ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΠΤΥΞΗ ΥΒΡΙΔΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΒΑΣΙΣΜΕΝΟΥ ΣΕ ΒΑΘΕΙΑ ΜΑΘΗΣΗ

Διπλωματική Εργασία

Ιωάννης Τσέλιος

Επιβλέπων: Μιχαήλ Βασιλακόπουλος

Σεπτέμβριος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΑΝΑΠΤΥΞΗ ΥΒΡΙΔΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΒΑΣΙΣΜΕΝΟΥ ΣΕ ΒΑΘΕΙΑ ΜΑΘΗΣΗ

Διπλωματική Εργασία

Ιωάννης Τσέλιος

Επιβλέπων: Μιχαήλ Βασιλακόπουλος

Σεπτέμβριος 2022

UNIVERSITY OF THESSALY

SCHOOL OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DEVELOPMENT OF A HYBRID RECOMMENDER SYSTEM BASED ON DEEP LEARNING

Diploma Thesis

Ioannis Tselios

Supervisor: Michael Vassilakopoulos

September 2022

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων **Μιχαήλ Βασιλακόπουλος**
Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Τουσίδου Ελένη**
Ε.ΔΙ.Π., Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Μέλος **Δασκαλοπούλου Ασπασία**
Επίκουρος Καθηγήτρια, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κύριο Βασιλακόπουλο Μιχαήλ που με βοήθησε και με συμβούλεψε καθ’όλη την διάρκεια της εκπόνησης της διπλωματικής εργασίας. Τέλος, θα ήθελα να ευχαριστήσω τις καθηγήτριες Δασκαλοπούλου Ασπασία και Τουσίδου Ελένη που δέχτηκαν να είναι μέλη της επιτροπής.

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΠΕΡΙ ΑΚΑΔΗΜΑΪΚΗΣ ΔΕΟΝΤΟΛΟΓΙΑΣ ΚΑΙ ΠΝΕΥΜΑΤΙΚΩΝ ΔΙΚΑΙΩΜΑΤΩΝ

«Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα διπλωματική εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Δηλώνω επίσης ότι τα αποτελέσματα της εργασίας δεν έχουν χρησιμοποιηθεί για την απόκτηση άλλου πτυχίου. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής».

Ο/Η Δηλών/ούσα

Ιωάννης Τσέλιος

Διπλωματική Εργασία

ΑΝΑΠΤΥΞΗ ΥΒΡΙΔΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ ΣΥΣΤΑΣΕΩΝ ΒΑΣΙΣΜΕΝΟΥ ΣΕ ΒΑΘΕΙΑ ΜΑΘΗΣΗ

Ιωάννης Τσέλιος

Περίληψη

Με την ραγδαία αύξηση της τεχνολογίας τα τελευταία 20 χρόνια, υπάρχουν δεκάδες εφαρμογές με μεγάλο όγκο χρηστών. Οι εφαρμογές αυτές με το πέρασμα του χρόνου έγιναν πιο εύχρηστες και πιο φιλικές για τον χρήστη. Αυτό έγινε με την υλοποίηση σημαντικών λειτουργιών. Μία από αυτές τις λειτουργίες είναι και τα συστήματα συστάσεων.

Ειδικότερα, τα συστήματα συστάσεων είναι τόσο σημαντικά τόσο για τον χρήστη όσο και για την εφαρμογή. Τα αρχικά μοντέλα ήταν αρκετά απλά σαν φιλοσοφία, ωστόσο αυτά που πρόσφεραν ήταν πολύ σημαντικά. Τα πρώτα μοντέλα βασιζόταν στο υλικό που παρακολουθεί ο χρήστης σε κάποια εφαρμογή και με αυτόν τον τρόπο γινόταν οι συστάσεις. Έπειτα, σε κάποιες εφαρμογές που υπήρχαν αρκετές αλληλεπιδράσεις με άλλους χρήστες δημιουργήθηκε ένας τύπος συστήματος βασισμένο πάνω σε αυτό. Προσπαθούσαν να προβλέψουν κατά πόσο ένας χρήστης μοιάζει με κάποιον άλλο. Τέλος, υπάρχουν και τα υβριδικά συστήματα συστάσεων, τα οποία καλύπτουν και τις δύο προηγούμενες κατηγορίες.

Σε αυτή την διπλωματική θα αναλύσουμε 4 υβριδικά μοντέλα συστάσεων. Τα μοντέλα αποτελούνται κυρίως από 4 κυρίως μέρη. Όλες αυτές οι κατηγορίες θα αναλυθούν αργότερα στο κεφάλαιο 3. Τα υβριδικά μοντέλα είναι η τελευταία εξέλιξη των συστημάτων συστάσεων. Συγκεκριμένα, θα παρουσιάσουμε 4 μοντέλα για 4 διαφορετικές βάσεις δεδομένων, θα αναφέρουμε αναλυτικά την διαδικασία προεπεξεργασίας για όλες τις βάσεις δεδομένων, Θα παρουσιαστούν λεπτομερώς όλα τα μοντέλα και στο τέλος θα παρουσιαστούν τα αποτελέσματα από την αξιολόγηση των μοντέλων.

Λέξεις-κλειδιά:

Υβριδικό σύστημα συστάσεων, Σύστημα βασισμένο στο περιεχόμενο, Σύστημα συνεργατικού φιλτραρίσματος, Σύστημα που βασίζεται στην γνώση, Ενσωματωμένα στρώματα.

Diploma Thesis

DEVELOPMENT OF A HYBRID RECOMMENDER SYSTEM BASED ON DEEP LEARNING

Ioannis Tselios

Abstract

With the rapid increase in technology over the past 20 years, there are plethora of applications with a large volume of users. These applications over time became more convenient and more user-friendly. This was done by implementing important functions of sons. One of these functions is the recommendation systems.

In particular, recommendation systems are as significant for both the user and the application. The first models that they used this technology were quite simple in philosophy, yet what they offered was very important. The first models were based on the object that the user likes in an application and in this way the recommendations were made. Then, there were soma applications that several interactions were made between users creating a system based on how similar one user is to another. Finally, there are the hybrid systems, which include both previous categories.

In this thesis we will analyze 4 hybrid recommendation models. The models mainly consist of 4 main parts content-based, collaborative-filtering, collaborative-filtering with neural network and knowledge-based. All these categories will be analyzed later in the paper3. Hybrid models are the latest evolution of recommendation systems. Specifically, we will present 4 models for 4 different datasets, we will show in detail the preprocessing of all datasets. All models will be presented in detail and at the end we will discuss the results from the evaluation of models.

Keywords:

Hybrid recommendation system, Content-based system, Collaborative-filtering system, Knowledge-based system, Embedding layers, LSTM.

