίπτωση χ . Αυτή η εργασία υποδηλώνεται με T= {y,f(x)} , η οποία δεν παρατηρείται αλλά μπορεί να μαθευτεί από τα εκπαιδευμένα δεδομένα , τα οποία αποτελούνται από ζεύγη $\{\chi_i,y_i\}$;όπου $\chi_i \epsilon$ X και $y_i \epsilon$ Y. Η συνάρτηση f() μπορεί να χρησιμοποιηθεί για να προβλέψει την αντίστοιχη ετικέτα f(x).

Για να γίνει καλύτερα αντιληπτός όμως ο ορισμός είναι καλό να αναλυθούν και οι όροι τομέας και εργασία.

Ορισμός: Ένας τομέας (domain) D αποτελείται από δύο μέρη: έναν χώρο χαρακτηριστικών X και την οριακή κατανομή $P(X)$. Με άλλα λόγια, $D = X, P(X)$ και το σύμβολο X υποδηλώνει ένα σύνολο περιπτώσεων, που ορίζεται ως $X = \{x \mid x_i \in X, i = 1,\ldots,n\}$. [2]

Γενικά εφαρμόζονται διαφορετικές στρατηγικές και τεχνικές μάθησης μεταφοράς με βάση τον τομέα της εφαρμογής, την εργασία που εκτελείται και τη διαθεσιμότητα των δεδομένων.

Πριν αποφασίσετε για τη στρατηγική της μεταφοράς μάθησης, είναι σημαντικό να έχετε μια απάντηση στις ακόλουθες ερωτήσεις:

- Ποιο μέρος της γνώσης μπορεί να μεταφερθεί από την πηγή στον στόχο για τη βελτίωση της απόδοσης της εργασίας-στόχου;
- Πότε να μεταφέρετε και πότε όχι, ώστε να βελτιώσετε την απόδοση/αποτελέσματα της εργασίας-στόχου και να μην τα υποβαθμίσετε;
- Πώς να μεταφέρετε τη γνώση που αποκτήθηκε από το μοντέλο πηγής με βάση τον τρέχοντα τομέα/εργασία μας;

Παραδοσιακά, οι στρατηγικές μάθησης μεταφοράς γνώσης εμπίπτουν σε τρεις μεγάλες κατηγορίες ανάλογα με τον τομέα εργασίας και την ποσότητα των δεδομένων με ετικέτα ή χωρίς ετικέτα που υπάρχουν και οι οποίες περιγράφονται παρακάτω πιο αναλυτικά .

- Επαγωγική Εκμάθηση Μεταφοράς γνώσης(**Inductive Transfer Learning**,): Η Εκμάθηση Επαγωγικής Μεταφοράς γνώσης απαιτεί οι τομείς προέλευσης και στόχος να είναι οι ίδιοι, αν και οι συγκεκριμένες εργασίες στις οποίες εργάζεται το μοντέλο είναι διαφορετικές. Οι αλγόριθμοι προσπαθούν να χρησιμοποιήσουν τη γνώση από το μοντέλο προέλευσης και να την εφαρμόσουν για να βελτιώσουν την εργασία-στόχο. Το προεκπαιδευμένο μοντέλο διαθέτει ήδη τεχνογνωσία στα χαρακτηριστικά του τομέα και βρίσκεται σε καλύτερη αφετηρία από ό,τι αν το εκπαιδεύαμε από την αρχή. Η επαγωγική μάθηση μεταφοράς γνώσης χωρίζεται περαιτέρω σε δύο υποκατηγορίες ανάλογα με το αν ο τομέας προέλευσης περιέχει δεδομένα με ετικέτα ή όχι. Αυτές περιλαμβάνουν τη μάθηση πολλαπλών εργασιών και την αυτοδιδάχθη, αντίστοιχα.
- **(Transductive Transfer Learning)**: Σενάρια όπου οι τομείς των εργασιών πηγής και στόχου δεν είναι ακριβώς οι ίδιοι αλλά αλληλένδετες χρησιμοποιούν τη στρατηγική Transductive Transfer Learning. Μπορεί κανείς να αντλήσει ομοιότητες μεταξύ των εργασιών πηγής και στόχου. Αυτά τα σενάρια έχουν συνήθως πολλά δεδομένα με ετικέτα στον τομέα προέλευσης, ενώ ο τομέας προορισμού έχει μόνο δεδομένα χωρίς ετικέτα.
- Η μάθηση χωρίς επίβλεψη μεταφοράς(**Unsupervised Transfer Learning**) είναι παρόμοια με την επαγωγική μάθηση μεταφοράς. Η μόνη διαφορά είναι ότι οι αλγόριθμοι επικεντρώνονται σε εργασίες χωρίς επίβλεψη και περιλαμβάνουν σύνολα δεδομένων χωρίς ετικέτα τόσο στις εργασίες προέλευσης όσο και στις εργασίες προορισμού.

Τώρα, θα αναλυθεί ένας άλλος τρόπος κατηγοριοποίησης των στρατηγικών μάθησης μεταφοράς γνώσης με βάση την ομοιότητα του τομέα και ανεξάρτητα από τον τύπο των δειγμάτων δεδομένων που υπάρχουν για εκπαίδευση.

- **Homogeneous Transfer Learning**: Οι προσεγγίσεις μάθησης ομογενούς μεταφοράς αναπτύσσονται και προτείνονται για να χειριστούν καταστάσεις όπου οι τομείς είναι του ίδιου χώρου χαρακτηριστικών. Στη

μάθηση ομογενούς μεταφοράς, οι τομείς έχουν μόνο μια μικρή διαφορά στις οριακές κατανομές. Αυτές οι προσεγγίσεις προσαρμόζουν τους τομείς διορθώνοντας τη μεροληψία επιλογής δείγματος ή τη μετατόπιση συμμεταβλητών.

- **Instance transfer:** Καλύπτει ένα απλό σενάριο στο οποίο υπάρχει μεγάλος αριθμός δεδομένων με ετικέτα στον τομέα προέλευσης και περιορισμένος αριθμός στον τομέα προορισμού. Τόσο οι τομείς όσο και οι χώροι χαρακτηριστικών διαφέρουν μόνο ως προς τις οριακές κατανομές.

- **Parameter transfer**: Η ιδέα πίσω από τις μεθόδους που βασίζονται σε παραμέτρους είναι ότι ένα καλά εκπαιδευμένο μοντέλο στον τομέα προέλευσης έχει μάθει μια καλά καθορισμένη δομή και εάν σχετίζονται δύο εργασίες, αυτή η δομή μπορεί να μεταφερθεί στο μοντέλο-στόχο. Σε γενικές γραμμές, υπάρχουν δύο τρόποι για να μοιραστείτε τα βάρη στα μοντέλα βαθιάς μάθησης: soft sharing και hard weight sharing.Στην κοινή κατανομή βάρους, το μοντέλο αναμένεται να είναι κοντά στα ήδη μαθημένα χαρακτηριστικά και συνήθως τιμωρείται εάν τα βάρη του αποκλίνουν σημαντικά από ένα δεδομένο σύνολο βαρών. Στην κοινή χρήση σκληρού βάρους, μοιραζόμαστε τα ακριβή βάρη μεταξύ διαφορετικών μοντέλων.

- **Relational-knowledge transfer:** Οι προσεγγίσεις μάθησης μεταφοράς με βάση τις σχέσεις επικεντρώνονται κυρίως στην εκμάθηση των σχέσεων μεταξύ της πηγής και ενός τομέα στόχου και στη χρήση αυτής της γνώσης για την εξαγωγή προηγούμενης γνώσης και τη χρήση της στο τρέχον πλαίσιο.

- **Heterogeneous Transfer Learning**: Η μάθηση μεταφοράς περιλαμβάνει την εξαγωγή αναπαραστάσεων από ένα προηγούμενο δίκτυο για την εξαγωγή σημαντικών χαρακτηριστικών από νέα δείγματα για μια αλληλοσχετιζόμενη εργασία. Ωστόσο, αυτές οι προσεγγίσεις ξεχνούν να λάβουν υπόψη τη διαφορά στους χώρους χαρακτηριστικών μεταξύ του τομέα προέλευσης και προορισμού. Αυτή η τεχνική στοχεύει να λύσει το ζήτημα των τομέων προέλευσης και προορισμού που έχουν διαφορετικούς χώρους χαρακτηριστικών και άλλες ανησυχίες, όπως διαφορετικές διανομές δεδομένων και χώρους ετικετών. Η ετερογενής μάθηση μεταφοράς εφαρμόζεται σε εργασίες μεταξύ τομέων, όπως η κατηγοριοποίηση κειμένου μεταξύ γλωσσών, η ταξινόμηση από κείμενο σε εικόνα και πολλές άλλες.

2.2 Έρευνες για μεταφορά γνώσης

Για την ανασκόπηση προηγούμενων εργασιών για τη μεταφορά μάθησης, χρησιμοποιούνται οι κατηγοριοποιήσεις της προηγούμενης ενότητας.

Πρώτον, οι σχετικές εργασίες εξετάζονται σε σχέση με τις ρυθμίσεις που βασίζονται σε ετικέτες και, δεύτερον, εξετάζονται σε βάθος οι ερμηνείες που βασίζονται σε δεδομένα και τα μοντέλα.

Στο πλαίσιο της επαγωγικής μάθησης μεταφοράς, έχουν δοκιμαστεί και οι τέσσερις τύποι προσεγγίσεων ως προς το «τι να μεταφέρω».

Στη ρύθμιση εκμάθησης μεταβιβαστικής μεταφοράς, έχουν δοκιμαστεί μόνο προσεγγίσεις μεταφοράς αναπαράστασης στιγμιότυπων και χαρακτηριστικών.

Τέλος, στη ρύθμιση εκμάθησης μεταφοράς χωρίς επίβλεψη, έχουν δοκιμαστεί μόνο προσεγγίσεις αναπαράστασης χαρακτηριστικών. [3]

Instance Weighting: Αυτή μέθοδος εφαρμόζει πολλά ευρετικά προσαρμογής με μια ενοποιημένη αντικειμενική συνάρτηση: (1) αφαίρεση παραπλανητικών περιπτώσεων εκπαίδευσης στην πηγή τομέα; (2) εκχώρηση περισσότερων βαρών σε εμφανίσεις-στόχους με ετικέτα από ό,τι σε εμφανίσεις πηγής με ετικέτα. (3) αύξηση των περιπτώσεων εκπαίδευσης με στιγμιότυπα-στόχους με προβλεπόμενες ετικέτες. [4]

Feature Transformation : Αυτή η προσέγγιση ανακαλύπτει υποκείμενες ουσιαστικές δομές μετατρέποντας και τους δύο τομείς σε έναν κοινό χώρο λανθάνοντος χαρακτηριστικού - συνήθως χαμηλής διάστασης - που έχει προγνωστικές ιδιότητες ενώ μειώνει την οριακή κατανομή μεταξύ των τομέων [5]. Αν και ο στόχος υψηλού επιπέδου πίσω από αυτές τις μεθόδους (βελτίωση της απόδοσης ενός εκπαιδευόμενου-στόχου) είναι πολύ διαφορετικός από τον στόχο στη μάθηση αναπαράστασης, η ιδέα πίσω από αυτές είναι αρκετά στενή. [6]

Feature mapping: είναι η περαιτέρω βελτίωση της απόδοσης τμηματοποίησης στιγμιότυπων μικρών αντικειμένων μέσω εκμάθησης μεταφοράς που βασίζεται σε χαρτογράφηση χαρακτηριστικών. Δεν μπορούμε να εκπαιδεύσουμε ένα συμβατικό δίκτυο τμηματοποίησης παρουσιών οπότε τμηματοποιούμε μικρά αντικείμενα λόγω έλλειψης μάσκας, δηλαδή ετικετών σε επίπεδο pixel, καθώς ένα μεγάλο ποσοστό μάσκας μικρών αντικειμένων που δημιουργούνται είναι άκυρα όπως συζητήθηκε προηγουμένως. Ωστόσο, μπορούμε να εκπαιδεύσουμε ένα δίκτυο ανίχνευσης αντικειμένων για να ανιχνεύει κάθε αντικείμενο για δεδομένες άφθονες ετικέτες κατηγορίας και πληροφορίες οριοθέτησης. Έτσι, μια διαισθητική λύση για τη βελτίωση της απόδοσης τμηματοποίησης των μικρών αντικειμένων είναι η δημιουργία γέφυρας μεταξύ της πρόβλεψης του

πλαισίου οριοθέτησης και της ανίχνευσης μάσκας. [7]

Προσαρμογή τομέα μέσω ανάλυσης στοιχείων μεταφοράς: Είναι μία νέα μέθοδος εκμάθησης, ανάλυση συνιστωσών μεταφοράς (TCA), για προσαρμογή τομέα. Το TCA προσπαθεί να μάθει ορισμένα στοιχεία μεταφοράς σε τομείς σε έναν αναπαραγόμενο χώρο του πυρήνα Hilbert χρησιμοποιώντας τη μέγιστη μέση αναντιστοιχία. Στον υποχώρο που εκτείνεται από αυτά τα στοιχεία μεταφοράς, οι ιδιότητες δεδομένων διατηρούνται και οι διανομές δεδομένων σε διαφορετικούς τομείς είναι κοντά η μία στην άλλη. Ως αποτέλεσμα, με τις νέες αναπαραστάσεις σε αυτόν τον υποχώρο, μπορούμε να εφαρμόσουμε τυπικές μεθόδους μηχανικής εκμάθησης για να εκπαιδεύσουμε ταξινομητές ή μοντέλα παλινδρόμησης στον τομέα προέλευσης για χρήση στον τομέα προορισμού. Επιπλέον, προκειμένου να αποκαλυφθεί η γνώση που κρύβεται στις σχέσεις μεταξύ των ετικετών δεδομένων από τους τομείς προέλευσης και προορισμού, επεκτείνουμε το TCA σε μια ημιεποπτευόμενη ρύθμιση εκμάθησης, η οποία κωδικοποιεί τις πληροφορίες ετικετών σε εκμάθηση στοιχείων μεταφοράς. Ονομάζουμε αυτή την επέκταση ημιεποπτευόμενη TCA. [8]

Clustering : Για το κομμάτι του clustering υπάρχουν αρκετές έρευνες όπως ένας αλγόριθμο ταξινόμησης με βάση τη συσσώρευση (CoCC). Η από κοινού ομαδοποίηση χρησιμοποιείται ως γέφυρα για τη διάδοση της δομής και της γνώσης της τάξης από τον εσωτερικό τομέα προς τον εξωτερικό τομέα. Μία άλλη είναι η αυτοδίδακτη ομαδοποίηση Self-Taught Clustering (STC) είναι μια περίπτωση μεταφοράς χωρίς επίβλεψη μάθηση, η οποία στοχεύει στη ομαδοποίηση μιας μικρής συλλογής δεδομένων χωρίς ετικέτα στόχου με τη βοήθεια μεγάλου όγκου βοηθητικών δεδομένων χωρίς ετικέτα.

Feature selection: Οι μέθοδοι επιλογής χαρακτηριστικών που βασίζονται σε στατιστικά περιλαμβάνουν την αξιολόγηση της σχέσης μεταξύ κάθε μεταβλητής εισόδου και της μεταβλητής στόχου χρησιμοποιώντας στατιστικά και την επιλογή εκείνων των μεταβλητών εισόδου που έχουν την ισχυρότερη σχέση με τη μεταβλητή στόχο. Αυτές οι μέθοδοι μπορεί να είναι γρήγορες και αποτελεσματικές, αν και η επιλογή των στατιστικών μέτρων εξαρτάται από τον τύπο δεδομένων τόσο των μεταβλητών εισόδου όσο και των μεταβλητών εξόδου. Μία έρευνα είναι βασισμένη στην Structural Correspondence Learning (SCL) από τον Blitzer. [9]

Feature Alignment: Η εργασία του Pan [10]προτείνει μια μάθηση μεταφοράς φασματικής ευθυγράμμισης χαρακτηριστικών (SFA), ένας αλγόριθμος που ανακαλύπτει μια νέα αναπαράσταση χαρακτηριστικών για τον τομέα

προέλευσης και προορισμού για την επίλυση των διαφορών οριακής κατανομής. Η μέθοδος SFA προϋποθέτει αφθονία επισημασμένων δεδομένων προέλευσης και περιορισμένου αριθμού δεδομένων στόχων με ετικέτα. Η προσέγγιση SFA προσδιορίζει χαρακτηριστικά για συγκεκριμένο και ανεξάρτητο τομέα και χρησιμοποιεί τα χαρακτηριστικά ανεξάρτητα από τον τομέα ως γέφυρα για τη δημιουργία ενός διμερούς γραφήματος που μοντελοποιεί τη σχέση συν-εμφάνισης μεταξύ των ανεξάρτητων από τον τομέα και των χαρακτηριστικών του τομέα. Αν το γράφημα εμφανίζει δύο χαρακτηριστικά του τομέα που έχουν συνδέσεις με ένα κοινό χαρακτηριστικό ανεξάρτητο από τον τομέα, τότε υπάρχει μεγαλύτερη πιθανότητα να ευθυγραμμιστούν τα χαρακτηριστικά του συγκεκριμένου τομέα. [11]

Parameter sharing: Η κοινή χρήση παραμέτρων λαμβάνει χώρα όταν δημιουργείται ένας χάρτης χαρακτηριστικών από το αποτέλεσμα της συνέλιξης μεταξύ ενός φίλτρου και των δεδομένων εισόδου από μια μονάδα εντός ενός επιπέδου στο επίπεδο μετατροπής. Όλες οι μονάδες σε αυτό το επίπεδο στρώματος μοιράζονται τα ίδια βάρη. γι' αυτό ονομάζεται κατανομή βάρους/παραμέτρου. Μία έρευνα έγινε από τον Zhuang [12] πρότεινε το Matrix TriFactorization Based Classification Framework (MTrick) για ταξινόμηση κειμένου.

Model Ensemble : Η εκμάθηση συνόλου είναι η εκπαίδευση πολλαπλών μοντέλων αντί για ένα μόνο μοντέλο και ο συνδυασμός των προβλέψεων από αυτά τα μοντέλα. Αυτό μειώνει τη διακύμανση των προβλέψεων και μειώνει τα σφάλματα γενίκευσης. Τα αποτελέσματα είναι προβλέψεις ότι αυτό είναι καλύτερο από οποιοδήποτε μεμονωμένο μοντέλο. Ένα παράδειγμα είναι του Gao [13]όπου που εστιάζει στη διαδικασία συνόλου των διαφόρων μαθητών, που θα μπορούσαν να κατασκευαστούν σε διαφορετικούς τομείς πηγών ή να δημιουργηθούν με εκτέλεση διαφορετικών αλγόριθμοων εκμάθησης σε έναν τομέα πηγής. Στο LWE συνήθως ανατίθεται ένας μαθητής με διαφορετικά βάρη κατά την ταξινόμηση διαφορετικών παρουσιών τομέα στόχου. [3]

Deep Learning Techniques: Ένα παράδειγμα είναι εργασία του Glorot [14] προτείνει έναν αλγόριθμο βαθιάς μάθησης για τη μεταφορά μάθησης ονομάζεται αυτόματος κωδικοποιητής στοίβαξης αποθορυβοποίησης (SDA) για την επίλυση των διαφορών οριακής κατανομής μεταξύ ενός τομέα προέλευσης με ετικέτα και ενός τομέα προορισμού χωρίς ετικέτα. Οι αλγόριθμοι βαθιάς μάθησης μαθαίνουν ενδιάμεσες αμετάβλητες έννοιες μεταξύ δύο πηγών δεδομένων, οι οποίες χρησιμοποιούνται για την εύρεση ενός κοινού λανθάνοντος συνόλου χαρακτηριστικών. Υπάρχουν προφανώς και πολλές ακόμα έρευνες πάνω στο deep learning κομμάτι καθώς έχουν

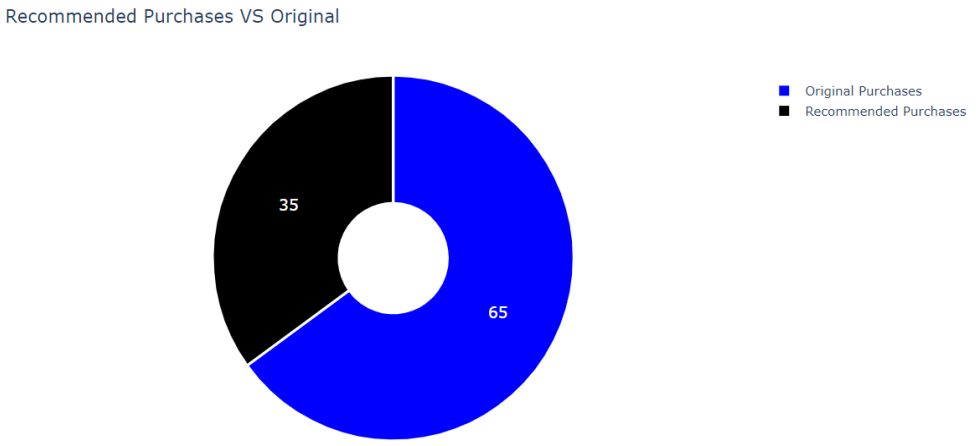
αναπτυχθεί και πάρα πολλοί αλγόριθμοι γύρω από αυτών εδώ αναλύθηκε ένας προηγουμένως.