Πίνακας περιεχομένων

Κατάλογος σχημάτων

Κατάλογος πινάκων

Συντομογραφίες

βλπ	βλέπε
κ.λπ.	και λοιπά
κ.α.	και άλλα
π.χ.	παραδείγματος χάρη
LSTM	Long short-term memory
RNN	Recurrent neural network
HRS	Υβριδικό σύστημα συστάσεων
MSE	Μέσο Τετραγωνικό Λάθος
RMSE	Μέση τετραγωνική ρίζα σφάλματος
MAE	Μέσο απόλυτο σφάλμα

Κεφάλαιο 1

Εισαγωγή

Από το 2000 και έπειτα έχει παρατηρηθεί τεράστια χρήση εφαρμογών, οι οποίες προσφέρουν διασκέδαση στους χρήστες, όπως διάφορα μέσα κοινωνικής δικτύωσης, όπως instagram, facebook και twitter ή εφαρμογές για παρακολούθηση video ή ταινιών, όπως tik-tok, youtube και netflix. Αυτό οφείλεται, κυρίως, στην μεγάλη τεχνολογική εξέλιξη των κινητών και των υπολογιστών. Ωστόσο, με το πέρασμα των χρόνων όλες αυτές οι εφαρμογές έγιναν πιο φιλικές για τους χρήστες, καθώς επίσης και για αυτούς που δημιουργούν κάποιο περιεχόμενο στις εφαρμογές αυτές. Αυτή η σημαντική εξέλιξη πραγματοποιήθηκε λόγω της εισαγωγής των συστημάτων συστάσεων, τα οποία είναι πολύ σημαντικά για να υπάρχει ισορροπία στις εφαρμογές. Αρχικά, τα συστήματα συστάσεων είναι συστήματα, τα οποία φιλτράρουν κάποιες πληροφορίες με τις οποίες το σύστημα αυτό θα προβλέψει κατά πόσο θα αξιολογούσε ένας χρήστης ένα αντικείμενο ή κατά πόσο θα ενδιέφερε τον χρήστη αυτό το αντικείμενο. Με λίγα λόγια είναι ένα σύστημα, το οποίο προτείνει αντικείμενα σε χρήστες σχετικά με την πληροφορία που έχει φιλτράρει για τους ίδιους τους χρήστες. Υπάρχουν διάφορες χρήσεις των συστημάτων συστάσεων, όπως αρχικά να προτείνει σχετικό περιεχόμενο στο κάθε χρήστη δυναμικά, όπως το netflix[1] και το youtube[2]. Επίσης, βοηθάει στην κατηγοριοποίηση των αντικειμένων βασισμένο στα χαρακτηριστικά του αντικειμένου αυτού. Επιπλέον, υπάρχουν διάφοροι τύποι συστημάτων συστάσεων, όπως να είναι βασισμένο στο περιεχόμενο(content-based), το συνεργατικό φιλτράρισμα(collaborative filtering) και το υβριδικό σύστημα συστάσεων(hybrid recommendation system). Κάθε σύστημα από τα παραπάνω έχει πλεονεκτήματα και μειονεκτήματα ανάλογα την χρήση τους. Σε αυτή την εργασία θα πραγματοποιηθεί κυρίως ανάλυση για υβριδικό σύστημα συστάσεων, το οποίο περιλαμβάνει και τους άλλους 2 τύπους συστημάτων.

1.1 Εργαλεία

Σε αυτή την ενότητα θα παρουσιάσουμε τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας.

-Pycharm[3]: Το pycharm είναι μία πάρα πολύ δυνατή cross-platform εφαρμογή, στην οποία μπορείς να γράψεις κώδικα. Είναι συμβατό με windows, linux και macOS και υποστηρίζει python 2 και 3. Το pycharm IDE περιέχει εργαλεία για ανάλυση, debugger και κάποια εργαλεία για να πραγματοποιήσεις κάποια test. Επίσης, το pycharm δημιουργήθηκε ειδικά για python, όμως μπορεί να υποστηρίξει αρκετές γλώσσες, όπως HTML, CSS, Javascript κ.α. Τέλος, το pycharm είναι συμβατό με ένα μεγάλο όγκο από βάσης δεδομένων.

-CUDA[4]: CUDA είναι ένα παράλληλο computing platform και προγραμματιστικό μοντέλο, το οποίο δημιουργήθηκε απο την NVIDIA. Με την τεχνολογία αυτή κατάφεραν να εκμεταλευτούν την ισχύ από την κάρτα γραφικών και σε συνδυασμό από 150 CUDA βιβλιοθηκών έκανα τις εφαρμογές και ολη την διαδικασία εκπαίδευσης των μοντέλων πιο γρήγορη. Είναι εργαλείο που χρησιμοποιείται για την έρευνα και την δημιουργία καινούριων deep learning αλγορίθμων.

-Cudnn[5]: Cudnn είναι μία σημαντική βιβλιοθήκη που χρησιμοποιείται για την εκπαίδευση deep learnong μοντέλων που χρησιμοποιεί σωστές παραμέτρους χωρίς να χρειάζεται να ασχοληθείς με low-level GPU tuning. Είναι συμβατό με πολλά deep learning frameworks, όπως Caffe2, Chainer, Keras, MATLAB, MxNet, PaddlePaddle, PyTorch και TensorFlow [6].

1.2 Βιβλιογραφική επισκόπηση

Είναι ευρέως γνωστό πως 3 είναι οι κυρίαρχες προσεγγίσεις για τα συστήματα συστάσεων και συγκεκριμένα, είναι τα συστήματα που βασίζονται στο περιεχόμενο(content-based) [7], είναι συστήματα που βασίζονται στην μέθοδο συνεργατικού φιλτραρίσματος και οι υβριδικές μέθοδοι. Οι μέθοδοι που βασίζονται στο περιεχόμενο χρησιμοποιούν τα χαρακτηριστικά των στοιχείων για τη δημιουργία λειτουργιών, οι οποίες θα ταιριάζουν με τα προφίλ χρηστών. Αυτές οι μέθοδοι αξιοποιούν τα χαρακτηριστικά των χρηστών και τα χαρακτηριστικά των αντικειμένων για να προβλέψετε την αξιολόγηση(rating) χρησιμοποιώντας διάφορες μεθόδους μηχανικής εκμάθησης. Οι μέθοδοι που βασίζονται στο περιεχόμενο, επίσης θα μπορούσαν να χρησιμοποιούν χώρο-χρονικά χαρακτηριστικά των χρηστών και των αντικειμένων [8].

Από την άλλη πλευρά, οι μέθοδοι συνεργατικού φιλτραρίσματος [9] χρησιμοποιούν τη σχέση μεταξύ χρηστών και στοιχείων που κωδικοποιούνται στον πίνακα αξιολογήσεων να κάνει προβλέψεις. Οι μέθοδοι αυτοί αξιοποιούν ομοιότητες μεταξύ χρήστες και αντικείμενα για την πρόβλεψη των αξιολογήσεων.

Τα υβριδικά συστήματα [10] συστάσεων είναι ένας συνδυασμός των 2 προηγούμενων συστημάτων. Παρόλα αυτά, αυτό είναι η βάση των υβριδικών συστημάτων. Τα σύγχρονα μοντέλα χρησιμοποιούν πολλές τεχνικές και αλγορίθμους, ώστε να βελτιώσουν την απόδοση των συστημάτων.

Υπάρχουν αρκετοί τρόποι για το πως θα αποφευχθεί η ψυχρή εκκίνηση(cold-start) [11], το οποίο είναι ένα πρόβλημα που συμβαίνει όταν δεν υπάρχει αρκετά μεγάλη πληροφορία για τον χρήστη, ώστε να πραγματοποιηθεί μία πρόβλεψη. Μία λύση είναι οι παράπλευρες πληροφορίες(side information), οι οποίες δίνουν πληροφορία στο σύστημα για τον χρήστη και βοηθάνε το σύστημα να ”μάθει” καλύτερα κάθε χρήστη που δεν υπάρχει αρκετή πληροφορία στην βάση δεδομένων. Για την δημιουργία των μοντέλων βασιστήκαμε κυρίως στο DNNRec: A novel deep learning based hybrid recommender system [12]. Το μοντέλο που έχουν δημιουργήσει στο άρθρο [12] χρησιμοποιεί παράπλευρη πληροφορία για να αποφύγει την ψυχρή εκκίνηση. Έτσι και σε αυτη την εργασία για να μπορέσουμε να το εξελίξουμε χρησιμοποιήσαμε επιπλέον πληροφορία από τις βάσεις δεδομένων. Έτσι εχουμε ακόμη περισσότερη πληροφορία για κάθε χρήστη. Μία άλλη λύση, την οποία χρησιμοποιούν το netflix [1] και το youtube [2] είναι πως κάθε καινούριος χρήστης, όταν χρησιμοποιήσει την εφαρμογή για πρώτη φορά πραγματοποιούνται κάποιες ερωτήσεις, οι οποίες θα δημιουργήσουν μία γρήγορη εικόνα στο σύστημα για τον χρήστη.