Τέλος υπάρχει και μία ακόμη τεχνική η **Relational Approaches** όπως αυτή της Markov Logic Networks (MLNs) τα MLN είναι ένας ισχυρός φορμαλισμός που συνδυάζει την εκφραστικότητα της λογικής πρώτης τάξης με την ευελιξία της πιθανότητας .Υπάρχουν δύο πτυχές για την εκμάθηση ενός MLN—η δομή και τα βάρη. Ενώ η μάθηση με βάρη είναι σχετικά γρήγορη, η μάθηση δομών είναι πολύ υπολογιστικά έντονη.

2.3 Συστήματα Συστάσεων

Σε αυτό το σημείο θα αναλυθεί η έννοια συστήματα συστάσεων και ο τρόπος με τον οποίο αυτά χρησιμοποιούνται στην καθημερινότητα.

Τα συστήματα συστάσεων έχουν γίνει ένα σημαντικό χαρακτηριστικό σε σύγχρονους ιστότοπους, π.χ. στο Amazon, Netflix ή Flickr. Τα ποσοστά, τα έσοδα και άλλα μέτρα επιτυχίας μπορεί να αυξηθούν με την εφαρμογή αποτελεσματικών συστημάτων συστάσεων. Το δύσκολο έργο είναι να προσδιοριστούν σχετικά στοιχεία ακόμα κι αν είναι γενικά μη δημοφιλής. Τα συστήματα συστάσεων αξιοποιούν το διαθέσιμο περιβάλλον, όπως πληροφορίες χρήστη, ώρα, τοποθεσία κ.λπ. για να φιλτράρουν σχετικά στοιχεία. [15] Εξ ορισμού, ένα σύστημα προτάσεων είναι ένα σύστημα που προσδιορίζει και παρέχει προτεινόμενο περιεχόμενο ή ψηφιακά στοιχεία για τους χρήστες χρησιμοποιώντας τα ενδιαφέροντα των χρηστών.



Εικόνα 3: Recommended purchases vs Original

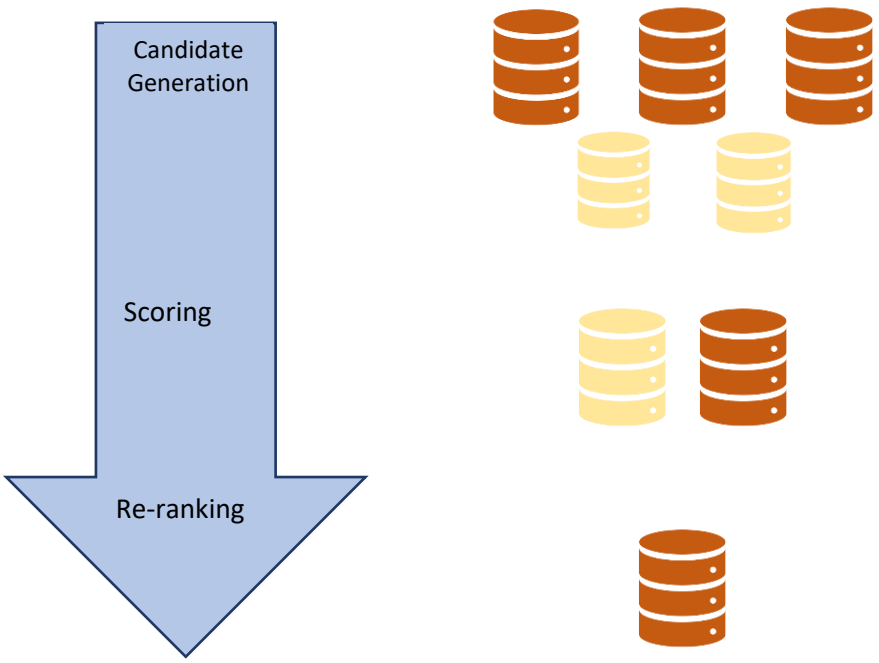
Αρχιτεκτονική Συστημάτων Συστάσεων:

Δημιουργία υποψηφίων: Συνήθως είναι διαθέσιμο ένα μεγάλο σύνολο δεδομένων, αλλά έχει μικρότερο αριθμό υποψηφίων ή κατηγοριών. Η δημιουργία και η διαφοροποίηση αυτών των κατηγοριών είναι το πρώτο βήμα για τη δημιουργία του συστήματος συστάσεων.

Βαθμολογία: Η βαθμολόγηση είναι το μοντέλο που δημιουργεί υποσύνολα και κατατάσσει τις κατηγορίες που θα εμφανιστούν στον χρήστη. Η προσαρμογή της βαθμολόγησης των κατηγοριών μπορεί να γίνει χρησιμοποιώντας συνθήκες ή ερωτήματα.

Επανακατάταξη: Το τελευταίο βήμα, η ανακατάταξη όπου λαμβάνει υπόψη όλους τους διαθέσιμους περιορισμούς και αφαιρεί τις ανεπιθύμητες ή τις συστάσεις με χαμηλή βαθμολογία. Η ανακατάταξη είναι εξαιρετικά

σημαντική καθώς φιλτράρει όλα τα περιττά στοιχεία.



Εικόνα 4: Αρχιτεκτονική ενός συστήματος συστάσεων

Προκειμένου να υλοποιήσει τη βασική του λειτουργία, προσδιορίζοντας τα χρήσιμα στοιχεία για τον χρήστη, ένα RS πρέπει να προβλέψει ότι αξίζει να προτείνετε ένα αντικείμενο. Για να γίνει αυτό, το σύστημα πρέπει να είναι σε θέση να προβλέψει τη χρησιμότητα ορισμένων αντικειμένων, ή τουλάχιστον να συγκρίνει τη χρησιμότητα τους ως προς ορισμένα στοιχεία και, στη συνέχεια, να αποφασίσει ποια στοιχεία θα προτείνετε με βάση αυτήν τη σύγκριση. [16]

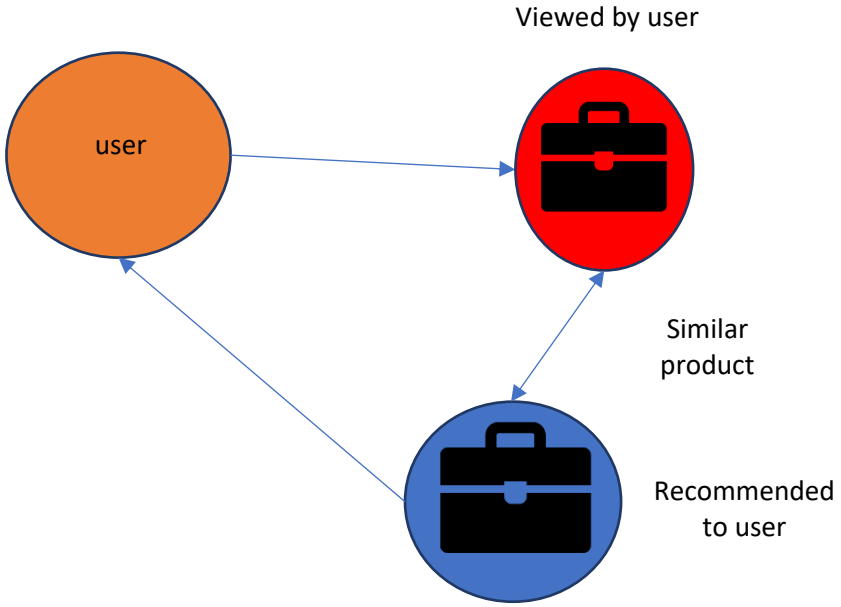
Γενικά υπάρχουν τρεις σημαντικοί τύποι συστημάτων συστάσεων οι οποίοι είναι οι Collaborative filtering , Content-Based Filtering, Hybrid Recommendation Systems.

Για να γίνει όμως καλύτερα αντιληπτά αυτά τα RS μια πρώτη επισκόπηση των διαφορετικών τύπων RS, είναι καλύτερο να παρουσιαστεί με μία ταξινόμηση που παρέχεται από το [17] που έχει γίνει ένας κλασικός τρόπος διάκρισης μεταξύ συστημάτων συστάσεων και αναφοράς σε αυτά. [17] και το οποίο κάνει διάκριση μεταξύ έξι διαφορετικών κατηγοριών για την προσέγγιση συστημάτων συστάσεων:

Content-based filtering: Αυτή η μέθοδος φιλτραρίσματος βασίζεται στην περιγραφή ενός στοιχείου και στο προφίλ των προτιμώμενων επιλογών του χρήστη. Σε ένα σύστημα συστάσεων που βασίζεται σε περιεχόμενο, οι λέξεις-κλειδιά χρησιμοποιούνται για την περιγραφή των στοιχείων. Η ιδέα του φιλτραρίσματος βάσει περιεχομένου είναι ότι αν σας αρέσει ένα αντικείμενο

θα σας αρέσει και ένα «παρόμοιο». Για παράδειγμα, όταν προτείνουμε το ίδιο είδος αντικειμένου όπως μια πρόταση ταινίας ή τραγουδιού. Αυτή η προσέγγιση έχει τις ρίζες της στην έρευνα ανάκτησης πληροφοριών και φιλτραρίσματος πληροφοριών.

Ένα σημαντικό ζήτημα με το φιλτράρισμα βάσει περιεχομένου είναι εάν το σύστημα μπορεί να μάθει τις προτιμήσεις των χρηστών από τις ενέργειες των χρηστών σχετικά με μια πηγή περιεχομένου και να τις αναπαράγει σε άλλους διαφορετικούς τύπους περιεχομένου. Όταν το σύστημα περιορίζεται στο να προτείνει το περιεχόμενο του ίδιου τύπου που χρησιμοποιεί ήδη ο χρήστης, η τιμή από το σύστημα συστάσεων είναι σημαντικά μικρότερη όταν μπορούν να προταθούν άλλοι τύποι περιεχομένου από άλλες υπηρεσίες. Για παράδειγμα, η σύσταση άρθρων ειδήσεων με βάση την περιήγηση ειδήσεων είναι χρήσιμη, αλλά δεν θα ήταν πολύ πιο χρήσιμη όταν η μουσική, τα βίντεο από διαφορετικές υπηρεσίες μπορούν να προταθούν με βάση την περιήγηση ειδήσεων.



Εικόνα 5:Contend base filtering

Collaborative filtering: Αυτή η μέθοδος φιλτραρίσματος βασίζεται συνήθως στη συλλογή και ανάλυση πληροφοριών σχετικά με τις συμπεριφορές των χρηστών, τις δραστηριότητες ή τις προτιμήσεις τους και την πρόβλεψη του τι θα τους αρέσει με βάση την ομοιότητα με άλλους χρήστες. Ένα βασικό πλεονέκτημα της προσέγγισης συνεργατικού φιλτραρίσματος είναι ότι δεν βασίζεται σε αναλύσιμο περιεχόμενο από μηχανήματα και επομένως είναι σε θέση να προτείνει με ακρίβεια σύνθετα αντικείμενα, όπως ταινίες, χωρίς να απαιτείται «κατανόηση» του ίδιου του στοιχείου. Το συνεργατικό φιλτράρισμα βασίζεται στην εφαρμογή αυτής της προσέγγισης [18]που συνιστά στον ενεργό χρήστη τα στοιχεία που άλλοι χρήστες έχουν παρόμοια γούστα με αυτά τα οποία τους άρεσαν στο παρελθόν. Παρακάτω δίνεται παράδειγμα για τον τρόπο με τον οποίο γίνεται προσπάθεια για να μετρηθεί η ομοιότητα. Αντί να

χρησιμοποιούμε χαρακτηριστικά στοιχείων για να προσδιορίσουμε την ομοιότητά τους, εστιάζουμε στην ομοιότητα των αξιολογήσεων των χρηστών για δύο στοιχεία. Δηλαδή, στη θέση του διανύσματος στοιχείου-προφίλ για ένα στοιχείο, χρησιμοποιούμε τη στήλη του στον πίνακα χρησιμότητας.

Το πρώτο ερώτημα που πρέπει να αντιμετωπίσουμε είναι πώς να μετρήσουμε την ομοιότητα των χρηστών ή των στοιχείων από τις γραμμές ή τις στήλες τους στον πίνακα βοηθητικών προγραμμάτων. Ο βοηθητικός πίνακας είναι ο πίνακας όπου οι σειρές αντιπροσωπεύουν αξιολογήσεις χρηστών ή οποιοδήποτε τέτοιο χαρακτηριστικό και οι στήλες αντιπροσωπεύουν ταινία/στοιχεία.

Παίρνοντας υπόψιν τον ακόλουθο πίνακα χρησιμότητας:

	Cry to Heaven	The Sea	Titanic	Toy story	Home alone
A	3		5	1	
B	4	5	4		1
C			3	1	5
D	1	3			3

Εδώ τα A,B,C,D είναι οι χρήστες και οι κεφαλίδες στηλών είναι οι ταινίες. Οι τιμές στον πίνακα υποδηλώνουν βαθμολογίες. Η ομοιότητα συνημιτονίου είναι η εύρεση της απόστασης συνημιτονίου μεταξύ δύο χρηστών. Είναι παρόμοιο με το γινόμενο κουκίδων δύο διανυσμάτων. Τα κενά αντιμετωπίζονται ως 0 και ο τύπος αυτής της ομοιότητας είναι ο παρακάτω.

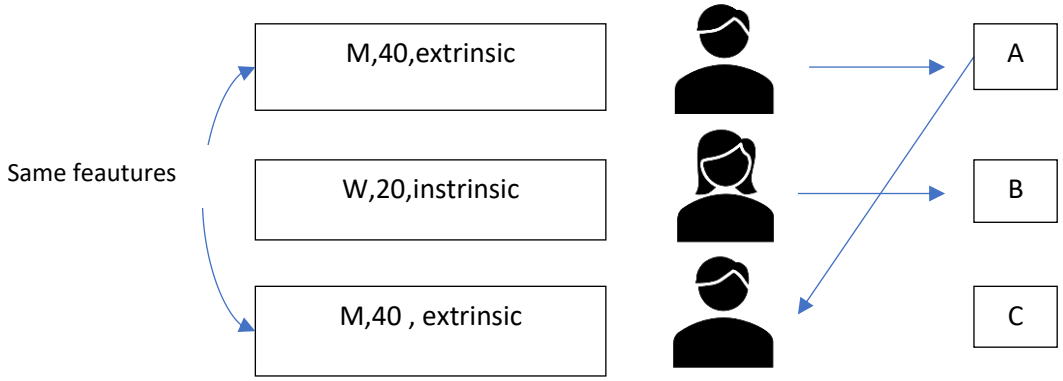
$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Εικόνα 6: Συνάρτηση ομοιότητας A , B

Οπότε αν υπολογίσουμε την ομοιότητα του A και του B έχουμε ότι η γωνία συν μεταξύ A και B είναι $\cos(\theta) = (3x4)+(5x4)/\text{sqrt}(3^2 + 5^2 + 1^2)*\text{sqrt}(4^2 + 5^2 + 4^2 + 1^2)$ και έτσι υπολογίζεται η ομοιότητα του συνημίτονου.

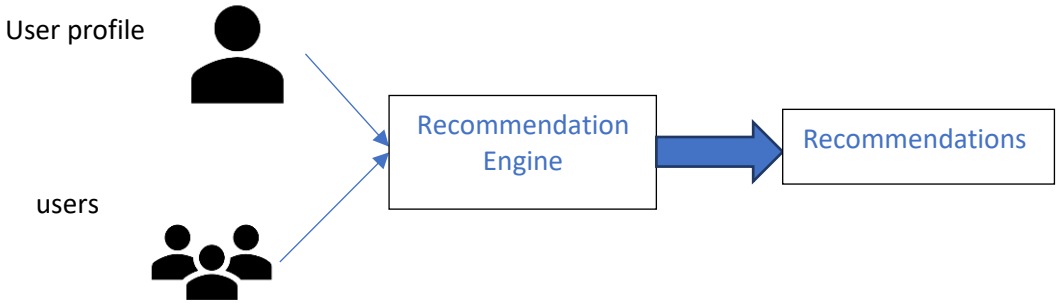
Demographic: Αυτός ο τύπος συστήματος προτείνει στοιχεία με βάση το δημογραφικό προφίλ του χρήστη. Η υπόθεση είναι ότι διαφορετικές συστάσεις πρέπει να δημιουργηθούν για διαφορετικές δημογραφικές θέσεις. Πολλές ιστοσελίδες υιοθετούν απλές και αποτελεσματικές λύσεις εξατομίκευσης με βάση τα δημογραφικά στοιχεία. Για παράδειγμα, οι χρήστες είναι ή αποστέλλονται σε συγκεκριμένους ιστότοπους με βάση τη γλώσσα ή τη χώρα τους. Οι προτάσεις μπορούν να προσαρμοστούν ανάλογα με την ηλικία του χρήστη. Ενώ αυτές οι προσεγγίσεις ήταν αρκετά δημοφιλείς στη βιβλιογραφία

του μάρκετινγκ, υπήρξε σχετικά μικρή κατάλληλη έρευνα RS σε δημογραφικά συστήματα [19]



Εικόνα 7: Demographic filtering

Knowledge-based: Τα συστήματα που βασίζονται στη γνώση προτείνουν στοιχεία που βασίζονται σε συγκεκριμένες γνώσεις τομέα σχετικά με τον τρόπο με τον οποίο ορισμένα χαρακτηριστικά στοιχεία ανταποκρίνονται στις ανάγκες και τις προτιμήσεις των χρηστών και, τελικά, πώς το στοιχείο είναι χρήσιμο για τον χρήστη. Τα αξιοσημείωτα συστήματα συστάσεων που βασίζονται στη γνώση βασίζονται σε περιπτώσεις [20], [21] Σε αυτά τα συστήματα μια συνάρτηση ομοιότητας υπολογίζει πόσο ταιριάζει ο χρήστης (περιγραφή προβλήματος) με τις συστάσεις (λύσεις του προβλήματος). Εδώ η βαθμολογία ομοιότητας μπορεί να ερμηνεύεται ως η χρησιμότητα της σύστασης για τον χρήστη. Τα συστήματα που βασίζονται στη γνώση τείνουν να λειτουργούν καλύτερα από άλλα στην αρχή της ανάπτυξής τους, αλλά εάν δεν είναι εξοπλισμένα με στοιχεία εκμάθησης μπορεί ξεπεραστούν από άλλες ρηχές μεθόδους που μπορούν να εκμεταλλευτούν τα αρχεία καταγραφής της αλληλεπίδρασης του ανθρώπου/υπολογιστή .



Εικόνα 8: knowledge based

Community-based: Αυτός ο τύπος συστήματος προτείνει στοιχεία με βάση

τις προτιμήσεις των φίλων των χρηστών. Αυτή η τεχνική ακολουθεί το επίγραμμα «Πες μου ποιος είναι ο φίλος σου, και θα σου πω ποιος είσαι». [22], [23]Τα στοιχεία δείχνουν ότι οι άνθρωποι τείνουν να βασίζονται περισσότερο σε συστάσεις από τους φίλους τους παρά σε συστάσεις από παρόμοια αλλά ανώνυμα άτομα [24]Αυτή η παρατήρηση, σε συνδυασμό με την αυξανόμενη δημοτικότητα των ανοιχτών κοινωνικών δικτύων, δημιουργεί αυξανόμενο ενδιαφέρον για συστήματα που βασίζονται στην κοινότητα ή όπως συνήθως αναφέρονται, κοινωνικά συστήματα συστάσεων [25]Αυτός ο τύπος RS μοντελοποιεί και αποκτά πληροφορίες για τις κοινωνικές σχέσεις των χρηστών και τις προτιμήσεις των φίλων του χρήστη. Η σύσταση βασίζεται σε αξιολογήσεις που δόθηκαν από τους φίλους του χρήστη. Στην πραγματικότητα αυτά τα RS ακολουθούν την άνοδο των κοινωνικών δικτύων και επιτρέπουν μια απλή και ολοκληρωμένη απόκτηση δεδομένων που σχετίζονται με τις κοινωνικές σχέσεις των χρηστών.

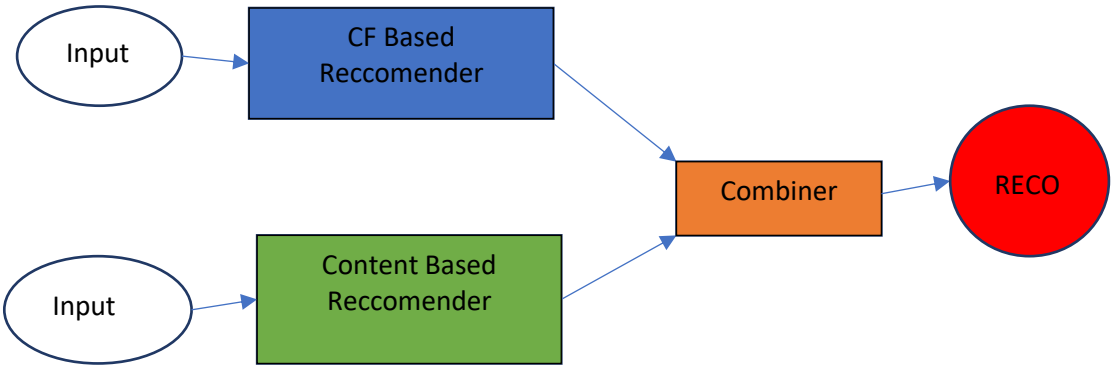
Η έρευνα σε αυτόν τον τομέα βρίσκεται ακόμη σε πρώιμο στάδιο και τα αποτελέσματα σχετικά με τα συστήματα και τις επιδόσεις τους είναι μικτές. Για παράδειγμα, οι [25], [26] αναφέρουν ότι οι γενικές συστάσεις που βασίζονται σε κοινωνικά δίκτυα δεν είναι πιο ακριβείς από αυτές που προέρχονται από τις παραδοσιακές προσεγγίσεις CF, εκτός από ειδικές περιπτώσεις, όπως όταν αξιολογήσεις χρηστών για ένα συγκεκριμένο στοιχείο έχουν μεγάλη ποικιλία (δηλαδή αμφιλεγόμενα στοιχεία) ή για καταστάσεις ψυχρής εκκίνησης, π.χ. Οι χρήστες δεν παρείχαν αρκετές βαθμολογίες για να υπολογίσουν την ομοιότητα με άλλους χρήστες. Οι υπόλοιποι έχουν δείξει ότι σε ορισμένες περιπτώσεις τα δεδομένα των κοινωνικών δικτύων παρέχουν καλύτερες συστάσεις από τα δεδομένα ομοιότητας προφίλ [27]και ότι η προσθήκη δεδομένων κοινωνικού δικτύου στα παραδοσιακά CF βελτιώνει τα αποτελέσματα των συστάσεων [28]

Hybrid recommender systems: Πρόσφατη έρευνα δείχνει ότι ο συνδυασμός συνεργατικών συστάσεων και συστάσεων που βασίζονται στο περιεχόμενο μπορεί να είναι πιο αποτελεσματικός. Οι υβριδικές προσεγγίσεις μπορούν να εφαρμοστούν κάνοντας προβλέψεις βάσει περιεχομένου και βάσει συνεργασίας χωριστά και στη συνέχεια συνδυάζοντάς τις. Περαιτέρω, προσθέτοντας δυνατότητες που βασίζονται στο περιεχόμενο σε μια προσέγγιση που βασίζεται σε συνεργασία και αντίστροφα ή με την ενοποίηση των προσεγγίσεων σε ένα μοντέλο.

Αρκετές μελέτες επικεντρώθηκαν στη σύγκριση της απόδοσης του υβριδίου με τις αμιγώς συνεργατικές και βασισμένες στο περιεχόμενο μεθόδους και καταδεικνύουν ότι οι υβριδικές μέθοδοι μπορούν να παρέχουν πιο ακριβείς συστάσεις από τις καθαρές προσεγγίσεις. Τέτοιες μέθοδοι μπορούν να χρησιμοποιηθούν για να ξεπεραστούν τα κοινά προβλήματα στα συστήματα συστάσεων, όπως η ψυχρή εκκίνηση και το πρόβλημα της έλλειψης δεδομένων.

Το Netflix είναι ένα καλό παράδειγμα χρήσης υβριδικών συστημάτων συστάσεων. Ο ιστότοπος κάνει συστάσεις συγκρίνοντας τις συνήθειες

παρακολούθησης και αναζήτησης παρόμοιων χρηστών (δηλαδή, συνεργατικό φιλτράρισμα), καθώς και προσφέροντας ταινίες που έχουν κοινά χαρακτηριστικά με ταινίες που ένας χρήστης έχει αξιολογήσει υψηλά (φιλτράρισμα βάσει περιεχομένου).



Εικόνα 9: Hybrid recommended

Τύποι δεδομένων για δημιουργία συστήματος προτάσεων: Υπάρχουν δύο είδη δεδομένων διαθέσιμα για την κατασκευή ενός συστήματος συστάσεων.

- **Types of Data for building recommendation systems:** Η ρητή ανατροφοδότηση είναι τα δεδομένα σχετικά με τη ρητή ανατροφοδότηση χρήστη (αξιολογήσεις κ.λπ.) σχετικά με ένα προϊόν. Ενημερώνει άμεσα ότι στους χρήστες αρέσει ένα προϊόν ή όχι.
- **Implicit feedback:** Στα σιωπηρά σχόλια, δεν έχουμε τα δεδομένα σχετικά με τον τρόπο με τον οποίο ο χρήστης αξιολογεί ένα προϊόν. Παραδείγματα σιωπηρών σχολίων είναι τα κλικ, οι ταινίες που παρακολουθήσατε, τα τραγούδια που παίχτηκαν, οι αγορές ή οι αντιστοιχισμένες ετικέτες.

Για την επιλογή τελικά ποιου συστήματος είναι καταλληλότερο για την επίλυση αυτής της εφαρμογής χρειάστηκε να γίνει η προηγούμενη μελέτη και τελικά επιλέχθηκε ο συνδυασμός συνεργατικών συστάσεων και συστάσεων που βασίζονται στο περιεχόμενο δηλαδή το υβριδικό σύστημα συστάσεων.

Το υβριδικό σύστημα συστάσεων είναι ένα ειδικό είδος συστάσεων που χρησιμοποιεί φιλτράρισμα τόσο συνεργασίας όσο και φιλτραρίσματος βάσει περιεχομένου για την υποβολή προτάσεων. Αυτό καθιστά το υβριδικό σύστημα συστάσεων μια πολύ ειδική και χρήσιμη μέθοδο για την κατασκευή συστήματος συστάσεων. Υπάρχουν όμως αρκετές τεχνικές και μέθοδοι για τη δημιουργία υβριδικού συστήματος συστάσεων.

3. Ορισμός προβλήματος και εργαλεία επίλυσης

3.1 Ορισμός προβλήματος

Σε αυτή την εργασία θα γίνει η προσπάθεια για την εκτέλεση εκμάθησης μεταφοράς γνώσης όπου το προ εκπαιδευμένο μοντέλο συνεργατικού φιλτραρίσματος εκπαιδεύεται πρώτα στο ml-latest-small σύνολο δεδομένων. Οι ήδη γνωστές αναπαραστάσεις και τα βάρη του μοντέλου χρησιμοποιούνται στη συνέχεια ως το σημείο εκκίνησης για εκπαίδευση στο σύνολο δεδομένων goodbooks-10k.

Στην εργασία αυτή το κοινό σημείο τομής μεταξύ των dataset είναι το ratings δηλαδή το μοντέλο που θα εκπαιδευτεί στα ml-latest-small θα είναι πάνω στα ratings και μετά αυτό το προ εκπαιδευμένο μοντέλο θα εφαρμοστεί πάνω στα goodbooks-10k καθώς αυτό είναι το κοινό τους πεδίο.

Αυτή η διαδικασία επαναχρησιμοποίησης του προεκπαιδευμένου μοντέλου και προσαρμογής του σε ένα σχετικό σύνολο δεδομένων είναι μια μορφή μεταφοράς μάθησης. Έπειτα θα γίνει η ίδια εργασία για τα goodbooks-10k χωρίς την μεταφορά γνώσης και θα συγκρίνουμε τα διαφορετικά rmse αλλά και τα βιβλία τα οποία προτείνουν τα δύο κομμάτια της εργασίας.

Θα γίνει χρήση αρκετών εργαλείων τα οποία θα αναλυθούν παρακάτω ώστε να μπορέσει να δουλέψει ορθά αυτή η εφαρμογή όπως το μοντέλο παραγοντοποιήσης υβριδικού πίνακα , το εργαλείο anaconda για την υποστήριξη του jupyter και τέλος της γλώσσας προγραμματισμού python όπως και βιβλιοθήκες της για το τελικό αποτέλεσμα.

3.2 Εργαλεία anaconda-Jupyter

Σε αυτό το σημείο θα αναλυθούν τα εργαλεία που χρησιμοποιήθηκαν για να επεξεργαστούν τα δεδομένα και να δώσουν στον χρήστη τα θεμιτά αποτελέσματα που θέλουμε.

Αρχικά χρειάστηκε ένα εργαλείο το οποίο να συγκεντρώσει όλες τις βιβλιοθήκες όπως την numpy, , time και άλλες εφαρμογές όπως το jupyter με τις οποίες θα υλοποιηθεί ή εργασία και θα μας εκτυπώσει τα απαραίτητά διαγράμματα και αποτελέσματα τα οποία θέλουμε να έχουμε .

Παρακάτω αναλύονται τα εργαλεία και οι βιβλιοθήκες που χρησιμοποιούνται για τη μεταφορά μάθησης στο πλαίσιο του συνεργατικού φιλτραρίσματος:

- Pandas: Το Pandas είναι μια ισχυρή βιβλιοθήκη χειρισμού και ανάλυσης δεδομένων στην Python. Παρέχει δομές δεδομένων όπως

DataFrames για αποτελεσματική εργασία με δομημένα δεδομένα. Σε αυτόν τον κώδικα, το Pandas χρησιμοποιείται για τη φόρτωση και το χειρισμό των συνόλων δεδομένων. Συγκεκριμένα, χρησιμοποιείται για την ανάγνωση των συνόλων δεδομένων ml-latest-small και goodbooks-10k από αρχεία CSV χρησιμοποιώντας τη συνάρτηση read_csv(). Τα Pandas DataFrames επιτρέπουν το εύκολο φιλτράρισμα και την επιλογή συγκεκριμένων στηλών ενδιαφέροντος, όπως η εξαγωγή του userId, movieId, βαθμολογία από ml-latest-small και user_id, book_id, βαθμολογία από goodbooks-10k σύνολα δεδομένων.

- Surprise: Το Surprise είναι μια βιβλιοθήκη Python που έχει σχεδιαστεί για τη δημιουργία και την ανάλυση συστημάτων συστάσεων. Παρέχει διάφορους αλγόριθμους συνεργατικού φιλτραρίσματος, μετρήσεις αξιολόγησης και βοηθητικά προγράμματα για το χειρισμό εργασιών που σχετίζονται με συστάσεις. Σε αυτόν τον κώδικα, το Surprise χρησιμοποιείται εκτενώς για διαφορετικές εργασίες συστήματος συστάσεων.

- Dataset Loading: Η **Dataset** από το Surprise χρησιμοποιείται για τη φόρτωση των συνόλων δεδομένων. Η συνάρτηση load_from_df() χρησιμοποιείται για τη δημιουργία του αντικειμένου Surprise Dataset από τα Pandas DataFrames. Παίρνει τα δεδομένα αξιολόγησης στοιχείων χρήστη από τα DataFrames και χρησιμοποιεί το αντικείμενο Reader για να ορίσει την κλίμακα αξιολόγησης.

- **Model Training**: Η βιβλιοθήκη Surprise παρέχει συνεργατικούς αλγόριθμους φιλτραρίσματος, όπως SVD (Singular Value Decomposition). Η κλάση SVD() χρησιμοποιείται για τη δημιουργία μιας παρουσίας του αλγορίθμου SVD. Αυτός ο αλγόριθμος χρησιμοποιείται ευρέως σε συστήματα συστάσεων για την παραγοντοποίηση της μήτρας αξιολόγησης στοιχείου χρήστη και τη σύλληψη λανθάνοντων παραγόντων. Στη συνέχεια, η μέθοδος fit() καλείται στο μοντέλο συνεργατικού φιλτραρίσματος (ml_algo) για να την εκπαιδεύσει στα σύνολα δεδομένων ml-latest-small και goodbooks-10k.

- **Model Evaluation**: Η βιβλιοθήκη Surprise προσφέρει επίσης μετρήσεις αξιολόγησης για την αξιολόγηση της απόδοσης των μοντέλων προτάσεων. Η μονάδα ακρίβειας παρέχει λειτουργίες όπως η rmse() για τον υπολογισμό του RMSE μεταξύ της προβλεπόμενης και της πραγματικής βαθμολογίας. Η συνάρτηση rmse() καλείται με τις προβλέψεις του μοντέλου (ml_predictions) για τον υπολογισμό του RMSE για το μοντέλο ml-latest-small.

- **Data Splitting**: Η συνάρτηση train_test_split() από τη βιβλιοθήκη

Surprise χρησιμοποιείται για να χωρίσει το σύνολο δεδομένων ml-latest-small σε σετ εκπαίδευσης και δοκιμής. Χωρίζει τυχαία τα δεδομένα σε δύο μέρη με βάση το παρεχόμενο μέγεθος δοκιμής. Αυτός ο διαχωρισμός είναι απαραίτητος για την αξιολόγηση της απόδοσης του μοντέλου σε μη ορατά δεδομένα.

- SVD (Singular Value Decomposition): Το SVD είναι μια τεχνική παραγοντοποίησης μητρών που χρησιμοποιείται συνήθως σε συστήματα συστάσεων που βασίζονται σε συνεργατικό φιλτράρισμα. Αποσυνθέτει τον πίνακα αξιολόγησης στοιχείων χρήστη σε πίνακες χαμηλότερης κατάταξης για να συλλάβει λανθάνοντες παράγοντες και να κάνει προβλέψεις. Ο αλγόριθμος SVD είναι ένας από τους αλγόριθμους συνεργατικού φιλτραρίσματος που εφαρμόζονται στο Surprise. Στον κώδικα, η κλάση SVD() χρησιμοποιείται για τη δημιουργία μιας παρουσίας του αλγορίθμου SVD (ml_algo), η οποία στη συνέχεια εκπαιδεύεται στα σύνολα δεδομένων ml-latest-small και goodbooks-10k.

- Reader: Η κλάση Reader στο Surprise βοηθά στην ανάλυση των συνόλων δεδομένων και στον καθορισμό της κλίμακας αξιολόγησης. Στον κώδικα, το αντικείμενο Reader() δημιουργείται με την επιθυμητή κλίμακα βαθμολογίας, η οποία ορίζεται ως (0,5, 5,0) χρησιμοποιώντας την παράμετρο rating_scale. Αυτό το εύρος χρησιμοποιείται για την ανάλυση και την κανονικοποίηση των αξιολογήσεων στα σύνολα δεδομένων σωστά.

Ο συνδυασμός των βιβλιοθηκών Pandas και Surprise επιτρέπει την αποτελεσματική επεξεργασία δεδομένων, την εκπαίδευση μοντέλων, την αξιολόγηση και τη δημιουργία προτάσεων σε αυτόν τον κώδικα. Το Pandas χειρίζεται τη φόρτωση, τον χειρισμό και το φιλτράρισμα δεδομένων, ενώ το Surprise παρέχει τους συνεργατικούς αλγόριθμους φιλτραρίσματος, τις μετρήσεις αξιολόγησης και τις βοηθητικές λειτουργίες που είναι ειδικά προσαρμοσμένες για συστήματα συστάσεων.

Αυτές οι βιβλιοθήκες παρέχουν λειτουργίες για τη φόρτωση, την προεπεξεργασία, την εκπαίδευση, την αξιολόγηση και την πραγματοποίηση προβλέψεων χρησιμοποιώντας συνεργατικά μοντέλα φιλτραρίσματος. Απλοποιούν την υλοποίηση της μάθησης μεταφοράς στο συνεργατικό φιλτράρισμα παρέχοντας αφαιρέσεις υψηλού επιπέδου και έτοιμους προς χρήση αλγόριθμους.

Χρησιμοποιώντας αυτά τα εργαλεία και τις βιβλιοθήκες, ο κώδικας δείχνει τη διαδικασία μεταφοράς μάθησης αρχικοποιώντας ένα μοντέλο με προ εκπαιδευμένα βάρη σε ένα σύνολο δεδομένων ("ml-latest-small") και στη συνέχεια ρυθμίζοντας το σε ένα άλλο σχετικό σύνολο δεδομένων ("goodbooks-

10k") για τη βελτίωση της απόδοσης των συστάσεων.