Επιπλέον, τεχνικές για την βελτίωση της λειτουργίας τον συστημάτων, όπως να χρησιμοποιεί ενσωμάτωσης για να αναπαραστήσει χρήστες και στοιχεία για εκμάθηση μη γραμμικών λανθάνοντων παραγόντων, έχουν προταθεί από το ίδιο άρθρο [12]. Το άρθρο [13] πρότεινε ένα σύστημα σύστασης που λαμβάνει υπόψη τη σειρά των αξιολογήσεων [14]. Για το σκοπό αυτό, το Item2vec χρησιμοποιείται για να μάθει την αρχική ενσωματωμένη διανυσματική αναπαράσταση των στοιχείων και στη συνέχεια να διανυσματίσει τις πληροφορίες χαρακτηριστικών περιεχομένου τους. Στη συνέχεια, τα βαθιά δίκτυα Bi-LSTM [13] αξιοποιούνται για την εκμάθηση της αναπαράστασης προτιμήσεων χρήστη αμφίδρομα από την ακολουθία στοιχείων χρήστη. Έτσι και σε αυτη την εργασία προσπαθούμε να ενσωματώσουμε τις τεχνικές αυτές, ώστε να μπορέσουμε να έχουμε ένα καλό αποτέλεσμα.

1.3 Οργάνωση του τόμου

Στο κεφάλαιο 1 γίνεται μία εισαγωγή που εξηγεί τι έχει υλοποιηθεί στην εργασία αυτή. Στο κεφάλαιο 2 γίνεται η παρουσίαση όλων των βάσεων δεδομένων, καθώς επίσης αναφέρεται αναλυτικά η διαδικασία επεξεργασίας των δεδομένων, ώστε να είναι έτοιμα για την εκπαίδευση των μοντέλων. Στο κεφάλαιο 3 αναφέρονται και εξηγούνται αναλυτικά τα μέρη ενός υβριδικού συστήματος συστάσεων, καθώς επίσης εξηγούνται αναλυτικά κάποια στρώματα(layers) που χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων. Στο κεφάλαιο 4 γίνεται η παρουσίαση των μοντέλων. Στο κεφάλαιο5 παρουσιάζονται η τρόποι αξιολόγησης των μοντέλων. Στο κεφάλαιο 6 παρουσιάζονται τα αποτελέσματα από τις μετρήσεις και πραγματοποιείται η σύγκριση με το DNNRec: A novel deep learning based hybrid recommender system [12]. Στο τελευταίο κεφάλαιο 7 αναφέρονται τα συμπεράσματα και κάποιες μελλοντικές επεκτάσεις.

Κεφάλαιο 2

Παρουσίαση και επεξεργασία των δεδομένων

2.1 Εισαγωγή

Σε αυτή την ενότητα θα μιλήσουμε για τα δεδομένα που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας. Αναλυτικότερα, χρησιμοποιήθηκαν 4 διαφορετικές βάσεις δεδομένων(datasets), το Movielens100k, το Movielens1m, το Book Crossing και το FilmTrust. Θα αναφέρουμε τι πληροφορία περιέχει κάθε βάση δεδομένων, καθώς επίσης θα αναλύσουμε βήμα βήμα την διαδικασία επεξεργασίας των δεδομένων, ώστε να είναι έτοιμα για την εκπαίδευση των μοντέλων

2.2 Παρουσίαση των βάσεων δεδομένων

2.2.1 Movielens100k

Συγκεκριμένα το Movielens100k αποτελείται από 3 αρχεία, οι αξιολογήσεις, τις ταινίες και τις ετικέτες(tags). Αρχικά, το αρχείο με τις αξιολογήσεις περιέχει λίγο περισσότερο από 100 χιλιάδες εγγραφές από τους χρήστες και κάθε χρήστης έχει αξιολογήσει τουλάχιστον 20 ταινίες. Σε αυτό το αρχείο οι πληροφορίες που έχουμε είναι το ID του χρήστη, το ID της ταινίας, η αξιολόγηση που έκανε ο χρήστης στην ταινία και τέλος την ημερομηνία για το πότε έγινε η αξιολόγηση της ταινίας. Το αρχείο με τις ταινίες αποτελείται από περίπου 10 χιλιάδες ταινίες. Οι πληροφορίες που παίρνουμε από το αρχείο αυτό είναι το ID της ταινίας,

τον τίτλο της ταινίας και το είδος της ταινίας. Τέλος, το αρχείο με τις ετικέτες μας δίνει κάποιες πληροφορίες/σχόλια από κάποιο χρήστη προς μία συγκεκριμένη ταινία.

2.2.2 Movielens1m

Στην συνέχεια, το Movielens1m είναι μία βάση δεδομένων αρκετά παρόμοια με τη προηγούμενη. Αποτελείται από 3 αρχεία, στο πρώτο είναι οι αξιολογήσεις, το αρχείο με τις ταινίες και το αρχείο με τις πληροφορίες του χρήστη. Το πρώτο και το δεύτερο αρχείο έχει σχεδόν την ίδια δομή με τη προηγούμενη βάση δεδομένων το Movielens100k. Ωστόσο, βασική διαφορά τους είναι ότι το αρχείο με τις αξιολογήσεις αποτελείται από 1 εκατομμύριο αξιολογήσεις και ότι το αρχείο με τις ταινίες αποτελείται από περίπου 4 χιλιάδες ταινίες, λιγότερες από την προηγούμενη βάση δεδομένων αν και οι αξιολογήσεις είναι πολύ περισσότερα. Τέλος, το αρχείο users μας δίνει πληροφορίες για τους χρήστες. Συγκεκριμένα το αρχείο περιέχει το ID , το φύλο, την ηλικία, το επάγγελμα και το ταχυδρομικό κώδικα του χρήστη.

2.2.3 Book Crossing

Το Book Crossing είναι ένα διαφορετικό από τις 2 προηγούμενες βάσεις δεδομένων. Είναι μία βάση δεδομένων που περιέχει αξιολογήσεις για βιβλία. Η βάση δεδομένων αποτελείται από 3 αρχεία το αρχείο με τις αξιολογήσεις, το αρχείο με την πληροφορία των βιβλίων και το αρχείο με την πληροφορία των χρηστών. Το αρχείο με τις αξιολογήσεις περιέχει περίπου 1,1 εκατομμύρια αξιολογήσεις και αποτελείται απο το ID του χρήστη, το ISBN του βιβλίου και την αξιολόγηση που έχει πραγματοποιήσει ο χρήστης στο βιβλίο. Έπειτα, το αρχείο που περιέχει την πληροφορία των βιβλίων περιέχει περίπου 270 χιλιάδες διαφορετικά βιβλία. Οι πληροφορίες που μας δίνει είναι το ISBN, τον τίτλο του βιβλίου, τον συγγραφέα του βιβλίου, τον χρόνο που κυκλοφόρησε, τον εκδότη και κάποιες εικόνες από το βιβλίο. Τέλος, το τελευταίο αρχείο περιέχει πληροφορίες σχετικά με του χρήστες, όπως το ID τους, την περιοχή που μένουν και την ηλικία τους.

2.2.4 FilmTrust

Τέλος, χρησιμοποίησα και το FilmTrust, το οποίο είναι πιο απλό από τις προηγούμενες βάσεις δεδομένων. Αρχικά, αποτελείται από 2 αρχεία, το πρώτο αρχείο περιέχει την βαθμολογία των χρηστών σε κάποια αντικείμενα. Το αρχείο αυτό περιέχει περίπου περίπου 36

χιλιάδες αξιολογήσεις. Επίσης, υπάρχει ακόμη ένα αρχείο, το οποίο περιέχει πληροφορίες σχετικά με τον ποιον εμπιστεύεται ο κάθε χρήστης. Το αρχείο αυτό περιέχει κάτω από 2 χιλιάδες εγγραφές. Είναι μία πληροφορία, η οποία στην ουσία είναι σαν του φίλους στο facebook ή σαν τους ακόλουθους στο instagram.

2.3 Επεξεργασία δεδομένων

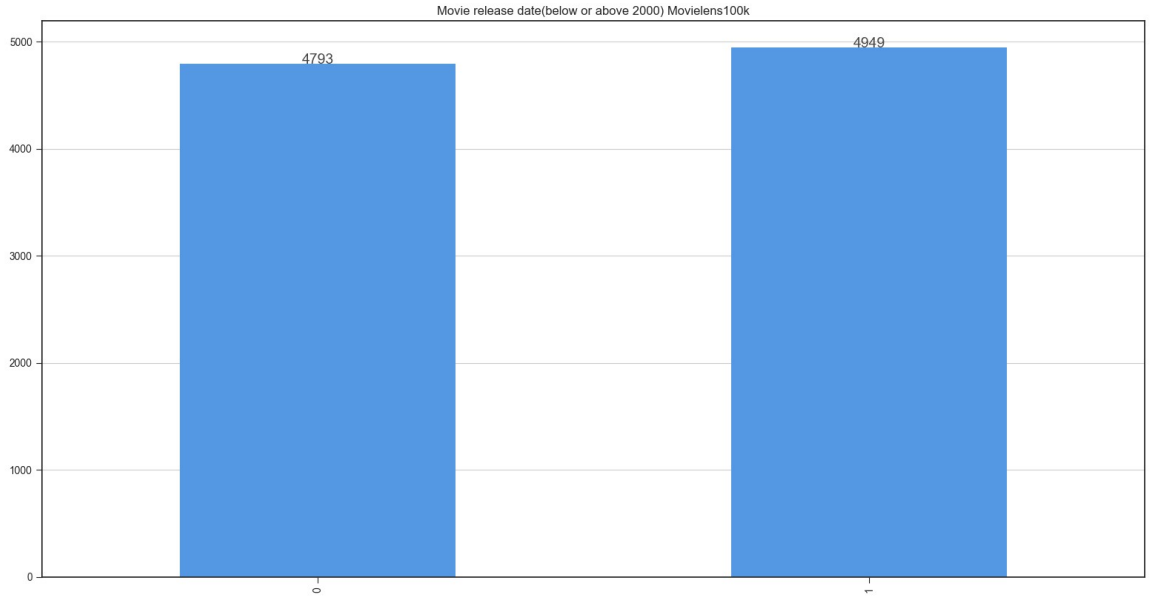
Η επεξεργασία των δεδομένων για νευρωνικά δίκτυα αποτελεί ένα απο τα κύρια κομμάτια στην υλοποίηση ενός αξιόπιστου και αποτελεσματικού μοντέλου. Το δεδομένα πρέπει να είναι καθαρά, να μην υπάρχουν δηλαδή ελλιπή στοιχεία, ώστε να μην επηρεάζει την αποτελεσματικότητα του μοντέλου. Ειδικότερα, στα συστήματα συστάσεων τα δεδομένα χωρίζονται ανάλογα με τις δυνατότητες του μοντέλου. Συγκεκριμένα αν έχουμε ένα μοντέλο που βασίζεται στο περιεχόμενο, οι πληροφορίες που μπορούμε να του δώσουμε ως είσοδο είναι ο τίτλος της ταινίας, τα είδη της ταινίας και μία περιγραφή της ταινίας. Αντίθετα αν έχουμε ένα μοντέλο συνεργατικού φιλτραρίσματα βασισμένο στους χρήστες οι πληροφορίες που πρέπει να δώσουμε στο μοντέλο πρέπει να είναι οι προτιμήσεις κάθε χρήστη. Το μοντέλο που υλοποίησα για την εργασία αυτή είναι ένα υβριδικό σύστημα συστάσεων το οποίο δέχεται και τις 2 προηγούμενες κατηγορίες δεδομένων ως είσοδο.

Αρχικά, τα υβριδικά συστήματα συστάσεων συνήθως αποτελούνται από 3 κατηγορίες δεδομένων. Η πρώτη κατηγορία αποτελείται από πληροφορίες σχετικά με την ταινία ή το βιβλίο. Όπως αναφέραμε και προηγουμένως μπορεί να περιέχει τον τίτλο, το είδος και μια περιγραφή της ταινίας ή του βιβλίου. Η επόμενη κατηγορία είναι πληροφορίες σχετικά με τις προτιμήσεις του χρήστη. Τέτοιου είδους πληροφορίες μπορεί να είναι βαθμολογίες που έχει βάλει κάποιος χρήστης για μία ταινία ή για ένα βιβλίο, καθώς επίσης θα μπορούσε να είναι και μία λίστα που θα περιέχει ταινίας ή βιβλία που θεωρεί ότι θα του άρεσε να δει ή να διαβάσει. Τέλος, η τρίτη κατηγορία αφορά τους χρήστης. Μπορεί να περιέχει πληροφορίες σχετικά με την περιοχή που μένουν, την γλώσσα που μιλάνε, το φύλο και ακόμη και δημογραφικές πληροφορίες, όπως ηλικία, εργασία κλπ.

Σε αυτή την ενότητα θα μιλήσουμε για την επεξεργασία των δεδομένων, ώστε να είναι έτοιμα για την είσοδό τους στο μοντέλο. Θα αναφέρουμε αναλυτικά της διαδικασία επεξεργασίας των δεδομένων σε 4 διαφορετικές βάσεις δεδομένων.