3.3 Χαρακτηριστικά του περιβάλλοντος Jupyter

Το Project Jupyter είναι ένα σύνολο έργων λογισμικού ανοιχτού κώδικα για δια δραστικούς και διερευνητικούς υπολογιστές που προκύπτουν από το IPython. Το κεντρικό στοιχείο που προσφέρει η Jupyter είναι το Σημειωματάριο Jupyter– μια διαδικτυακή δια δραστική πλατφόρμα υπολογιστών. [29] Το επιτρέπει στους χρήστες να δημιουργούν αφηγήσεις που βασίζονται σε δεδομένα και κώδικα που συνδυάζουν ζωντανό (επαναεκτελέσιμο) κώδικα, εξισώσεις, αφηγηματικό κείμενο, δια δραστικούς πίνακες εργαλείων και άλλα πλούσια μέσα. Το σημειωματάριου Jupyter παρέχει μια πλήρη και εκτελέσιμη εγγραφή του που μπορεί να μοιραστεί με άλλους με τρόπο που δεν ήταν δυνατό στο παρελθόν. [30]

Ένα σημειωματάριο Jupyter [2] είναι ταυτόχρονα και δια δραστικό εγγράμματο έγγραφο προγραμματισμού και μια εφαρμογή που εκτελεί έγγραφα. Ένα σημειωματάριο αποτελείται από κελιά, τα οποία μπορεί να είναι τριών κατηγοριών : κωδικός, μαρκάρισμα και ακατέργαστο. Ένα κελί κώδικα περιέχει εκτελέσιμο κώδικα που χρησιμοποιείται για την παραγωγή αποτελεσμάτων. Ένα κελί σήμανσης περιέχει μορφοποιημένο κείμενο. Τέλος, ένα ακατέργαστο κελί περιέχει κείμενο που είναι ούτε κώδικας ούτε μορφοποιημένο κείμενο. Εργαλεία που μετατρέπουν σημειωματάρια σε άλλες μορφές χρησιμοποιούν ακατέργαστα κελιά για διαμόρφωση. Ο Jupyter χρησιμοποιεί έναν πυρήνα για την εκτέλεση κελιών κώδικα. Όταν ο Jupyter στέλνει ένα κελί κώδικα για εκτέλεση, επισημαίνει το κελί ως εκτελούμενο εκχωρώντας το στον μετρητή εκτέλεσης κελιών. Μετά την εκτέλεση, ο πυρήνας εκχωρεί έναν αριθμό στον μετρητή, ο οποίος υποδεικνύει την εντολή εκτέλεσης. Οι χρήστες μπορούν να εκτελέσουν στα κελιά οποιαδήποτε παραγγελία και ένα δεδομένο κελί μπορεί να εκτελεστεί πολλές φορές.

Είναι δυνατή η αποθήκευση εκτελεσμένων ή μη σημειωματάριων. Ένα μη εκτελεσμένο σημειωματάριο περιέχει μόνο προοπτικές δεδομένα [31]δηλαδή ο τίτλος του σημειωματάριου και ο ορισμός των κελιών του. Το εκτελεσμένο σημειωματάριο περιέχει πιθανά δεδομένα συν αναδρομικά δεδομένα [31] που προέρχονται από την εκτέλεση των κελιών του σημειωματάριου και τους μετρητές εκτέλεσής τους. Η εκτέλεση ενός notebook δεν απαιτεί καθαρισμό των εξόδων των προηγούμενων εκτελέσεων. Έτσι, ένα εκτελεσμένο σημειωματάριο μπορεί να περιέχει αναδρομικά δεδομένα πολλαπλών εκτελέσεων.

Ακόμη μέσα στο περιβάλλον όπου γίνεται η επίλυση αυτής της εφαρμογής ο χρήστης χρησιμοποιεί την γλώσσα προγραμματισμού python όπως προαναφέρθηκε καθώς αυτή διαθέτει πολλές βιβλιοθήκες και εργαλεία τα οποία είναι χρήσιμα τόσο για την επίλυση της εφαρμογής αλλά και για την καλύτερη παρουσίαση των αποτελεσμάτων.

3.4 Datasets

Σε αυτό το σημείο θα αναφερθούμε στα δεδομένα τα οποία χρησιμοποιήθηκαν για την υλοποίηση της εφαρμογής .

- Αρχικά είναι το **ml-latest-small dataset** : Το σύνολο δεδομένων ml-latest-small είναι ένα σύνολο δεδομένων αξιολόγησης ταινιών που περιέχει αξιολογήσεις χρηστών για ταινίες. Είναι μια μικρότερη έκδοση του ml-latest συνόλου δεδομένων .
  Το σύνολο δεδομένων περιέχει τις ακόλουθες κατηγορίες :

  - ratings.csv: Αυτό το αρχείο περιέχει τις αξιολογήσεις ταινιών που παρέχονται από τους χρήστες. Συνήθως περιλαμβάνει τις ακόλουθες στήλες:
    - userId: Ένα αναγνωριστικό για κάθε χρήστη.
    - movieId: Ένα αναγνωριστικό για κάθε ταινία.
    - rating: Η βαθμολογία που δίνει ο χρήστης για την ταινία (συνήθως σε κλίμακα από 0,5 έως 5,0).
    - timestamp: Η χρονική σήμανση που υποδεικνύει πότε δόθηκε η βαθμολογία..

- Το σύνολο δεδομένων goodbooks-10k είναι ένα σύνολο δεδομένων αξιολόγησης βιβλίων που περιέχει αξιολογήσεις χρηστών για βιβλία. Είναι μια συλλογή αξιολογήσεων και μεταδεδομένων από τον ιστότοπο Goodreads.
  Το σύνολο δεδομένων περιέχει τις ακόλουθες κατηγορίες:

  - ratings.csv: Αυτό το αρχείο περιέχει τις αξιολογήσεις βιβλίων που παρέχονται από τους χρήστες. Συνήθως περιλαμβάνει τις ακόλουθες στήλες:

    - user_id: Ένα αναγνωριστικό για κάθε χρήστη.
    - book_id: Ένα αναγνωριστικό για κάθε βιβλίο.
    - rating: Η βαθμολογία που δίνει ο χρήστης για το βιβλίο (συνήθως σε κλίμακα από 0,5 έως 5,0).

  - books.csv: Αυτό το αρχείο περιέχει πληροφορίες μεταδεδομένων σχετικά με τα βιβλία στο σύνολο δεδομένων. Συνήθως περιλαμβάνει τις ακόλουθες στήλες:

    - book_id: Ένα αναγνωριστικό για κάθε βιβλίο.
    - title: Ο τίτλος του βιβλίου.
    - authors: Ο/οι συγγραφέας του βιβλίου.

Ενδέχεται να υπάρχουν και άλλα πεδία μεταδεδομένων, όπως έτος δημοσίευσης, μέση βαθμολογία και ούτω καθεξής.
Μία αναπαράσταση είναι η παρακάτω των δεδομένων του goodbooks-10k.

	book_id	goodreads_book_id	best_book_id	work_id	books_count	isbn	isbn13	authors	original_publication_year	original_title	...	rati
0	1	2767052	2767052	2792775	272	439023483	9.780439e+12	Suzanne Collins	2008.0	The Hunger Games	...	
1	2	3	3	4640799	491	439554934	9.780440e+12	J.K. Rowling, Mary GrandPré	1997.0	Harry Potter and the Philosopher's Stone	...	
2	3	41865	41865	3212258	226	316015849	9.780316e+12	Stephenie Meyer	2005.0	Twilight	...	
3	4	2657	2657	3275794	487	61120081	9.780061e+12	Harper Lee	1960.0	To Kill a Mockingbird	...	
4	5	4671	4671	245494	1356	743273567	9.780743e+12	F. Scott Fitzgerald	1925.0	The Great Gatsby	...	
...	...	...	...	...	...	...	...	...	...	...	...	
9995	9996	7130616	7130616	7392860	19	441019455	9.780441e+12	Ilona Andrews	2010.0	Bayou Moon	...	
9996	9997	208324	208324	1084709	19	067973371X	9.780680e+12	Robert A. Caro	1990.0	Means of Ascent	...	
9997	9998	77431	77431	2393986	60	039330762X	9.780393e+12	Patrick O'Brian	1977.0	The Mauritius Command	...	
9998	9999	8565083	8565083	13433613	7	61711527	9.780062e+12	Peggy Orenstein	2011.0	Cinderella Ate My Daughter: Dispatches from th...	...	
9999	10000	8914	8914	11817	31	375700455	9.780376e+12	John Keegan	1998.0	The First World War	...	

10000 rows × 23 columns

Εικόνα 10: Books data presentation

ratings_count	work_ratings_count	work_text_reviews_count	ratings_1	ratings_2	ratings_3	ratings_4	ratings_5	image_url	
4780653	4942365	155254	66715	127936	560092	1481305	2706317	https://images.gr-assets.com/books/1447303603m...	assets.cc
4602479	4800065	75867	75504	101676	455024	1156318	3011543	https://images.gr-assets.com/books/1474154022m...	assets.cc
3866839	3916824	95009	456191	436802	793319	875073	1355439	https://images.gr-assets.com/books/1361039443m...	assets.cc
3198671	3340896	72586	60427	117415	446835	1001952	1714267	https://images.gr-assets.com/books/1361975680m...	assets.cc
2683664	2773745	51992	86236	197621	606158	936012	947718	https://images.gr-assets.com/books/1490528560m...	assets.cc
...	...	...	...	...	...	...	...	...	
17204	18856	1180	105	575	3538	7860	6778	https://images.gr-assets.com/books/1307445460m...	assets.cc
12582	12952	395	303	551	1737	3389	6972	https://s.gr-assets.com/assets/nophoto/book/11...	assets.com/a
9421	10733	374	11	111	1191	4240	5180	https://images.gr-assets.com/books/1455373531m...	assets.cc
11279	11994	1988	275	1002	3765	4577	2375	https://images.gr-assets.com/books/1279214118m...	assets.cc
9162	9700	364	117	345	2031	4138	3069	https://images.gr-assets.com/books/1403194704m...	assets.cc

Εικόνα 11: Υπόλοιπα πεδία του goodreads

Στον κώδικα, το σύνολο δεδομένων ml-latest-small χρησιμοποιείται αρχικά για την εκπαίδευση ενός μοντέλου συλλογικού φιλτραρίσματος. Στη συνέχεια, το μοντέλο ρυθμίζεται με ακρίβεια στο σύνολο δεδομένων goodbooks-10k για να βελτιωθεί η απόδοσή του στην πρόταση βιβλίων. Τα σύνολα δεδομένων φορτώνονται χρησιμοποιώντας Panda και συγκεκριμένες στήλες (userId, movieId, rating από το ml-latest-small and user_id, book_id, rating από το

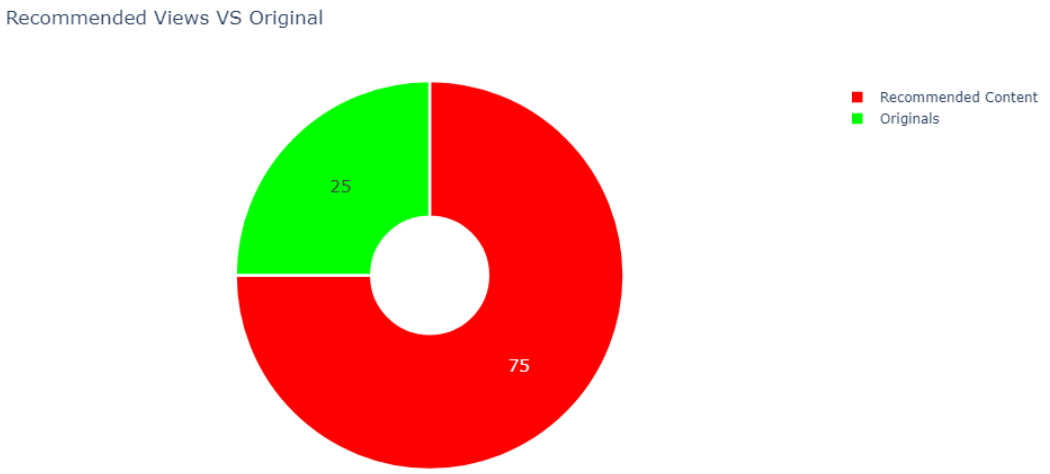
goodbooks-10k) χρησιμοποιούνται για τις αντίστοιχες εργασίες εκπαίδευσης και fine-tuning του μοντέλου.

4. Ανάπτυξη προβλήματος και αποτελέσματα.

Ο σκοπός της παρούσας εργασίας ήταν να προσπαθήσει να βρει ένα τρόπο να μεταφέρει γνώση από ένα πεδίο σε ένα άλλο με την χρήση διαφόρων βιβλιοθηκών και τεχνικών ώστε να έχει ως αποτέλεσμα για τον χρήστη για τον οποίο γνωρίζουμε ότι έχει αγαπημένες ταινίες και τις έχει αξιολογήσεις με ένα βαθμό ανάμεσα στο 1 και στο 5 , να του προταθούν νέες ταινίες τις οποίες δεν έχει ακόμη αξιολογήσει.

Για την ανάλυση του προβλήματος θα χρειαστεί να γίνει αναφορά τόσο στα δεδομένα που χρησιμοποιήθηκαν αλλά και σε τεχνικές που είναι χρήσιμες για την σύγκριση αλλά και για την παρουσίαση των αποτελεσμάτων.

Είναι γεγονός ότι το 75% από αυτά που τελικά παρακολουθούν οι χρήστες στο Netflix προέρχεται από συστάσεις, και τα συστήματα συστάσεων είναι υπεύθυνα για το 70% του χρόνου που αφιερώνουν οι άνθρωποι παρακολουθώντας βίντεο στο YouTube.



Εικόνα 12: Recommended view vs Original

Όπως προαναφέρθηκε πρέπει αρχικά να γίνει η ανάλυση των δεδομένων που θα χρησιμοποιηθούν . Σε αυτή την έρευνα χρησιμοποιήθηκαν δύο σύνολα δεδομένων , τα δεδομένα από το Movielens και πιο συγκεκριμένα το ml-latest-small σύνολο δεδομένων και τα δεδομένα από το goodbooks-10k.

Το σύνολο δεδομένων ml-latest-small περιέχει αξιολογήσεις ταινιών, ενώ το σύνολο δεδομένων goodbooks-10k περιέχει αξιολογήσεις βιβλίων. Παρά τις διαφορές μεταξύ ταινιών και βιβλίων, τα δύο σύνολα δεδομένων σχετίζονται καθώς και τα δύο περιλαμβάνουν στοιχεία αξιολόγησης χρηστών.

Αυτό είναι και το πιο σημαντικό ώστε να γίνει η χρήση της μεταφοράς γνώσης από ένα πεδίο σε ένα άλλο καθώς έχουν ένα κοινό σημείο έτσι προσαρμόζοντας παρακάτω το μοντέλο στο σύνολο δεδομένων goodbooks-10k, το μοντέλο μπορεί να μάθει από τα πρόσθετα δεδομένα και ενδεχομένως να βελτιώσει την απόδοσή του στην εργασία της σύστασης βιβλίων.

4.1 Ανάλυση δεδομένων

Ξεκινώντας για την επίλυση αυτού του προβλήματος χρειάζεται αρχικά η ανάλυση των δύο συνόλων δεδομένων ως προς τα δεδομένα που θα χρησιμοποιήσουμε αλλά και αν αυτά έχουν ομοιότητες όπως αναφέρθηκε προηγουμένως .

Για την εκτέλεση εκμάθησης μεταφοράς, το προεκπαιδευμένο μοντέλο συνεργατικού φιλτραρίσματος εκπαιδεύεται πρώτα στο ml-latest-small σύνολο δεδομένων. Οι ήδη γνωστές αναπαραστάσεις και τα βάρη του μοντέλου χρησιμοποιούνται στη συνέχεια ως το σημείο εκκίνησης για εκπαίδευση στο σύνολο δεδομένων goodbooks-10k. Αυτή η διαδικασία επαναχρησιμοποίησης του προεκπαιδευμένου μοντέλου και προσαρμογής του σε ένα σχετικό σύνολο δεδομένων είναι μια μορφή μεταφοράς μάθησης.

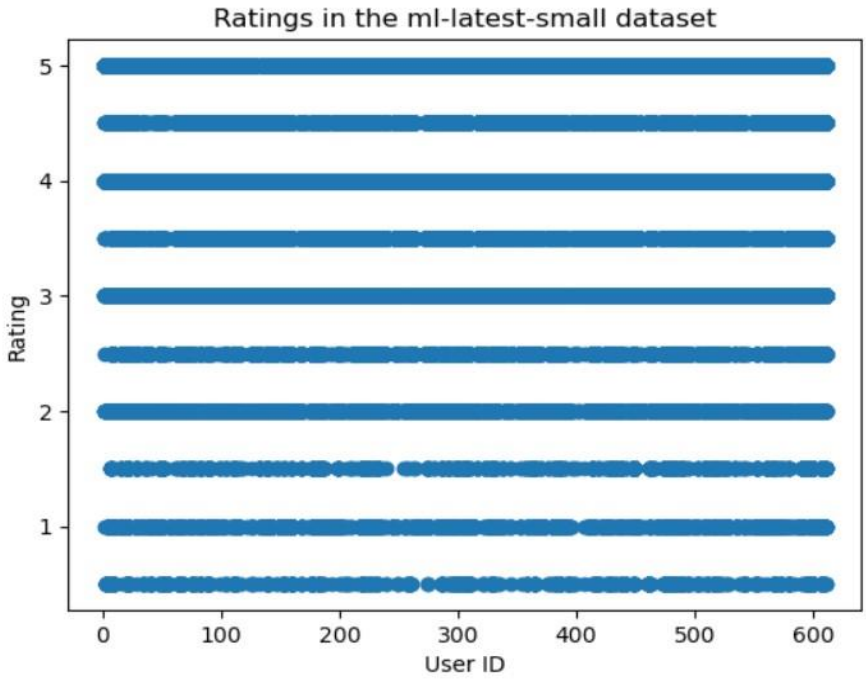
Για αυτούς τους λόγους σε αυτό το κομμάτι χρειάζεται να κρατηθούν οι στήλες με τα id και τα ratings καθώς είναι τα κοινά σημεία αναφοράς και για τα δύο σύνολα δεδομένων και να απαλειφθούν οι υπόλοιπες.

	book_id	average_rating	original_publication_year	ratings_count	language_code
0	1	4.34	2008.0	4780653	eng
1	2	4.44	1997.0	4602479	eng
2	3	3.57	2005.0	3866839	en-US
3	4	4.25	1960.0	3198671	eng
4	5	3.89	1925.0	2683664	eng
...	...	...	...	...	...
9995	9996	4.09	2010.0	17204	eng
9996	9997	4.25	1990.0	12582	eng
9997	9998	4.35	1977.0	9421	eng
9998	9999	3.65	2011.0	11279	eng
9999	10000	4.00	1998.0	9162	NaN

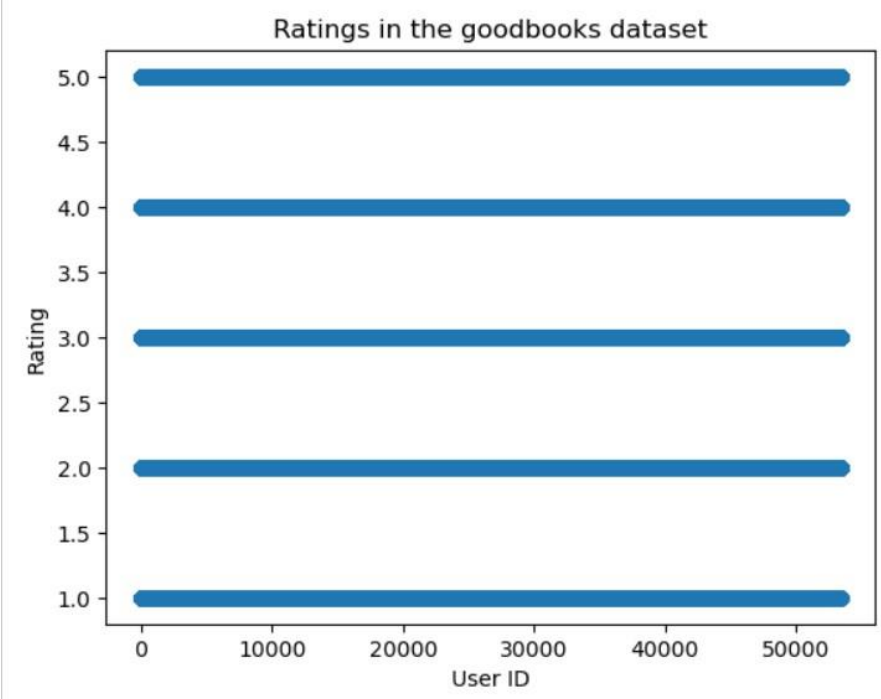
10000 rows × 5 columns

Εικόνα 13: Όλες οι στήλες από το goodbooks dataset

Παρακάτω αναπαρίστανται τα δεδομένα για τα ratings τόσο του ml-latest-small συνόλου δεδομένων αλλά και τα δεδομένα από το goodbooks-10k.