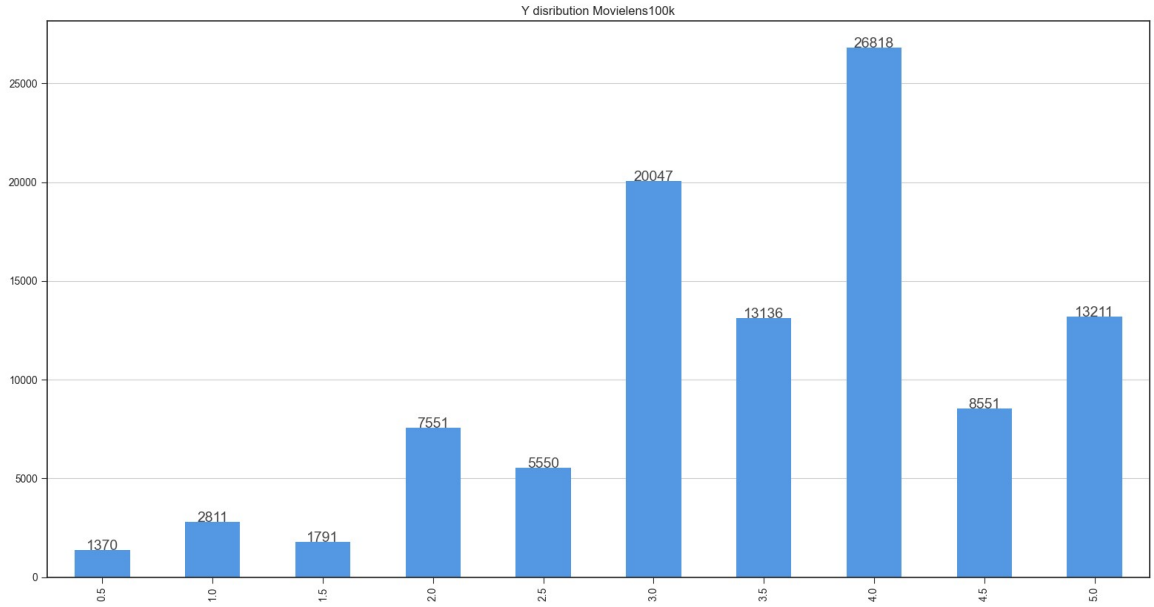
2.3.1 Προεπεξεργασία Movielens100k

Για το movielens100k πρώτα επεξεργαζόμαστε το αρχείο με τις ταινίες. Υπάρχει ένα σημαντικό πρόβλημα με τα ID τον ταινιών. Τα ID των ταινιών δεν είναι αριθμημένα με την σειρά, γι αυτό το λόγο δημιούργησα μία ακόμη στήλη με τα ID να ξεκινάνε από το 1 και να συνεχίζουν με την σειρά μέχρι και την τελευταία ταινία. Στην συνέχεια, ο τίτλος της ταινίας στις περισσότερες ταινίες περιέχει και την ημερομηνία κυκλοφορίας της. Έτσι δημιούργησα μια στήλη με το καθαρό τίτλο της ταινίας και μία στήλη με την ημερομηνία κυκλοφορίας της. Τέλος, για να δημιουργήσουμε μία νέα πληροφορία για το σύστημα θεώρησα σημαντικό να χωρίσω τις ταινίες, οι οποίες είναι παλιές και ταινίες που είναι πιο πρόσφατες. Ο χωρισμός των ταινιών έγινε πολύ απλά. Οι ταινίες που κυκλοφόρησαν πριν το 2000 τις θεώρησα ως παλιές και τις ταινίες που κυκλοφόρησαν μετά το 2000 τις θεώρησα πιο πρόσφατες. Αυτό φαίνεται στο σχήμα 2.1 (0 για ταινίες που κυκλοφόρησαν πριν το 2000 και 1 για τις υπόλοιπες ταινίες).



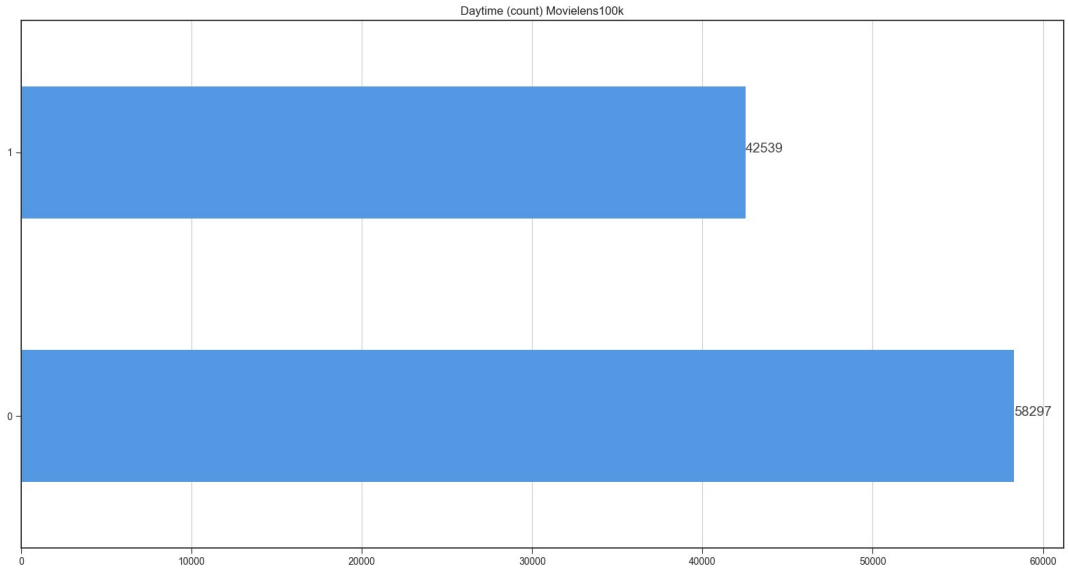
Σχήμα 2.1: Ημερομηνία κυκλοφορίας της ταινίας(πριν ή μετά το 2000) Movielens100k

Τώρα θα περάσουμε στην επεξεργασία στο αρχείο των χρηστών. Αρχικά, οι πληροφορίες που παίρνουμε από το αρχείο αυτό είναι η βαθμολογία των χρηστών στις ταινίες που έχουν παρακολουθήσει. Εκεί δεν χρειάζεται να κάνουμε κάποια αλλαγή. Το σχήμα 2.2 δείχνει ποιες είναι οι συνηθέστερες βαθμολογίες των χρηστών για τα συγκεκριμένα δεδομένα.



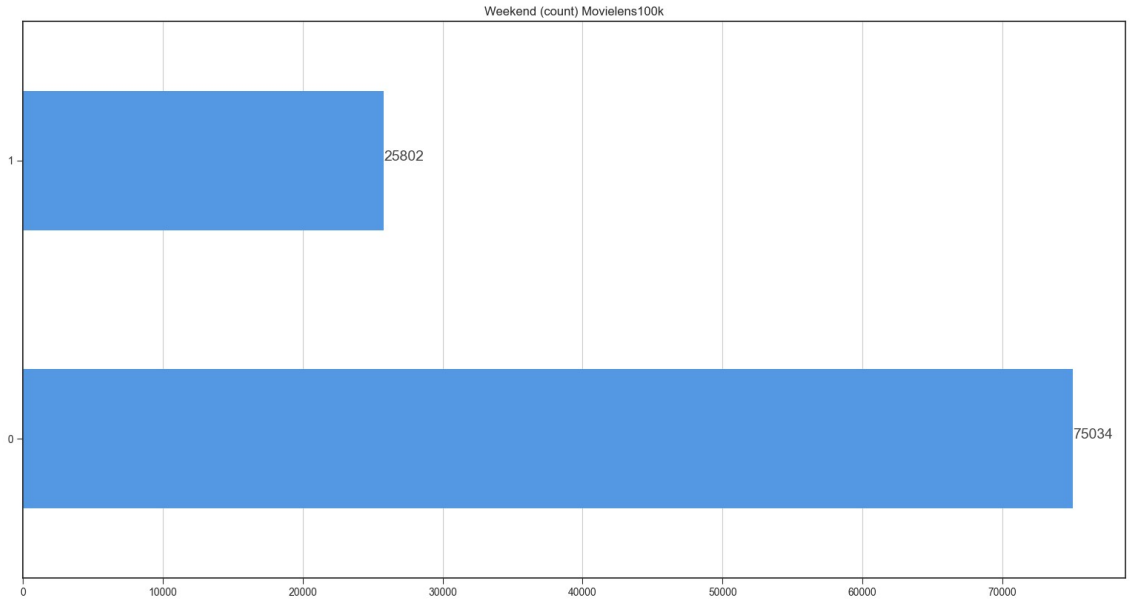
Σχήμα 2.2: Συνηθέστερες βαθμολογίες χρηστών Movielens100k

Για να μπορέσουμε να εκμεταλλευτούμε πλήρως τις πληροφορίες που μας δίνει αυτό το αρχείο μπόρεσα να βγάλω άλλες 2 χρήσιμες πληροφορίες. Η πρώτη πληροφορία είναι ποια ώρα της ημέρας έκανε ο χρήστης την βαθμολόγηση θεωρώντας πως τότε ο χρήστης παρακολούθησε την ταινία. Έτσι, δημιούργησα μία ακόμη στήλη που αν η βαθμολόγηση της ταινίας έγινε από τις πρωινές έως τις απογευματινές ώρες παίρνει την τιμή 1 και τις υπόλοιπες ώρες τις τιμές 0,όπως φαίνεται στο σχήμα 2.3.



Σχήμα 2.3: Ώρα παρακολούθησης ταινιών Movielens100k

Επιπλέον, μία ακόμη χρήσιμη πληροφορία που πήρα από την ημερομηνία βαθμολόγησης της ταινίας είναι αν οι βαθμολόγηση της ταινίας έγινε το σαββατοκύριακο. Με τον ίδιο τρόπο, όπως προηγουμένως, χώρισα ταινίες που τις έχουν παρακολουθήσει το σαββατοκύριακο και σε ταινίες που έχουν παρακολουθήσει τις υπόλοιπες μέρες, το οποίο το βλέπουμε στο σχήμα 2.4.



Σχήμα 2.4: Ημέρα παρακολούθησης ταινιών (Μέσα στην εβδομάδα ή το σαββατοκύριακο) Movielens100k

Επιπρόσθετα, αφού έχουν πραγματοποιηθεί αυτές οι διαδικασίες πρέπει να αξιοποιήσουμε την πληροφορία που μας δίνει για το είδος της ταινίας. Κάποιες ταινίες μπορεί να αντιστοιχούν σε περισσότερα απο ένα είδος, όπως για παράδειγμα Adventure|Children|Fantasy, έτσι για να μπορέσει το μοντέλο να αξιοποιήσει την πληροφορία αυτή, δημιούργησα στήλες με όλα τα δυνατά είδη ταινιών βάζοντας 1 και 0 ανάλογα με το είδος κάθε ταινίας. Έπειτα, είναι γνωστό πως για τα μοντέλα γενικότερα είναι πιο εύκολο και πιο γρήγορο να έχουν ως είσοδο μικρούς αριθμούς, καθώς οι πράξεις γίνονται πιο γρήγορα. Έτσι έκανα scale τις τιμές των βαθμολογήσεων από 0 έως 1.

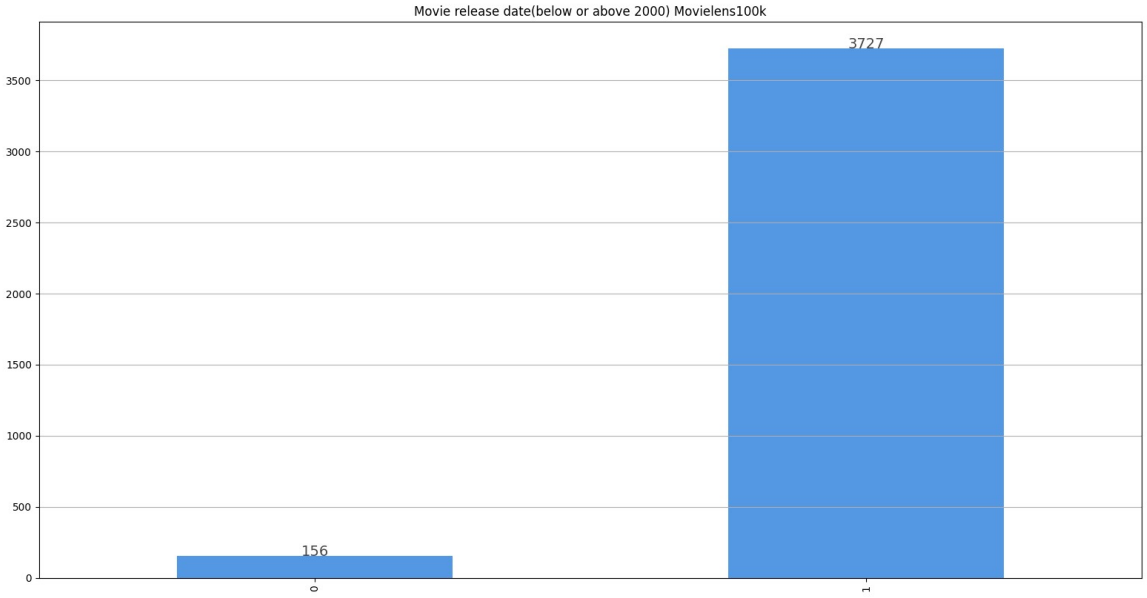
Σε αυτό το σημείο πραγματοποιείται η ένωση όλων των παραπάνω πληροφοριών και ,επίσης πραγματοποιείται και ο χωρισμός των δεδομένων που θα εκπαιδεύσουν το μοντέλο και των δεδομένων που θα γίνει η αξιολόγηση του. Όλες οι πληροφορίες που έχουμε για την είσοδο του μοντέλου είναι αριθμοί από 0 έως 1 εκτός από το ID των χρηστών και το ID των ταινιών και το τίτλο της ταινίας. Για να εκμεταλλευτούμε και τον τίτλο της ταινίας

χρησιμοποίησα το feature extraction από την sklearn, το οποίο στην ουσία παίρνει συγκεκριμένα τις 600 πρώτες πιο συνηθισμένες λέξεις από τους τίτλους των ταινιών και δημιουργεί ένα καινούριο είδος. Τέλος, αφού έχουμε ενώσει όλες τις πληροφορίες που μας έχει δώσει η βάση δεδομένων είναι λογικό να υπάρχουν κενές τιμές σε κάποια πεδία. Αυτό μπορεί να επηρεάσει αρκετά την αξιοπιστία του μοντέλου μας, οπότε αφαιρούμε κάθε γραμμή που περιέχει κενές τιμές σε κάποια στήλη. Η βάση δεδομένων είναι έτοιμη για την εκπαίδευση του μοντέλου. Οι πληροφορίες θα χωριστούν σε 4 μέρη, καθώς το μοντέλο για το συγκεκριμένη βάση δεδομένων έχει 4 εισόδους.

2.3.2 Προεπεξεργασία Movielens1m

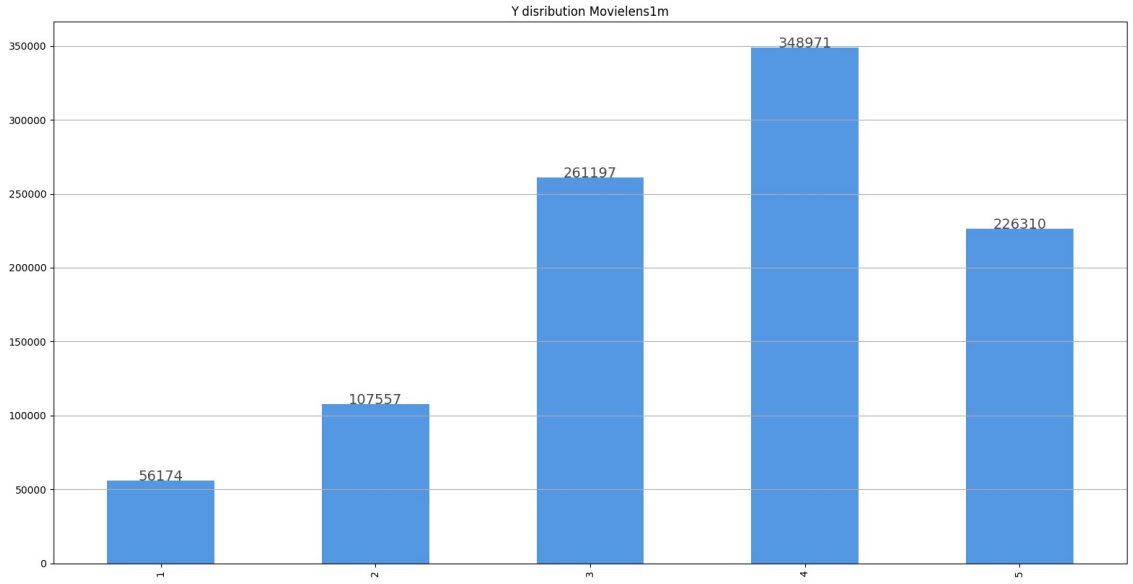
Το movielens1m είναι αρκετά παρόμοιο με το movielens100k και οι αλλαγές που χρειάστηκαν να γίνουν είναι σχεδόν ίδιες, ώστε να είναι έτοιμο για το μοντέλο. Οι διαφορές είναι ότι έχουμε περισσότερες αξιολογήσεις ταινιών και κάποιες επιπλέον πληροφορίες για τους χρήστες. Παρόλα αυτά, δεν αξιοποίησα τις συγκεκριμένες πληροφορίες καθώς υπήρχαν πολλές ελλιπής πληροφορίες και αυτό τελικά θα δημιουργούσε ένα πολύ μικρή βάση δεδομένων, η οποία δεν θα ήταν καθόλου αξιοποιήσιμη σε ένα τόσο μεγάλο και πολύπλοκο μοντέλο. Για τα προβλήματα που δημιούργησαν οι ελλιπής πληροφορίες θα αναφερθούμε στο τέλος αυτής της ενότητα. Τα στάδια της διαδικασίας επεξεργασίας δεδομένων είναι ακριβώς τα ίδια με το movilens100k, δηλαδή στο τέλος έχουμε ένα μεγάλο αρχείο που χωρίζεται και αυτό σε 4 μέρη για να είναι έτοιμο για την είσοδό του στο μοντέλο.