Εικόνα 14: ml-dataset



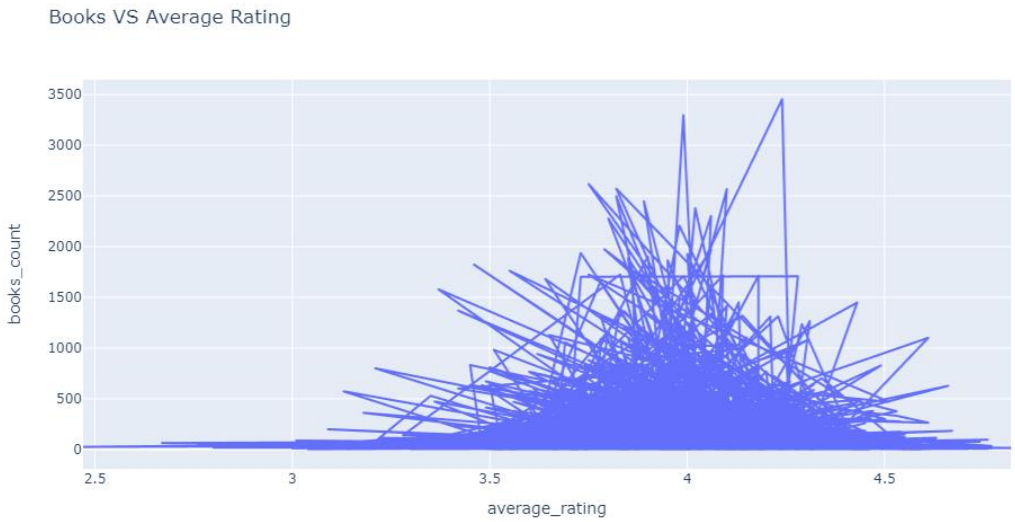
Εικόνα 15: goodbooks dataset

Σε αυτά τα διαγράμματα παρατηρείτε ότι έχουν κοινα σημεία αξιολόγησης το καθένα για το είδος του και έτσι θα το εκμεταλευτούμε αυτό ώστε να δούμε τι αποτελέσματα θα πάρουμε.

Γενικά σε αυτή την εργασία, η εκμάθηση μεταφοράς γνώσης χρησιμοποιείται για τη βελτίωση της απόδοσης ενός μοντέλου συλλογικού φιλτραρίσματος (Collaborative Filtering) που έχει εκπαιδευτεί στο σύνολο δεδομένων ml-latest-small προσαρμόζοντάς το στο σύνολο δεδομένων goodbooks-10k. Το συνεργατικό φιλτράρισμα (Collaborative Filtering) είναι μια δημοφιλής τεχνική που χρησιμοποιείται σε συστήματα συστάσεων που περιλαμβάνει τη δημιουργία προτάσεων με βάση την ομοιότητα της συμπεριφοράς ή των προτιμήσεων των χρηστών. Το σύνολο δεδομένων ml-latest-small αποτελείται από αξιολογήσεις ταινιών, ενώ το σύνολο δεδομένων goodbooks-10k αποτελείται από αξιολογήσεις βιβλίων. Αν και τα σύνολα δεδομένων είναι διαφορετικά, μοιράζονται την ίδια υποκείμενη εργασία σύστασης.

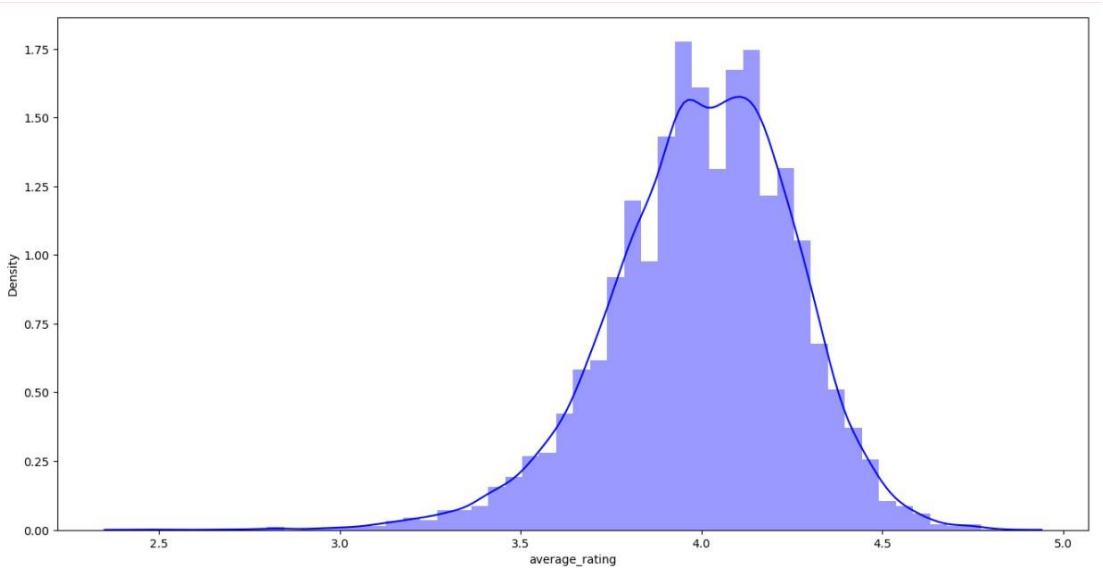
Οι μαθημένες αναπαραστάσεις και τα βάρη του προεκπαιδευμένου μοντέλου επαναχρησιμοποιούνται και πραγματοποιείται περαιτέρω εκπαίδευση για την προσαρμογή του μοντέλου στο νέο σύνολο δεδομένων.

Στην συνέχεια βλέπουμε τα στοιχεία από τα goodbooks όπου έχουμε αρχικά και τα οποία όπως προαναφέρθηκε θα επεξεργαστούμε για να έχουμε ως τελικό αποτέλεσμα για τον χρήστη να του προτείνουμε από τα βιβλία τα οποία δεν έχει αξιολογήσει.



Εικόνα 16: Βιβλία με βάση την μέση αξιολόγηση

Καθώς μπορούμε να δούμε και με μία άλλη αναπαράσταση τα δεδομένα για το rating ως προς τα που κυμαίνονται.



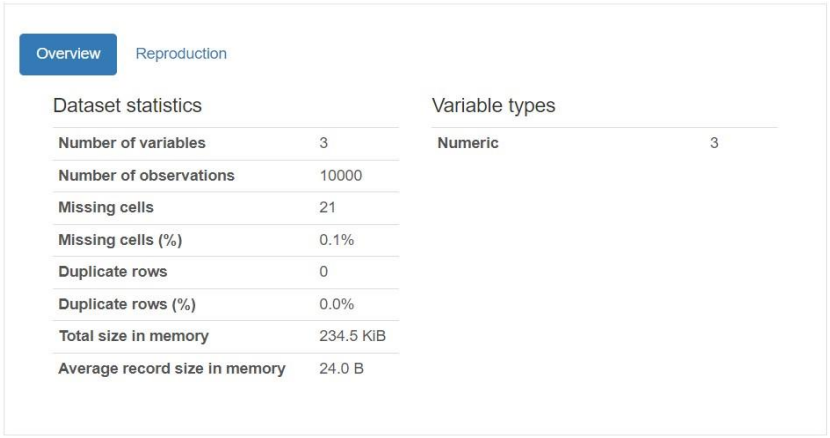
Εικόνα 17: Μέση τιμή αξιολόγησης βιβλίων.

Για την ανάδειξη όλων αυτών των σχημάτων και αυτών που θα δούμε παρακάτω έχουν χρησιμοποιηθεί και οι ανάλογες βιβλιοθήκες οι οποίες προαναφέρθηκαν στο κομμάτι της θεωρίας αλλά είναι σημαντικό να αναφερθούν και εδώ επιγραμματικά όπως οι pandas, matplotlib αλλά και βιβλιοθήκες με σημαντικές συναρτήσεις για την επίλυση του προβλήματος όπως οι SVD, Dataset, Reader από την βιβλιοθήκη surprise.

Ακόμη ένα ισχυρό εργαλείο που χρησιμοποιήθηκε ώστε να γίνει αποφυγή λαθών είναι το εργαλείο panda.

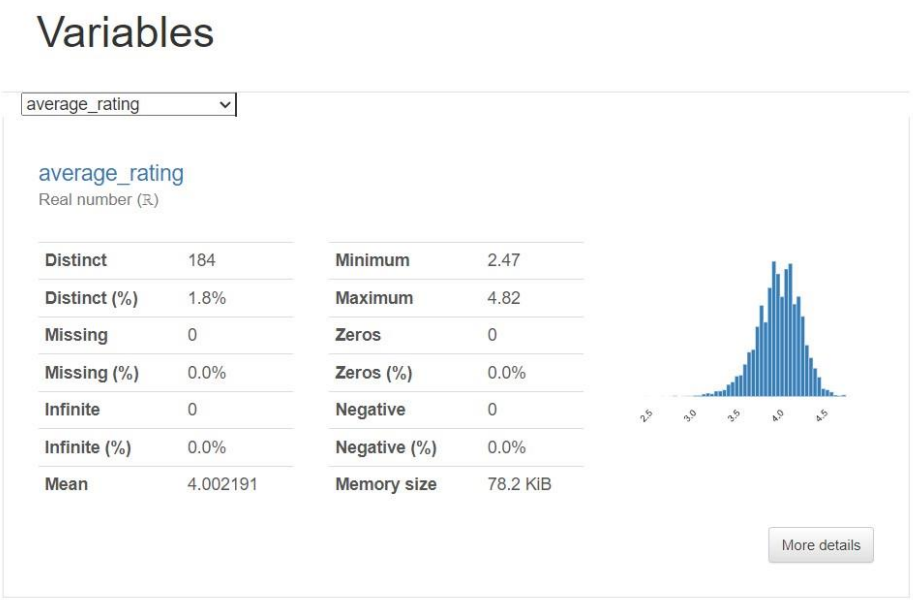
Γενικά τα στοιχεία που μας δίνει είναι 5 και αναλύονται παρακάτω

1. Η Επισκόπηση αποτελείται από συνολικά στατιστικά στοιχεία. Αυτό περιλαμβάνει τον αριθμό των μεταβλητών (χαρακτηριστικά ή στήλες του πλαισίου δεδομένων), τον αριθμό των παρατηρήσεων (σειρές του πλαισίου δεδομένων), τα κελιά που λείπουν, το ποσοστό των κελιών που λείπουν, τις διπλότυπες σειρές, το ποσοστό των διπλότυπων σειρών και το Συνολικό μέγεθος στη μνήμη.



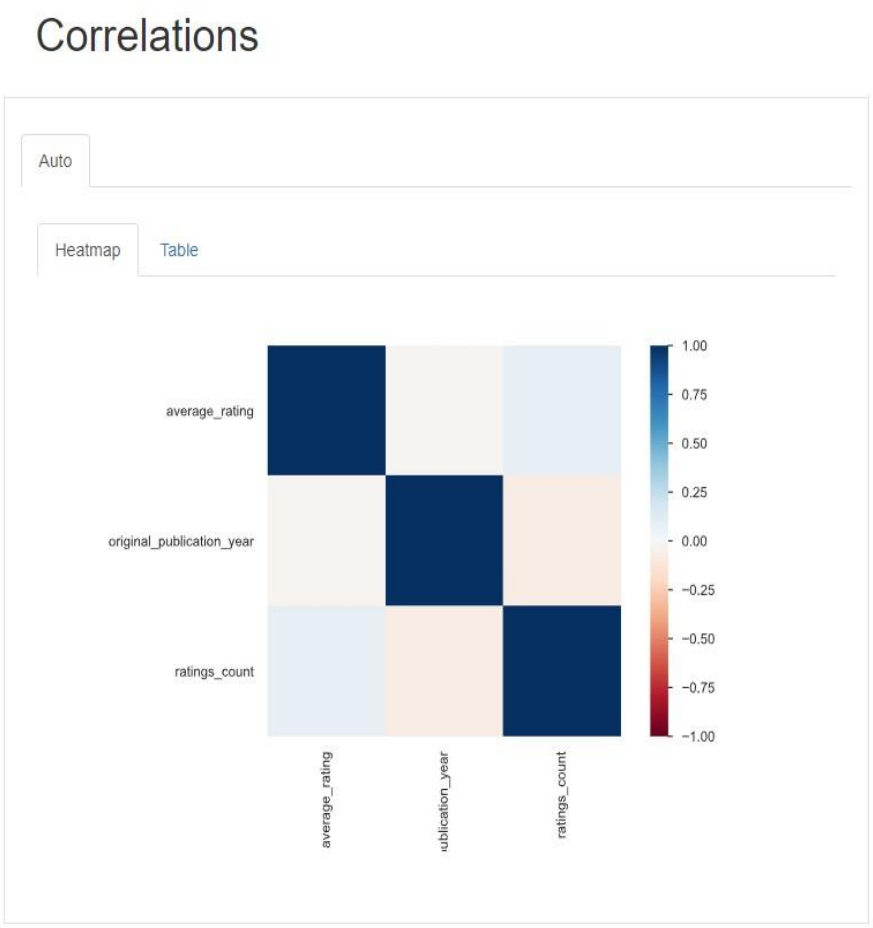
Εικόνα 18: Overview

2. Αυτή η ενότητα της αναφοράς παρέχει μια λεπτομερή ανάλυση όλων των μεταβλητών/στηλών/χαρακτηριστικών του συνόλου δεδομένων. Οι πληροφορίες που παρουσιάζονται ποικίλλουν ανάλογα με τον τύπο δεδομένων της μεταβλητής.



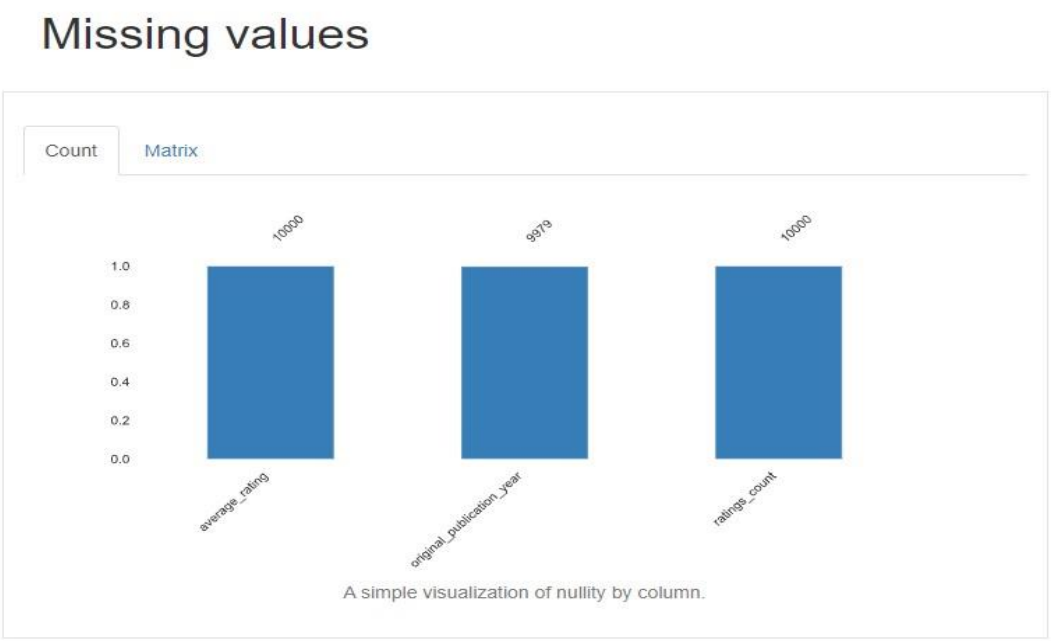
Εικόνα 19:Variables

3. Η συσχέτιση χρησιμοποιείται για να περιγράψει τον βαθμό στον οποίο δύο μεταβλητές κινούνται σε συντονισμό μεταξύ τους



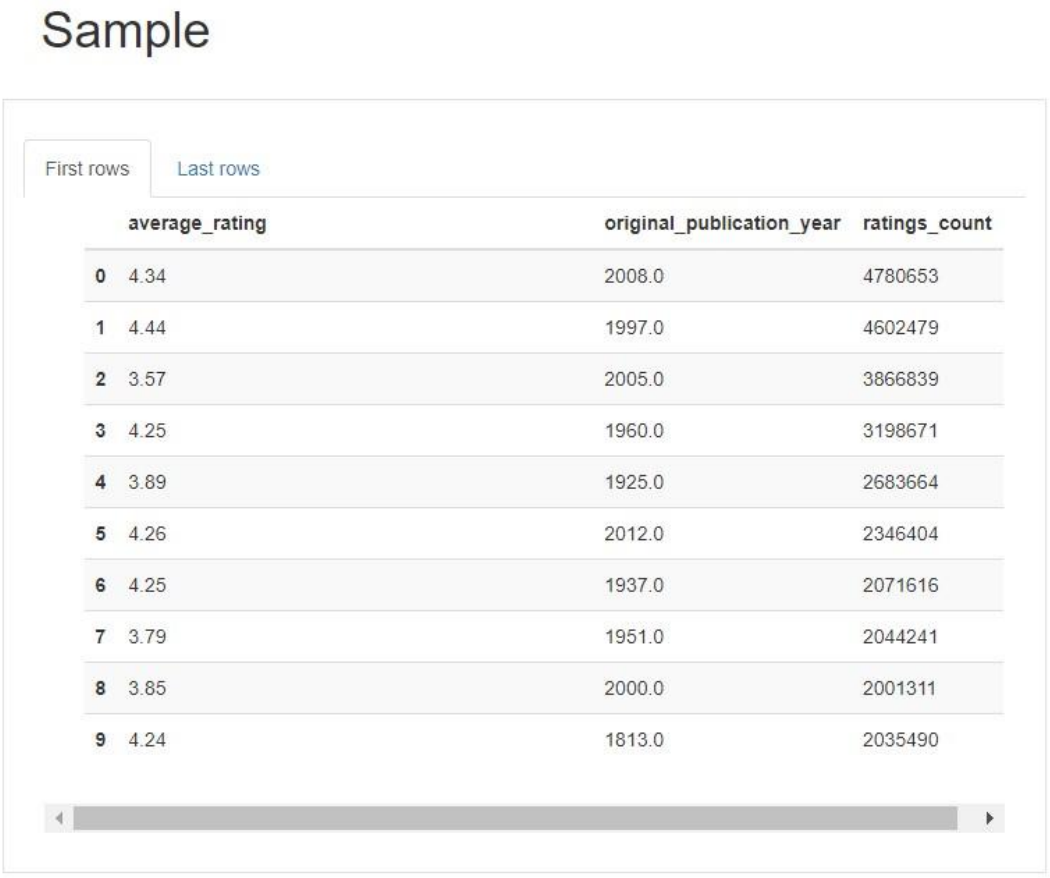
Εικόνα 20: correlation

4. Η αναφορά που δημιουργήθηκε περιέχει επίσης τις απεικονίσεις για τις τιμές που λείπουν που υπάρχουν στο σύνολο δεδομένων. Παίρνετε 2 τύπους γραφικής παράστασης: Αρίθμηση, μήτρα. Το διάγραμμα μέτρησης είναι ένα βασικό διάγραμμα ράβδων με έναν άξονα x ως ονόματα στηλών και το μήκος της ράβδου αντιπροσωπεύει τον αριθμό των παρόντων τιμών (χωρίς μηδενικές τιμές). Παρόμοια είναι η μήτρα.



Εικόνα 21: Missing values

5. Αυτή η ενότητα εμφανίζει την πρώτη και τις τελευταίες 10 σειρές του συνόλου δεδομένων.



	average_rating	original_publication_year	ratings_count
0	4.34	2008.0	4780653
1	4.44	1997.0	4602479
2	3.57	2005.0	3866839
3	4.25	1960.0	3198671
4	3.89	1925.0	2683664
5	4.26	2012.0	2346404
6	4.25	1937.0	2071616
7	3.79	1951.0	2044241
8	3.85	2000.0	2001311
9	4.24	1813.0	2035490

Εικόνα 22: Rating data

4.2 Ροή προβλήματος

Αρχικά το πρόγραμμα εισάγει τις διάφορες βιβλιοθήκες που έχουν αναφερθεί και αναλυθεί προηγουμένως κα είναι οι surprise και pandas με τις βιβλιοθήκες τους .

Η πρώτη ενότητα του κώδικα φορτώνει το ml-latest-small σύνολο δεδομένων και εκπαιδεύει ένα μοντέλο συνεργατικού φιλτραρίσματος χρησιμοποιώντας Singular Value Decomposition (SVD) από τη βιβλιοθήκη Surprise. Το αρχείο ratings.csv διαβάζεται χρησιμοποιώντας Pandas και μετατρέπεται σε αντικείμενο Surprise Dataset χρησιμοποιώντας την κλάση Reader. Στη συνέχεια, τα δεδομένα χωρίζονται σε ένα σύνολο εκπαίδευσης χρησιμοποιώντας τη μέθοδο build_full_trainset(). Όλη αυτή η διεργασία γίνεται για να χρησιμοποιηθεί το μοντέλο έπειτα στο σύνολο δεδομένων του goodbooks . Μετά την εκπαίδευση, το μοντέλο έχει μάθει για τις σχέσεις μεταξύ των χρηστών και των ταινιών και έχει αποκτήσει γνώσεις για το πώς να κάνει συστάσεις με βάση αυτές τις σχέσεις.

Στη συνέχεια, το μοντέλο ρυθμίζεται με ακρίβεια (fine-tune) στο σύνολο δεδομένων goodbooks-10k, το οποίο περιέχει αξιολογήσεις βιβλίων. Το fine-tune αναφέρεται στη διαδικασία περαιτέρω εκπαίδευσης του προεκπαιδευμένου μοντέλου στο νέο σύνολο δεδομένων για την προσαρμογή του στη νέα εργασία. Προσαρμόζοντας το προεκπαιδευμένο μοντέλο στο νέο σύνολο δεδομένων, το μοντέλο μπορεί να αξιοποιήσει τη γνώση που έχει μάθει από το ml-latest-small σύνολο δεδομένων για να βελτιώσει την απόδοσή του στη νέα εργασία της πρότασης βιβλίων.

Έπειτα ρυθμίζει με ακρίβεια το μοντέλο συνεργατικού φιλτραρίσματος στο σύνολο δεδομένων goodbooks-10k φορτώνοντας το σύνολο δεδομένων, δημιουργώντας ένα νέο σύνολο και προσαρμόζοντας το μοντέλο στο προ εκπαιδευμένο σύνολο χρησιμοποιώντας τη μέθοδο fit().

Για να γίνει fine-tune του προεκπαιδευμένου μοντέλου SVD στο σύνολο δεδομένων goodbooks-10k, τα δεδομένα αξιολογήσεων(ratings) από αυτό το σύνολο δεδομένων φορτώνονται και μετατρέπονται σε μορφή που είναι συμβατή με τη βιβλιοθήκη Surprise που χρησιμοποιείται για τη δημιουργία και την εκπαίδευση του μοντέλου SVD. Η κλίμακα αξιολόγησης ορίζεται επίσης σε (0,5, 5,0), η οποία είναι η ίδια κλίμακα που χρησιμοποιείται στο σύνολο δεδομένων ml-latest-small.

Πριν δούμε για ένα συγκεκριμένο χρήστη τις προτάσεις καλό είναι να δούμε ότι το rmse για το συγκεκριμένο σύνολο το οποίο κρατώντας ένα train set της τάξης του 20% οποίο διαλέγεται τυχαία και random state 42, όπως θα γίνει και στο κομμάτι χωρίς μεταφοράς γνώσης, το αποτέλεσμα του rmse είναι 0,62 .

Αυτό μπορεί να γίνει αντιληπτό και από την εικόνα 23 όπου πάνω αναγράφεται για τα δεδομένα που προανέφερα η εκτύπωση του προγράμματος για το rmse.

Έπειτα καλό θα ήταν να γίνει αναφορά ότι το κοινό σημείο και στα δύο dataset είναι το ratings. Πιο συγκεκριμένα η μεταφορά γνώσης σε αυτόν το πρόβλημα περιλαμβάνει τη χρήση της γνώσης που αποκτήθηκε από το σύνολο δεδομένων ml-latest-small για τη βελτίωση της απόδοσης του μοντέλου συλλογικού φιλτραρίσματος στο σύνολο δεδομένων goodbooks-10k. Προσαρμόζοντας το εκ των προτέρων εκπαιδευμένο μοντέλο στο νέο σύνολο δεδομένων, το μοντέλο μπορεί να μάθει από τα νέα δεδομένα και να προσαρμοστεί στα χαρακτηριστικά του συνόλου δεδομένων στόχου.

Οι μαθημένες αναπαραστάσεις και τα βάρη του προ-εκπαιδευμένου μοντέλου επαναχρησιμοποιούνται ως το σημείο εκκίνησης για τη λεπτομερή ρύθμιση στο νέο σύνολο δεδομένων. Αυτή η διαδικασία επιτρέπει στο μοντέλο να μάθει καλύτερες αναπαραστάσεις των προτιμήσεων των χρηστών και να βελτιώσει την ακρίβεια των προτάσεων και αυτό μπορεί να οδηγήσει σε καλύτερη απόδοση στο στόχο της πρότασης βιβλίων στους χρήστες με βάση τις προηγούμενες αξιολογήσεις τους.

Αυτό δεν σημαίνει ότι με την ύπαρξη παραπάνω παραμέτρων πέρα από το rating δεν θα ήταν πιο κατανοητή η ύπαρξη της μεταφοράς γνώσης αντιθέτως όπως αναφέρεται και στα συμπεράσματα η εξέλιξη της εργασίας, η εκτύπωση των αποτελεσμάτων αλλά και το rmse θα μπορούσε να ήταν καλύτερο αν υπήρχε Dataset Expansion , η οποία έννοια αναφέρεται παρακάτω στα συμπεράσματα.

Αφού γίνει το fine tune το μοντέλο στο σύνολο δεδομένων goodbooks-10k, μπορεί να χρησιμοποιηθεί για την υποβολή προτάσεων για έναν συγκεκριμένο χρήστη (user_id = 42) με βάση τα βιβλία που δεν έχουν ακόμη βαθμολογήσει. Για να γίνει αυτό, ο κώδικας ανακτά πρώτα τη λίστα των βιβλίων που έχει ήδη βαθμολογήσει ο χρήστης και δημιουργεί ένα σύνολο από όλα τα βιβλία στο σύνολο δεδομένων. Στη συνέχεια υπολογίζει τη διαφορά συνόλου μεταξύ των δύο συνόλων για να αποκτήσει τη λίστα των βιβλίων χωρίς αξιολόγηση. Για καθένα από αυτά τα βιβλία, ο κώδικας χρησιμοποιεί τη μέθοδο πρόβλεψης του αλγόριθμου SVD για την εκτίμηση της βαθμολογίας του χρήστη. Οι προβλέψεις αποθηκεύονται σε μια λίστα πλειάδων, όπου κάθε πλειάδα περιέχει το αναγνωριστικό βιβλίου και την προβλεπόμενη βαθμολογία.

Ακόμη το μοντέλο συνεργατικού φιλτραρίσματος αξιολογείται στο δοκιμαστικό σύνολο χρησιμοποιώντας τη μέτρηση RMSE από τη βιβλιοθήκη Surpise. Αυτό είναι σημαντικό γιατί αργότερα θα συγκρίνουμε αυτό το RMSE με το RMSE ενός μοντέλου όπου δεν έχει χρησιμοποιηθεί το fine tuning.

Το επόμενο βήμα είναι να δημιουργήσετε συστάσεις για έναν συγκεκριμένο χρήστη (user_id = 42) με βάση τα βιβλία που δεν έχουν ακόμη βαθμολογήσει. Για να γίνει αυτό, ο κώδικας ανακτά πρώτα τη λίστα των βιβλίων που έχει ήδη βαθμολογήσει ο χρήστης και δημιουργεί ένα σύνολο από όλα τα βιβλία στο σύνολο δεδομένων. Στη συνέχεια υπολογίζει τη διαφορά συνόλου μεταξύ των δύο συνόλων για να αποκτήσει τη λίστα των βιβλίων χωρίς αξιολόγηση. Για καθένα από αυτά τα βιβλία, ο κώδικας χρησιμοποιεί τη μέθοδο πρόβλεψης του αλγόριθμου SVD για την εκτίμηση της βαθμολογίας του χρήστη. Οι προβλέψεις αποθηκεύονται σε μια λίστα πλειάδων, όπου κάθε πλειάδα περιέχει το αναγνωριστικό βιβλίου και την προβλεπόμενη βαθμολογία.

Στη συνέχεια, αυτές οι προβλεπόμενες βαθμολογίες ταξινομούνται με φθίνουσα σειρά για να αποκτηθούν τα κορυφαία προτεινόμενα βιβλία.

Τέλος, εκτυπώνει τα 10 καλύτερα προτεινόμενα βιβλία στον χρήστη, συμπεριλαμβανομένου του τίτλου του βιβλίου, του συγγραφέα, της προβλεπόμενης βαθμολογίας.

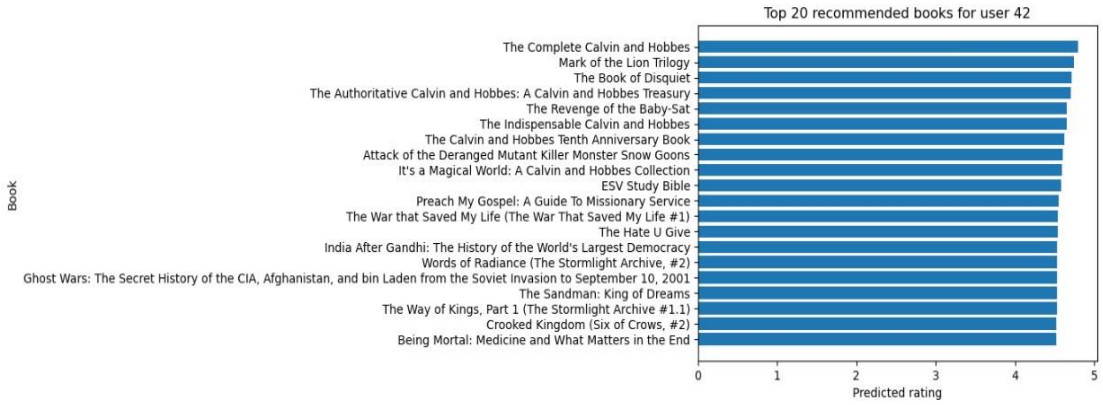
Τα αποτελέσματα για έναν τυχαίο χρήστη , συγκεκριμένα εδώ τον χρήστη 42 καθώς και το rmse φαίνονται στις παρακάτω εικόνες.

```
RMSE: 0.6451

Top 10 recommended books:
The Complete Calvin and Hobbes by Bill Watterson (predicted rating: 4.80, original rating: [])
Mark of the Lion Trilogy by Francine Rivers (predicted rating: 4.75, original rating: [])
The Book of Disquiet by Fernando Pessoa, Richard Zenith (predicted rating: 4.72, original rating: [])
The Authoritative Calvin and Hobbes: A Calvin and Hobbes Treasury by Bill Watterson (predicted rating: 4.71, original rating:
[])
The Revenge of the Baby-Sat by Bill Watterson (predicted rating: 4.66, original rating: [])
The Indispensable Calvin and Hobbes by Bill Watterson (predicted rating: 4.66, original rating: [])
The Calvin and Hobbes Tenth Anniversary Book by Bill Watterson (predicted rating: 4.62, original rating: [])
Attack of the Deranged Mutant Killer Monster Snow Goons by Bill Watterson (predicted rating: 4.61, original rating: [])
It's a Magical World: A Calvin and Hobbes Collection by Bill Watterson (predicted rating: 4.60, original rating: [])
ESV Study Bible by Anonymous, Lane T. Dennis, Wayne A. Grudem (predicted rating: 4.59, original rating: [])
Preach My Gospel: A Guide To Missionary Service by The Church of Jesus Christ of Latter-day Saints (predicted rating: 4.55, ori
ginal rating: [])
The War that Saved My Life (The War That Saved My Life #1) by Kimberly Brubaker Bradley, مرضیه ورشوساز (predicted rating: 4.54, o
riginal rating: [])
The Hate U Give by Angie Thomas (predicted rating: 4.54, original rating: [])
India After Gandhi: The History of the World's Largest Democracy by Ramachandra Guha (predicted rating: 4.54, original rating:
[])
Words of Radiance (The Stormlight Archive, #2) by Brandon Sanderson (predicted rating: 4.53, original rating: [])
Ghost Wars: The Secret History of the CIA, Afghanistan, and bin Laden from the Soviet Invasion to September 10, 2001 by Steve C
oll (predicted rating: 4.53, original rating: [])
The Sandman: King of Dreams by Alisa Kwitney, Neil Gaiman (predicted rating: 4.53, original rating: [])
The Way of Kings, Part 1 (The Stormlight Archive #1.1) by Brandon Sanderson (predicted rating: 4.53, original rating: [])
Crooked Kingdom (Six of Crows, #2) by Leigh Bardugo (predicted rating: 4.53, original rating: [])
Being Mortal: Medicine and What Matters in the End by Atul Gawande (predicted rating: 4.52, original rating: [])
```

Εικόνα 23: Τα δέκα καλύτερα προτεινόμενα βιβλία με αξιολόγηση

Για την καλύτερη σύγκριση μεταξύ των αποτελεσμάτων του μοντέλου που έχει εκπαιδευτεί αλλά και του μοντέλου που θα αναλυθεί παρακάτω και δεν έχει μεταφορά γνώσης στο επόμενο διάγραμμα είναι 20 ταινίες ώστε να μπορεί όποιος θέλει να συγκρίνει τα αποτελέσματα και να δει ομοιότητες και διαφορές.



Εικόνα 24: Ραβδοειδής αναπαράσταση αποτελεσμάτων

Πριν ξεκινήσει η ανάλυση και του δεύτερου μοντέλου χωρίς μεταφορά γνώσης καλό θα ήταν να αναφερθούμε πιο αναλυτικά για τον τρόπο με το οποίο χρησιμοποιείται αυτή μέσα στην ροή του προβλήματος.

Η μεταφορά γνώσης αναφέρεται σε μια τεχνική μηχανικής μάθησης όπου ένα μοντέλο εκπαιδεύεται σε μια εργασία και στη συνέχεια μεταφέρεται ή ρυθμίζεται με ακρίβεια για να εκτελέσει μια άλλη εργασία. Στο πλαίσιο των συστημάτων συστάσεων, η μάθηση μεταφοράς μπορεί να χρησιμοποιηθεί για τη βελτίωση της απόδοσης ενός μοντέλου συστάσεων σε ένα σύνολο δεδομένων στόχου, αξιοποιώντας τη γνώση που αποκτήθηκε από ένα σύνολο δεδομένων πηγής.

4.3 Μοντέλο χωρίς μεταφορά γνώσης.

Πρώτον, φορτώνεται το σύνολο δεδομένων goodbooks-10k που περιέχει πληροφορίες σχετικά με βιβλία και τις αξιολογήσεις τους. Χρησιμοποιεί panda για τη φόρτωση δύο αρχείων CSV: ratings.csv που περιέχει πληροφορίες σχετικά με τους χρήστες και τις αξιολογήσεις βιβλίων τους και το books.csv που περιέχει πληροφορίες για τα ίδια τα βιβλία.

Στη συνέχεια, ο κώδικας της εργασίας χωρίζει το σύνολο δεδομένων σε σύνολα εκπαίδευσης και δοκιμής . Ο διαχωρισμός πραγματοποιείται τυχαία, με το 20% των δεδομένων να δεσμεύεται για δοκιμή.

Έπειτα, στην εργασία εκπαιδεύεται ένα μοντέλο συνεργατικού φιλτραρίσματος χρησιμοποιώντας τον αλγόριθμο Singular Value Decomposition (SVD) που εφαρμόζεται στη βιβλιοθήκη έκπληξης. Η κλάση Reader χρησιμοποιείται για τον καθορισμό της κλίμακας αξιολόγησης (σε αυτήν την περίπτωση, από 0,5 έως 5,0) και η κλάση Dataset χρησιμοποιείται για τη φόρτωση των δεδομένων εκπαίδευσης σε μια μορφή που μπορεί να χρησιμοποιηθεί από τον αλγόριθμο SVD.

Μετά την εκπαίδευση του μοντέλου, αξιολογείται η απόδοσή του στο δοκιμαστικό σύνολο χρησιμοποιώντας τη μέτρηση ριζικού μέσου τετραγωνικού σφάλματος (RMSE) από τη βιβλιοθήκη Surpise.

Στη συνέχεια, προσδιορίζεται από το πρόβλημα, τα βιβλία που ο χρήστης με αναγνωριστικό 42 δεν έχει ακόμη βαθμολογήσει, συγκρίνοντας τη λίστα των βιβλίων στο σύνολο δεδομένων με τη λίστα βιβλίων που έχει βαθμολογήσει ο χρήστης.

Τέλος, δημιουργούνται προβλέψεις για τον χρήστη στα βιβλία χωρίς αξιολόγηση χρησιμοποιώντας το εκπαιδευμένο μοντέλο SVD. Κάνει βρόχο στη λίστα των βιβλίων χωρίς αξιολόγηση και για κάθε βιβλίο καλεί τη μέθοδο πρόβλεψης του μοντέλου SVD για να λάβει μια προβλεπόμενη βαθμολογία για τον χρήστη. Στη συνέχεια, αποθηκεύει αυτές τις προβλέψεις σε μια λίστα πλειάδων, όπου κάθε πλειάδα περιέχει το αναγνωριστικό βιβλίου και την προβλεπόμενη βαθμολογία.

Αυτό έχεις ως αποτέλεσμα, η λίστα με τις προβλέψεις να ταξινομείται με φθίνουσα σειρά της προβλεπόμενης βαθμολογίας και οι 10 κορυφαίες προτάσεις να εκτυπώνονται στην κονσόλα. Για κάθε προτεινόμενο βιβλίο, εκτυπώνεται ο τίτλος, ο συγγραφέας και η προβλεπόμενη βαθμολογία, μαζί με την αρχική βαθμολογία του χρήστη, εάν έχει ήδη βαθμολογήσει το βιβλίο. Το τελευταίο που αναφέρεται ως προς το αν έχει βαθμολογηθεί το βιβλίο είναι μία μέθοδος αποφυγής λάθους καθώς αν μας δείχνει τιμή τότε το πρόγραμμα παίρνει και ήδη βαθμολογημένα βιβλία για να προτείνει στον χρήστη το οποίο

θα ήταν λάθος.