Στην συνέχεια θα δούμε τα αντίστοιχα γραφήματα όπως και στο Movielens100k. Το πρώτο σχήμα 2.5 δείχνει το πότε κυκλοφόρησαν οι ταινίες (0 για ταινίες που κυκλοφόρησαν πριν το 2000 και 1 για τις υπόλοιπες ταινίες).



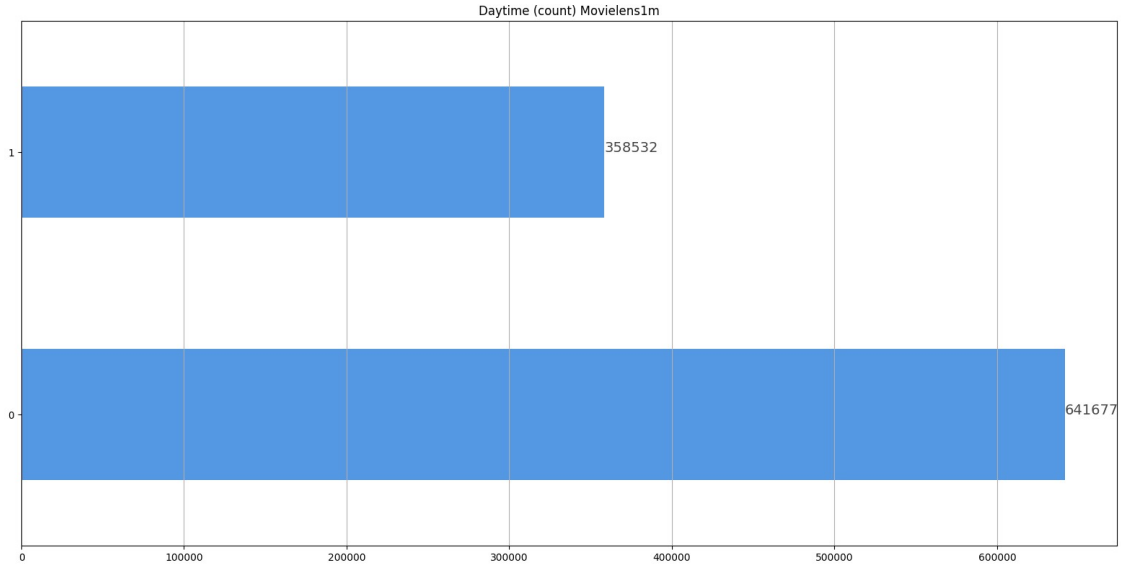
Σχήμα 2.5: Ημερομηνία κυκλοφορίας της ταινίας Movielens1m

Το δεύτερο σχήμα 2.6 δείχνει ποιες είναι οι βαθμολογίες των όλων των χρηστών.



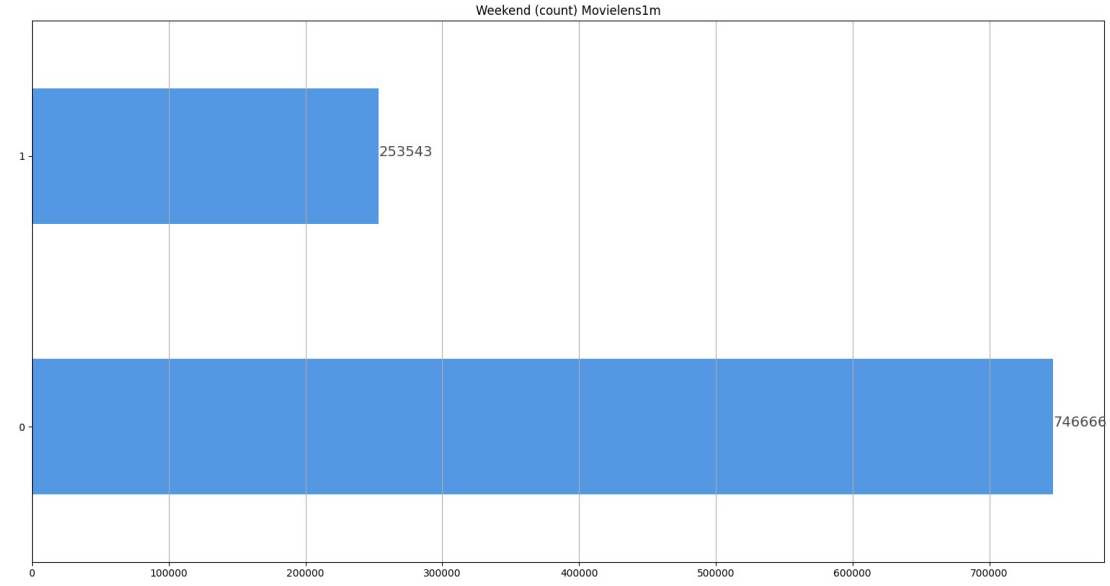
Σχήμα 2.6: Συνηθέστερες βαθμολογίες χρηστών Movielens1m

Έπειτα, ακολουθεί το σχήμα 2.7 το οποίο μας δείχνει πότε έγινε η αξιολόγηση της ταινίας.



Σχήμα 2.7: Ώρα παρακολούθησης ταινιών Movielens1m

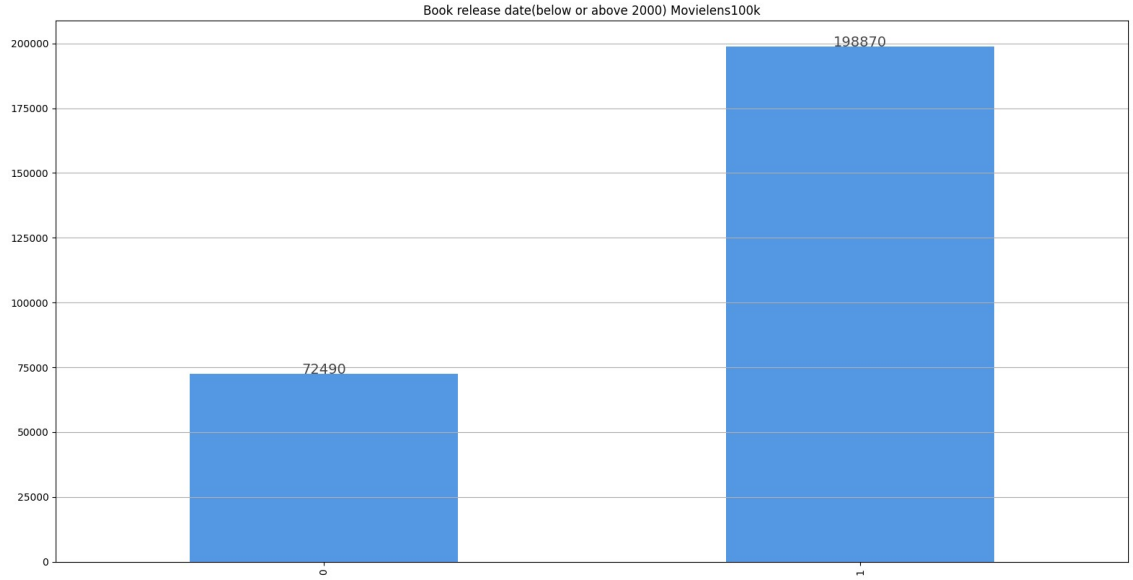
Τέλος, το σχήμα 2.8 δείχνει ποια μέρα έγινε η αξιολόγηση.