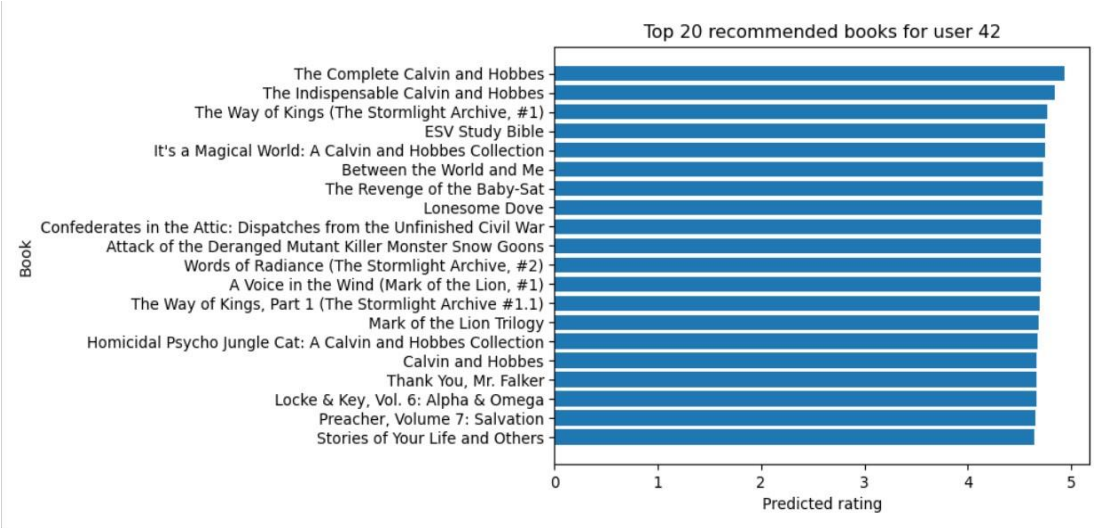
Παρακάτω υπάρχουν τα αποτελέσματα για αυτό το κομμάτι της εργασίας ως προς το RMSE αλλά και τα προτεινόμενα βιβλία.

Αρχικά έχουμε πάλι τις προτάσεις μαζί με την ανάδειξη του RMSE.

```
RMSE: 0.8294
Top 10 recommended books:
The Complete Calvin and Hobbes by Bill Watterson (predicted rating: 4.93, original rating: [])
The Indispensable Calvin and Hobbes by Bill Watterson (predicted rating: 4.84, original rating: [])
The Way of Kings (The Stormlight Archive, #1) by Brandon Sanderson (predicted rating: 4.77, original rating: [])
ESV Study Bible by Anonymous, Lane T. Dennis, Wayne A. Grudem (predicted rating: 4.75, original rating: [])
It's a Magical World: A Calvin and Hobbes Collection by Bill Watterson (predicted rating: 4.75, original rating: [])
Between the World and Me by Ta-Nehisi Coates (predicted rating: 4.73, original rating: [])
The Revenge of the Baby-Sat by Bill Watterson (predicted rating: 4.73, original rating: [])
Lonesome Dove by Larry McMurtry (predicted rating: 4.72, original rating: [])
Confederates in the Attic: Dispatches from the Unfinished Civil War by Tony Horwitz (predicted rating: 4.71, original rating:
[])
Attack of the Deranged Mutant Killer Monster Snow Goons by Bill Watterson (predicted rating: 4.70, original rating: [])
Words of Radiance (The Stormlight Archive, #2) by Brandon Sanderson (predicted rating: 4.70, original rating: [])
A Voice in the Wind (Mark of the Lion, #1) by Francine Rivers, Richard Ferrone (predicted rating: 4.70, original rating: [])
The Way of Kings, Part 1 (The Stormlight Archive #1.1) by Brandon Sanderson (predicted rating: 4.70, original rating: [])
Mark of the Lion Trilogy by Francine Rivers (predicted rating: 4.69, original rating: [])
Homicidal Psycho Jungle Cat: A Calvin and Hobbes Collection by Bill Watterson (predicted rating: 4.67, original rating: [])
Calvin and Hobbes by Bill Watterson, G.B. Trudeau (predicted rating: 4.67, original rating: [])
Thank You, Mr. Falker by Patricia Polacco (predicted rating: 4.66, original rating: [])
Locke & Key, Vol. 6: Alpha & Omega by Joe Hill, Gabriel Rodríguez (predicted rating: 4.66, original rating: [])
Preacher, Volume 7: Salvation by Garth Ennis, Steve Dillon (predicted rating: 4.66, original rating: [])
Stories of Your Life and Others by Ted Chiang (predicted rating: 4.65, original rating: [])
```

Εικόνα 25: Αποτελέσματα μοντέλου χωρίς transfer learning

Καθώς και σε διάγραμμα όπως προηγουμένως.



Εικόνα 26: Αποτελέσματα σε διάγραμμα.

Τέλος είναι σημαντικό να αναφερθεί ότι έχουν χρησιμοποιηθεί τα ίδια dataset για την εκπαίδευση όλων των μοντέλων. Εάν η προσέγγιση με μεταφορά γνώσης είχε ως αποτέλεσμα ένα μοντέλο με χαμηλότερο RMSE από το μοντέλο

που εκπαιδεύτηκε από την αρχή στο σύνολο δεδομένων goodbooks-10k, τότε η προσέγγιση εκμάθησης μεταφοράς γνώσης είναι καλύτερη για αυτήν τη συγκεκριμένη εργασία. Ωστόσο, η πραγματική απόδοση μπορεί να εξαρτάται από πολλούς παράγοντες.

4.4 Σύγκριση του μοντέλου με το μοντέλο χωρίς μεταφορά γνώσης.

Όταν συγκρίνουμε την προσέγγιση μεταφοράς γνώσης με την εκπαίδευση ενός μοντέλου από την αρχή, θα πρέπει να αξιολογήσουμε την απόδοση και των δύο προσεγγίσεων στο ίδιο σύνολο δοκιμών.

Γενικά, η μεταφορά γνώσης μπορεί να οδηγήσει σε καλύτερη απόδοση από την εκπαίδευση ενός μοντέλου από την αρχή, ειδικά όταν το σύνολο δεδομένων στόχου είναι μικρό και το σύνολο δεδομένων πηγής είναι μεγάλο και παρόμοιο. Ωστόσο, η πραγματική απόδοση εξαρτάται από πολλούς παράγοντες, όπως η ποιότητα των συνόλων δεδομένων πηγής και στόχου, η ομοιότητα μεταξύ των συνόλων δεδομένων, η επιλογή του μοντέλου και των υπερπαραμέτρων και οι μετρήσεις αξιολόγησης που χρησιμοποιούνται.

Οπότε συγκρίνουμε τα αποτελέσματα αυτών των εργασιών τόσο ως προς το RMSE όσο και ως προς τα 20 καλύτερα βιβλία που προτείνουν στον χρήστη.

Το πρώτο κομμάτι της εργασίας είναι ένα παράδειγμα μεταφοράς γνώσης, όπου ένα προ εκπαιδευμένο μοντέλο είναι fine tune σε ένα νέο σύνολο δεδομένων. Το προ εκπαιδευμένο μοντέλο εκπαιδεύεται στο ml-latest-small σύνολο δεδομένων και, στη συνέχεια, ρυθμίζεται με ακρίβεια στο σύνολο δεδομένων goodbooks-10k. Στη συνέχεια, το βελτιωμένο μοντέλο χρησιμοποιείται για να κάνει συστάσεις για έναν συγκεκριμένο χρήστη.

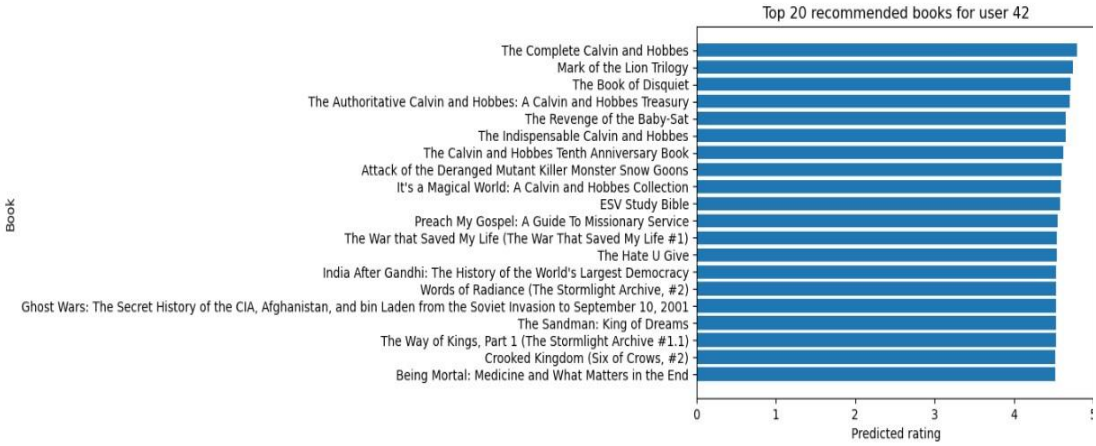
Το δεύτερο κομμάτι , από την άλλη πλευρά, εκπαιδεύει ένα μοντέλο συνεργατικού φιλτραρίσματος από την αρχή στο σύνολο δεδομένων goodbooks-10k. Στη συνέχεια, το μοντέλο χρησιμοποιείται για να κάνει συστάσεις για έναν συγκεκριμένο χρήστη.

Όσον αφορά την αξιολόγηση, και οι δύο εργασίες χρησιμοποιούν τη μέτρηση RMSE για να αξιολογήσουν την απόδοση του μοντέλου συνεργατικού φιλτραρίσματος στο δοκιμαστικό σύνολο.

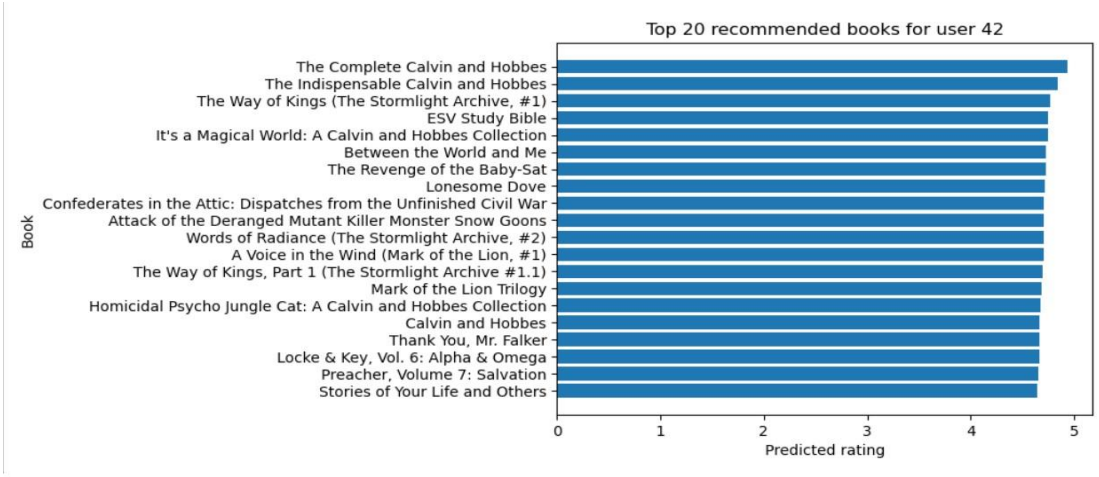
Τέλος, και οι δύο κώδικες κάνουν προβλέψεις για έναν συγκεκριμένο χρήστη στα μη βαθμολογημένα στοιχεία και τα ταξινομούν κατά φθίνουσα σειρά της προβλεπόμενης βαθμολογίας.

Οπότε είναι η στιγμή να συγκριθούν και των δύο τα αποτελέσματα και αν δούμε αν έχουν κοινά βιβλία καθώς και τι rating δίνουν και πόσο RMSE έχει το καθένα.

Αρχικά έχουμε τις εικόνες από πριν οι οποίες θα μπουν δίπλα η μία στην άλλη ώστε να είναι πιο εύκολη η σύγκριση.



Εικόνα 27: goodbooks with transfer learning



Εικόνα 28: goodbooks without transfer learning

Όπως μπορεί να γίνει αντιληπτό το the complete calvin and hobes είναι το ίδιο και για τους δύο . Ενώ άλλες ταινίες είναι είτε σε άλλη σειρά ή καθόλου μέσα στις top 20.

Αυτό εν μέρη μπορεί να δικαιολογηθεί και από το RMSE καθώς διαφέρουν τα goodbooks με transfer learning και αυτά χωρίς transfer learning καθώς έχουν RMSE 0,64 και 0,82 αντίστοιχα.

Ακόμη και για τα δύο μοντέλα ο διαχωρισμός ο οποίος έγινε στα δεδομένα τους είναι της τάξης του 20% και random state 42 περισσότερο αυτό γίνεται για να υπάρχει όσο το δυνατόν μεγαλύτερη ομοιότητα μέσα στους κώδικες και έτσι το τελικό rmse να είναι όσο πιο κοντά γίνεται σε ακρίβεια χωρίς πολλούς άλλους παράγοντες . Παρακάτω έχουμε και ένα πίνακα όπου αναφέρεται μετά από

αρκετές δοκιμές πόσα αποτελέσματα παρατηρήσαμε ότι εμφανίζονται τα ίδια στα τοπ 10 για τον ίδιο χρήστη παρότι τα rmse δεν έχουν μεγάλες μεταβολές.

Με Transfer learning Rmse	Χωρίς Transfer Learning Rmse	Και στα δύο έχουμε test dataset	Ομοιότητα μεταξύ προτάσεων βιβλίων με και χωρίς μεταφοράς γνώσης στα τοπ 20
0,64	0,82	20%	7 στα 20

Βασικό σε αυτό το σημείο είναι ότι τα 7 στα 20 περίπου δεν εκτυπώνονται στην ίδια σειρά πέρα από το πρώτο που βγαίνει το ίδιο και στα δύο, τα υπόλοιπα βγαίνουν σε διαφορετική σειρά με απόκλιση συνήθως όχι μεγαλύτερη από 6 θέσεις . Στο παράδειγμα στην εικόνα 27 και 28 προφανώς υπάρχει και παράδειγμα με απόσταση μεγαλύτερη από 6 θέσεις όπως το mark of the lion αλλά αυτό δεν είναι σύνηθες.

4.5 Αξιολόγηση αποτελεσμάτων

Το μοντέλο χρησιμοποιείται για να κάνει προβλέψεις στο σύνολο δοκιμών χρησιμοποιώντας τη μέθοδο test(). Αυτές οι προβλέψεις εκτιμούν τις βαθμολογίες που θα έδιναν οι χρήστες στα στοιχεία του δοκιμαστικού σετ.

Υπολογισμός RMSE: Οι προβλέψεις που γίνονται από το μοντέλο συγκρίνονται με τις πραγματικές αξιολογήσεις στο σύνολο δοκιμών. Η συνάρτηση rmse() από τη **accuracy** στο Surprise χρησιμοποιείται για τον υπολογισμό της τιμής RMSE. Το RMSE μετρά τη μέση διαφορά μεταξύ της προβλεπόμενης και της πραγματικής βαθμολογίας, παρέχοντας μια συνολική αξιολόγηση της ακρίβειας του μοντέλου.

Ερμηνεία RMSE: Μια χαμηλότερη τιμή RMSE υποδηλώνει καλύτερη απόδοση, καθώς σημαίνει ότι οι προβλέψεις του μοντέλου είναι πιο κοντά στις πραγματικές αξιολογήσεις. Η τιμή RMSE μπορεί να ερμηνευθεί ως το μέσο σφάλμα πρόβλεψης, με χαμηλότερες τιμές που υποδεικνύουν καλύτερη ακρίβεια στην καταγραφή των προτιμήσεων των χρηστών.

Επιπλέον, το πρόγραμμα δημιουργεί εξατομικευμένες προτάσεις βιβλίων για συγκεκριμένο χρήστη με βάση το βελτιωμένο μοντέλο. Αυτές οι προτάσεις ταξινομούνται με βάση τις προβλεπόμενες αξιολογήσεις και παρουσιάζονται στον χρήστη. Ωστόσο, οι αρχικές βαθμολογίες που παρέχονται από τον χρήστη για αυτά τα βιβλία εμφανίζονται επίσης μαζί με τις προβλεπόμενες βαθμολογίες για αναφορά και αποφυγή σφαλμάτων.

Συνολικά, η διαδικασία αξιολόγησης σε αυτό το πρόγραμμα εστιάζει στον ποσοτικό προσδιορισμό της ακρίβειας του μοντέλου συνεργατικού φιλτραρίσματος χρησιμοποιώντας το RMSE και στη δημιουργία εξατομικευμένων συστάσεων με βάση το λεπτομερώς συντονισμένο μοντέλο. Παρέχει πληροφορίες για την απόδοση του μοντέλου και την ικανότητά του να κάνει ακριβείς προβλέψεις για τα δεδομένα .

Παρόλα αυτά δύο ακόμη πολύ δημοφιλείς μετρήσεις είναι η ακρίβεια και η ανάκληση για την αξιολόγηση των συστημάτων ανάκτησης πληροφοριών. Το 1968, ο Cleverdon τα πρότεινε ως βασικές μετρήσεις [40], και έχουν κρατήσει από τότε. Για την αξιολόγηση των συστημάτων συστάσεων, έχουν χρησιμοποιηθεί από τους Billsus και Pazzani , [41]Οι Basu et al. [42]και Sarwar et al. [43].

Η ακρίβεια και η ανάκληση υπολογίζονται από έναν πίνακα 2 × 2 . Το σύνολο στοιχείων πρέπει να χωριστεί σε δύο κατηγορίες — σχετικές ή μη σχετικό. Δηλαδή, εάν η κλίμακα αξιολόγησης δεν είναι ήδη δυαδική, πρέπει να μετασχηματίσουμε σε δυαδική κλίμακα. Για παράδειγμα, το σύνολο δεδομένων MovieLens [44] έχει κλίμακα βαθμολογίας 1–5 και συνήθως μετατρέπεται σε δυαδική κλίμακα από μετατροπή κάθε βαθμολογίας 4 ή 5 σε

"σχετική" και όλες οι αξιολογήσεις 1–3 σε "μη σχετικές". Για ακρίβεια και ανάκληση, πρέπει επίσης να διαχωρίσουμε το σετ στοιχείου στο σύνολο που επιστράφηκε στον χρήστη (επιλεγμένο/προτεινόμενο) και το σύνολο που δεν ήταν. Υποθέτουμε ότι ο χρήστης θα εξετάσει όλα τα στοιχεία που ανακτώνται. Η ακρίβεια ορίζεται ως η αναλογία των σχετικών στοιχείων που επιλέγονται προς τον αριθμό των αντικειμένων επιλεγμένο, και η εξίσωση της φαίνεται παρακάτω.

$$P = \frac{N_{rs}}{N_s}$$

Η ακρίβεια αντιπροσωπεύει την πιθανότητα ότι ένα επιλεγμένο στοιχείο είναι σχετικό. Ανάκληση, φαίνεται στην εξίσωση παρακάτω , ορίζεται ως ο λόγος των σχετικών στοιχείων που επιλέγονται προς το σύνολο αριθμός των σχετικών διαθέσιμων στοιχείων. Η ανάκληση αντιπροσωπεύει την πιθανότητα να επιλεγεί το σχετικό στοιχείο.

$$R = \frac{N_{rs}}{N_s}$$

Η ακρίβεια και η ανάκληση εξαρτώνται από το διαχωρισμό σχετικών και μη σχετικών είδη. Ο ορισμός της «συνάφειας» και ο σωστός τρόπος υπολογισμού της ήταν μια σημαντική πηγή επιχειρημάτων στο πεδίο της ανάκτησης πληροφοριών. Οι περισσότερες αξιολογήσεις ανάκτησης πληροφοριών έχουν επικεντρωθεί σε μια αντικειμενική εκδοχή της συνάφειας, όπου η συνάφεια ορίζεται σε σχέση με ένα ερώτημα και είναι ανεξάρτητα από τον χρήστη. Ομάδες ειδικών μπορούν να συγκρίνουν έγγραφα με ερωτήματα και να καθορίσετε ποια έγγραφα σχετίζονται με ποια ερωτήματα. Ωστόσο, η αντικειμενική συνάφεια δεν έχει νόημα στα συστήματα συστάσεων. Συστήματα συστάσεων προτείνουν αντικείμενα με βάση την πιθανότητα ότι θα συναντήσουν έναν συγκεκριμένο χρήστη γεύση ή ενδιαφέρον. Αυτός ο χρήστης είναι το μόνο άτομο που μπορεί να καθορίσει εάν ένα στοιχείο ανταποκρίνεται στις γευστικές του απαιτήσεις. Επομένως, η συνάφεια είναι πιο εγγενώς υποκειμενική στα συστήματα συστάσεων σε σχέση με την παραδοσιακή ανάκτηση εγγράφων. [39]

Εδώ όμως χρησιμοποιείται μόνο το rmse και πάμε να δούμε το γιατί.

Το RMSE και το Precision@K είναι διαφορετικές μετρήσεις αξιολόγησης με διαφορετικούς σκοπούς:

Το RMSE παρέχει ένα συνολικό μέτρο της ακρίβειας πρόβλεψης λαμβάνοντας υπόψη ολόκληρη την εργασία πρόβλεψης αξιολόγησης. Αξιολογεί πόσο καλά προβλέπει το ίδιο το σύστημα τις αξιολογήσεις.

Το Precision@K, από την άλλη, εστιάζει στην ποιότητα των κορυφαίων K συστάσεων. Αξιολογεί την ικανότητα του συστήματος να προτείνει σχετικά στοιχεία μεταξύ ενός μικρότερου υποσυνόλου προτάσεων.

Και οι δύο μετρήσεις είναι πολύτιμες για την αξιολόγηση διαφορετικών πτυχών ενός συστήματος συστάσεων. Ενώ το RMSE μετρά την ακρίβεια των προβλέψεων αξιολόγησης, το Precision@K αξιολογεί την ικανότητα του συστήματος να παρέχει σχετικές συστάσεις. Παρέχουν συμπληρωματικές πληροφορίες για την απόδοση του συστήματος και την ικανοποίηση των χρηστών.

Εμείς χρειαζόμασταν να δούμε πόσο καλά προβλέπει το ίδιο το σύστημα τις αξιολογήσεις αλλά ήταν σημαντικό να γίνει αναφορά και σε άλλες μεθόδους που όπως θα αναλυθεί και στα συμπεράσματα μπορούν να είναι πάτημα για μετέπειτα έλεγχο και εξέλιξη του κώδικα .

5. Συμπεράσματα

Ο παρεχόμενος κώδικας δείχνει την εφαρμογή ενός συστήματος συστάσεων που βασίζεται σε συλλογικό φιλτράρισμα χρησιμοποιώντας τις βιβλιοθήκες Pandas και Surprise. Εκπαιδεύει ένα μοντέλο συνεργατικού φιλτραρίσματος στο ml-latest-small σύνολο δεδομένων, προσαρμόζει το μοντέλο στο σύνολο δεδομένων goodbooks-10k και αξιολογεί την απόδοσή του ,χρησιμοποιώντας τη μέτρηση RMSE. Δημιουργεί επίσης εξατομικευμένες προτάσεις βιβλίων για συγκεκριμένο χρήστη με βάση το βελτιωμένο μοντέλο.

Ο κώδικας δείχνει τη δύναμη των Panda για φόρτωση, χειρισμό και φιλτράρισμα δεδομένων και αξιοποιεί τις δυνατότητες της βιβλιοθήκης Surprise για εργασίες συστήματος συστάσεων. Η χρήση του αλγορίθμου SVD από το Surprise επιτρέπει την παραγοντοποίηση μήτρας, τη σύλληψη λανθανόντων παραγόντων και την πραγματοποίηση ακριβών προβλέψεων.

Η αξιολόγηση της απόδοσης του μοντέλου χρησιμοποιώντας το RMSE παρέχει πληροφορίες για την ακρίβεια των προβλέψεων αξιολόγησης, υποδεικνύοντας πόσο καλά το μοντέλο αποτυπώνει τις προτιμήσεις των χρηστών. Οι χαμηλότερες τιμές RMSE υποδηλώνουν καλύτερη ακρίβεια στην πρόβλεψη αξιολογήσεων.

Επιπλέον, ο κώδικας δείχνει τη δημιουργία εξατομικευμένων προτάσεων βιβλίων κάνοντας προβλέψεις για βιβλία χωρίς αξιολόγηση για έναν συγκεκριμένο χρήστη. Οι προτάσεις ταξινομούνται με βάση τις προβλεπόμενες βαθμολογίες, παρέχοντας πολύτιμες προτάσεις για τις μελλοντικές επιλογές βιβλίων του χρήστη.

Μια αξιοσημείωτη πτυχή του κώδικα είναι η εφαρμογή της μάθησης μεταφοράς γνώσης . Η μάθηση μεταφοράς γνώσης αναφέρεται στη διαδικασία αξιοποίησης της γνώσης από έναν τομέα ή ένα σύνολο δεδομένων για τη βελτίωση της απόδοσης σε έναν άλλο τομέα ή σύνολο δεδομένων. Σε αυτόν τον κώδικα, το μοντέλο συνεργατικού φιλτραρίσματος εκπαιδεύεται αρχικά στο σύνολο δεδομένων ml-latest-small, το οποίο περιέχει αξιολογήσεις ταινιών. Στη συνέχεια, το μοντέλο ρυθμίζεται με ακρίβεια χρησιμοποιώντας το σύνολο δεδομένων goodbooks-10k, το οποίο αποτελείται από αξιολογήσεις βιβλίων. Προσαρμόζοντας το προ εκπαιδευμένο μοντέλο στον νέο τομέα, το σύστημα συστάσεων επωφελείται από την προηγούμενη γνώση που αποκτήθηκε από τις αξιολογήσεις ταινιών, ενισχύοντας την ικανότητά του να κάνει ακριβείς προβλέψεις για βιβλία.

Η προσέγγιση εκμάθησης μεταφοράς γνώσης που χρησιμοποιείται στον κώδικα αξιοποιεί την κατανόηση των προτιμήσεων των χρηστών και των

χαρακτηριστικών στοιχείων από το μοντέλο συνεργατικού φιλτραρίσματος από την αρχική εκπαίδευση στο σύνολο δεδομένων αξιολογήσεων ταινιών. Στη συνέχεια, αυτή η γνώση εφαρμόζεται στον τομέα αξιολογήσεων βιβλίων, με αποτέλεσμα ένα πιο αποτελεσματικό σύστημα συστάσεων για βιβλία. Προσαρμόζοντας το μοντέλο στο νέο σύνολο δεδομένων, μαθαίνει να αποτυπώνει τις συγκεκριμένες αποχρώσεις και μοτίβα που υπάρχουν στις αξιολογήσεις βιβλίων, οδηγώντας σε βελτιωμένες προτάσεις.

Η εκμάθηση μεταφοράς σε αυτόν τον κώδικα επιτρέπει στο μοντέλο να ξεπεράσει τον περιορισμό της ύπαρξης περιορισμένων ή αραιών δεδομένων στον νέο τομέα. Αντί να ξεκινά από το μηδέν, αξιοποιεί την υπάρχουσα γνώση που αποκτήθηκε από έναν σχετικό τομέα για να επιταχύνει τη μάθηση και να βελτιώσει την απόδοση του συστήματος συστάσεων. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη όταν η διαθεσιμότητα δεδομένων είναι περιορισμένη ή όταν η εκπαίδευση ενός μοντέλου από την αρχή σε ένα νέο σύνολο δεδομένων μπορεί να είναι υπολογιστικά ακριβή.

Συνολικά, ο κώδικας αποτελεί παράδειγμα των πλεονεκτημάτων της μεταφοράς μάθησης σε συστήματα συστάσεων, δείχνοντας πώς τα προ εκπαιδευμένα μοντέλα μπορούν να βελτιστοποιηθούν και να εφαρμοστούν σε νέους τομείς για να αξιοποιήσουν την προηγούμενη γνώση και να βελτιώσουν την ακρίβεια των συστάσεων.

Ο παρεχόμενος κώδικας μπορεί να χρησιμεύσει ως αφετηρία για διάφορες περαιτέρω χρήσεις και επεκτάσεις:

- ** Dataset Expansion**: Ο κώδικας μπορεί να προσαρμοστεί ώστε να λειτουργεί με μεγαλύτερα ή διαφορετικά σύνολα δεδομένων. Μπορείτε να εξερευνήσετε άλλα σύνολα δεδομένων ταινιών ή βιβλίων και να τα ενσωματώσετε στη διαδικασία εκπαίδευσης ή τελειοποίησης. Αυτό θα επέτρεπε στο σύστημα συστάσεων να έχει ευρύτερο πεδίο εφαρμογής και ενδεχομένως να βελτιώσει την απόδοσή του.

- ** Hyperparameter Tuning**: Ο κώδικας μπορεί να βελτιωθεί με την ενσωμάτωση τεχνικών συντονισμού υπερπαραμέτρων για τη βελτιστοποίηση της απόδοσης του μοντέλου συνεργατικού φιλτραρίσματος. Τεχνικές όπως η αναζήτηση πλέγματος ή η τυχαία αναζήτηση μπορούν να εφαρμοστούν για να βρεθεί ο καλύτερος συνδυασμός υπερπαραμέτρων για τον αλγόριθμο SVD ή άλλους αλγόριθμους συστάσεων που είναι διαθέσιμοι στη βιβλιοθήκη Surprise.

- ** Cold Start Problem**: Ο κωδικός μπορεί να τροποποιηθεί για να χειριστεί το πρόβλημα ψυχρής εκκίνησης, το οποίο προκύπτει όταν

υπάρχουν περιορισμένες ή καθόλου διαθέσιμες πληροφορίες για νέους χρήστες ή στοιχεία. Τεχνικές όπως το φιλτράρισμα βάσει περιεχομένου ή οι υβριδικές προσεγγίσεις μπορούν να ενσωματωθούν για την αντιμετώπιση αυτής της πρόκλησης και την παροχή ουσιαστικών συστάσεων ακόμη και για νέους χρήστες ή αντικείμενα.

- ** Real-Time Recommendations**: Ο κώδικας μπορεί να προσαρμοστεί για να παρέχει συστάσεις σε πραγματικό χρόνο ενημερώνοντας συνεχώς το μοντέλο προτάσεων με βάση τις αλληλεπιδράσεις ή τα σχόλια νέων χρηστών. Αυτό θα συνεπαγόταν τη δημιουργία ενός αγωγού για την απορρόφηση δεδομένων, την επανεκπαίδευση μοντέλων και τη δημιουργία ενημερωμένων προτάσεων.

- ** Integration with Other Systems**: Το σύστημα προτάσεων μπορεί να ενσωματωθεί με άλλα συστήματα ή εφαρμογές, όπως πλατφόρμες ηλεκτρονικού εμπορίου ή υπηρεσίες ροής περιεχομένου, για να παρέχει εξατομικευμένες προτάσεις στους χρήστες με βάση το ιστορικό περιήγησής τους, τη συμπεριφορά αγορών ή περιεχόμενο που παρακολουθήσατε.

Αυτές είναι λίγες μόνο δυνατότητες για περαιτέρω αξιοποίηση του κώδικα. Η συγκεκριμένη κατεύθυνση θα εξαρτηθεί από τις απαιτήσεις , τα διαθέσιμα δεδομένα και την επιθυμητή λειτουργικότητα του συστήματος συστάσεων.

6. Πίνακας ανάλυσης συμβόλων

Table 1:Πίνακας ανάλυσης συμβόλων

Symbol	Definition
n	Number of instances
m	Number of domains
D	Domain
T	Task
X	Feature space
y	Label space
x	Feature vector
υ	Label vector
a	Weighting coefficient
β	Weighting coefficient
λ	Tradeoff parameter
δ	Parameter/Error
b	Bias
N	Iteration/Kernel number
f	Decision function
η	Scale parameter
Θ	Model parameters

7. Εικόνες

8. Βιβλιογραφία

[1] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, Qing He Fuzhen Zhuang, A Comprehensive Survey on Transfer Learning, June 2020.

[2] Γ. Μαρία, Μεταφορά Γνώσης με χρήση Νευρωνικών Δικτύων, 2021.

[3] C. Zografou, TRANSFER LEARNING IN RECOMMENDER SYSTEMS.

[4] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, Alex Smola, Correcting sample selection bias by unlabeled data. In NIPS, pages 601–608, 2006.

[5] Maxime Oquab, Léon Bottou, Ivan Laptev, Josef Sivic, Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1717–1724)., 2014.

[6] Yaroslav Ganin, Victor Lempitsky, Unsupervised domain adaptation by backpropagation. arXiv preprint arXiv:1409.7495., 2014.

[7] Shisha Liao, Chenqiang Gao, Chengjuan Xie, Feng Yang, Yue Zhao, Atsushi Sagata,Yongqing Sun, Weakly Supervised Instance Segmentation Based on Two-Stage Transfer Learning,IEEE Access PP(99):1-1, January 2020.

[8] Sinno Jialin Pan , Ivor W Tsang, James T Kwok, Qiang Yang, "Domain Adaptation via Transfer Component Analysis," in IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 199-210,doi: 10.1109/TNN.2010.2091281., Feb. 2011.

[9] John Blitzer, Ryan McDonald, Fernando Pereira, Domain adaptation with structural correspondence in Proc. Conference on Empirical Methods in Natural learning Language Processing, Sydney, pp. 120–128, Jul. 2006.

[10] Sinno Jialin Pan , Xiaochuan Ni , Jian-Tao Sun , Qiang Yang and Zheng Chen, Cross-domain sentiment classification via spectral feature alignment. In: Proceedings of the 19th international conference on world wide web p. 751–60, 2010.

[11] Karl Weiss, Taghi M. Khoshgoftaar & DingDing Wang, A survey of transfer learning,p.12,2016.

[12] Fuzhen Zhuang,Ping Luo,Hui Xiong,Qing He,Yuhong Xiong, Zhongzhi Shi, .Exploiting associations between word clusters and document classes for crossdomain text categorization,vol. 4, no. 1, pp. 100–114, 2011.

[13] Jing Gao,Aalborg University,Wei Fan,Jing Jiang, Jiawei Han, .Knowledge transfer via multiple model local structure mapping, in Proc. 14th ACM SIGKDD International Conference onKnowledge Discovery and Data Mining, Las Vegas, pp. 283–291, Aug. 2008.

[14] Xavier Glorot,Antoine Bordes, Y. Bengio, . Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the twenty-eight international conference on machine learning, vol. 27,. p., 2011.

[15] S Rendle, C Freudenthaler, WSDM '14: Proceedings of the 7th ACM international conference on Web search and data miningFebruary 2014,Pages 273–282

[16] Francesco Ricci,Lior Rokach ,Bracha Shapira, Paul B. Kantor, Introduction to Recommender Systems, Chapter 1.

[17] R. Burke, Hybrid web recommender systems. In: The Adaptive Web, pp. 377–408, Springer Berlin / Heidelberg, 2007.

[18] J. Schafer, Dan Frankowski, Jonathan L. Herlocker, Shilad Sen, Collaborative filtering

recommender systems. In: The Adaptive Web,pp. 291–324, Springer Berlin / Heidelberg , 2007.

[19] Francesco Ricci F. Mahmood, Towards learning user-adaptive state models in a conversational recommender system In: A. Hinneburg (ed.) LWA 2007: Lernen - Wissen - Adaption, Halle,Workshop Proceedings, pp. 373–378, 2007.

[20] DEREK BRIDGE,MEHMET H. GÖKER,LORRAINE McGINTY and BARRY SMYTH, Case-based recommender systems. The Knowledge Engineering review 20(3), 315–320, 2006.

[21] Francesco Ricci,D. Cavada,N. Mirzadeh, A. Venturini, Case-based travel recommendations. In:D.R. Fesenmaier, K. Woeber, H. Werthner (eds.) Destination Recommendation Systems: Behavioural Foundations and Applications, pp. 67–93, 2006.

[22] Ofer Arazy,Nanda Kumar, Bracha Shapira, Improving social recommender systems. IT Professional, 2007.

[23] David Ben-Shimon, Alexander Tsikinovsky, Lior Rokach, Amnon Meisles,Guy Shani, and Lihi Naamani, Recommender system from personal social networks,Advances in Soft Computing, vol. 43, pp. 47–55, 2007.

[24] Rashmi Sinha and Kirsten Swearingen, Comparing recommendations made by online systems andfriends. In: DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries, 2007.

[25] J. Golbeck, Generating predictive movie recommendations from trust in social networks,Trust Management, 4th International Conference, iTrust 2006, Pisa, Italy, May 16-19, 2006,pp. 93–104, 2006.

[26] Paolo Massa and Paolo Avesani, Trust-aware collaborative filtering for recommender systems,Proceedings of the International Conference on Cooperative Information Systems, CoopIS,pp. 492–508, 2004.

[27] Ido Guy, Naama Zwerdling, David Carmel, Inbal Ronen, Erel Uziel, Sivan Yogev, Shila Ofek-Koifman, Personalized recommendation of social software items based on social relations. In: RecSys '09: Proceedings of the third ACM conference on Recommender systems, pp. 53–60, 2009.

[28] Georg Groh,Christian Ehmig, Recommendations in taste related domains: collaborative filtering vssocial filtering. In: GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work, pp. 127–136, 2007.

[29] Marijan Beg, Juliette Taka, Thomas Kluyver, Alexander Konovalov, Min Ragan-Kelley, Nicolas M. Thiéry, Hans Fangohr, Using Jupyter for reproducible scientific workflows.

[30] Ask Hjorth Larsen 1, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, Karsten W Jacobsen,The atomic simulation environment—a python library for working with atoms, 2017.

[31] Juliana Freire,David Koop,Emanuele Santos, Claudio Silva, "Provenance for computational tasks: A survey," Computing in Science & Engineering, vol. 10,no. 3, pp. 11–21, 2008.

[32] M. Kula, Metadata Embeddings for User and Item Cold-start Recommendations, 30 Jul

2015.

[33] Martin Saveski, Amin Mantrach, Item cold-start recommendations: learning local collective embeddings. In Proceedings of the 8th ACM Conference on Recommender systems, pages 89–96., 2014.

[34] S. Rendle, BPR: Bayesian personalized ranking from implicit feedback.” Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009.

[35] Jason Weston and Samy Bengio and Nicolas Usunier2: Scaling up to large vocabulary image annotation.” IJCAI. Vol. 11, 2011.

[36] D G Altman , J M Bland, tests 3: receiver operating characteristic plots, 1994.

[37] Enrique Schisterman, Neil J Perkins, Aiyi Liu,Howard D Bondell, Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. Epidemiology,pg 73–81, 2005.

[38] Jason Weston, Hector Yee, Ron J. Weiss, Learning to Rank Recommendations with the k-Order Statistic Loss. 2013

[39] Michael D. Ekstrand, John T. Riedland Joseph A. Konstan, Evaluating Collaborative Filtering Recommender Systems,Vol. 4, No. 2 (2010) 81–17

[40] Cleverdon, Cyril W, Mills, Jack, Keen, Michael, Factors Determining the Performance of Indexing Systems., 1968.

[41] Michael J. Pazzani,Daniel BILLSUS, Learning collaborative information filters. In Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98). C. Rich, and J. Mostow, Eds.AAAI Press, Menlo Park, Calif., 46–53, 1998.

[42] Chumki Basu,Haym Hirsh,William Cohen, Recommendation as classification: using social and content-based information in recommendation. In Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98). C. Rich, and J. Mostow, Eds. AAAI Press, Menlo Park, Calif.,714–720, 1998.

[43] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, Analysis of recommendation algorithms for E-commerce. In Proceedings of the 2nd ACM Conference on Electronic Commerce (EC’00). ACM, New York. 285–295., 2001.

[44] B. Dahlen, J. Konstan, J. Herlocker, N. Good, A. Borchers, and J. Riedl., Jumpstarting movielens: User benefits of starting a collaborative filtering system with “dead data”. TR 98-017. University of Minnesota, 1998.