Σχήμα 2.8: Ημέρα παρακολούθησης ταινιών Movielens1m

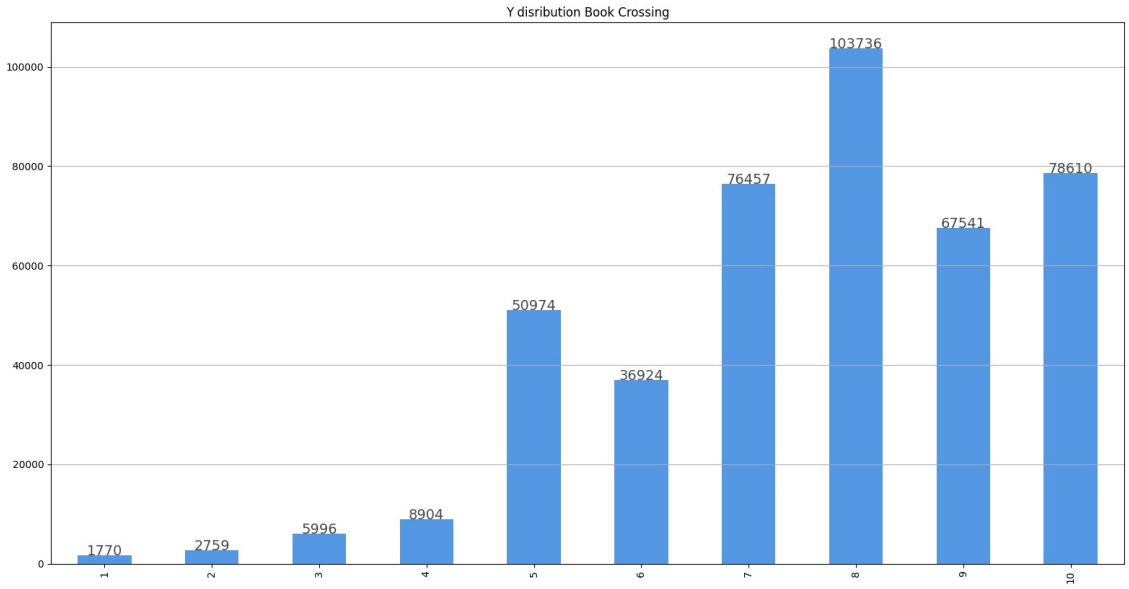
2.3.3 Προεπεξεργασία Book-Crossing

Για τη συγκεκριμένη βάση δεδομένων η διαδικασία ήταν λίγο πιο απλή αφού τα περισσότερα δεδομένα ήταν έτοιμα από την ίδια την βάση. Αρχικά, επεξεργαζόμαστε το αρχείο που περιέχει τα βιβλία. Το πρώτο πράγμα που κάνουμε είναι να διώξουμε γραμμές που δεν έχουν ISBN, δηλαδή το ID του βιβλίου. Στην συνέχεια, από την ημερομηνία κυκλοφορίας του βιβλίου δημιούργησα μία στήλη που μας δίνει την πληροφορία αν το βιβλίο είναι παλιό ή πιο πρόσφατο. Η διαδικασία γίνεται ακριβώς με τον ίδιο τρόπο όπως στο movielens100k και στο movielens1m. Τα βιβλία που κυκλοφόρησαν πριν το 2000 θεωρούνται παλιά, ενώ τα βιβλία που κυκλοφόρησαν μετά το 2000 θεωρούνται πιο πρόσφατα. Όπως και στις προηγούμενες βάσεις δεδομένων το σχήμα 2.9 μας δείχνει πόσα βιβλία είναι πιο παλιά και πόσα βιβλία είναι πιο πρόσφατα.



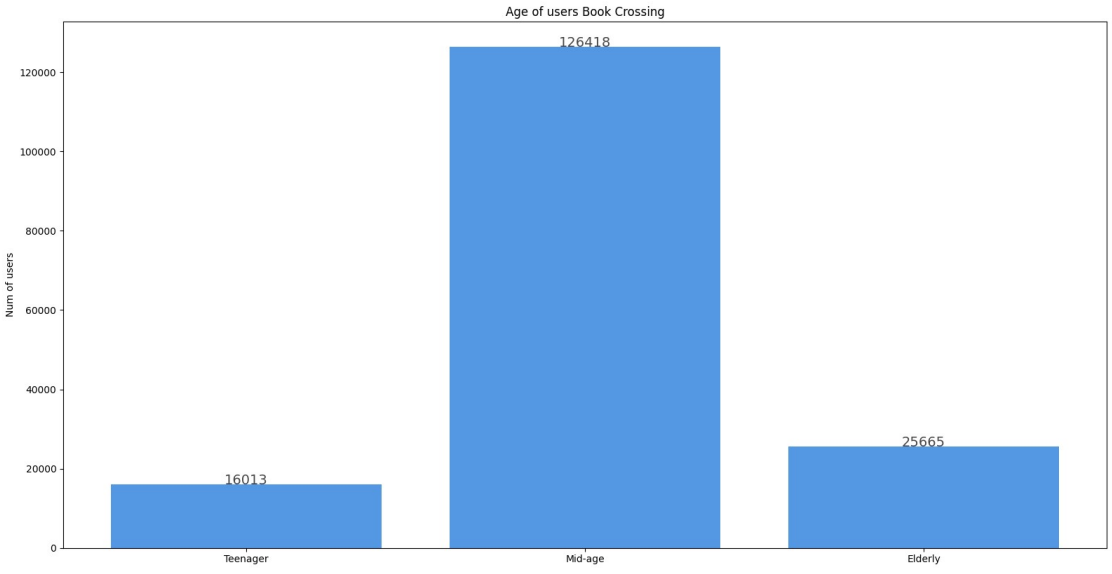
Σχήμα 2.9: Ημερομηνία κυκλοφορίας του βιβλίου(πριν ή μετά το 2000) Book-Crossing

Η επεξεργασία του αρχείου με τους χρήστες είναι ακόμη πιο απλή. Το μόνο που χρειάστηκε να κάνουμε ήταν να ελέγξουμε αν δεν υπάρχει κάποιο ID από κάποιο χρήστη και να μετονομάσουμε κάποιες στήλες. Στο σχήμα 2.10 θα δείτε τις βαθμολογίες των χρηστών στα βιβλία. Είναι διαφορετικές από τις βαθμολογίες των ταινιών, καθώς στις ταινίες έχουμε βαθμολογία από 0 έως 5 και στα βιβλία έχουμε από 0 έως 10.



Σχήμα 2.10: Συνηθέστερες βαθμολογίες χρηστών Book-Crossing

Για να μπορέσουμε να αξιοποιήσουμε τις περισσότερες πληροφορίες που μας δίνει η βάση δεδομένων, ως επιπλέον πληροφορία χρησιμοποίησα την ηλικία κάθε χρήστη. Έτσι, υπάρχουν 3 στήλες, στις οποίες μπαίνουν οι τιμές 0 ή 1. Η πρώτη από τις 3 στήλες ονομάζεται teenager και εκεί μπαίνουν οι χρήστες που είναι κάτω των 18 σε ηλικία. Η δεύτερη στήλη αφορά τους χρήστες που η ηλικία είναι ανάμεσα απο 18 και 50. Τέλος, η τρίτη στήλη αφορά όλους τους υπόλοιπους που είναι πάνω από 50 χρονών. Στο σχήμα 2.11 αυτό έχουμε την ηλικία των χρηστών με τον χωρισμό που κάναμε παραπάνω.

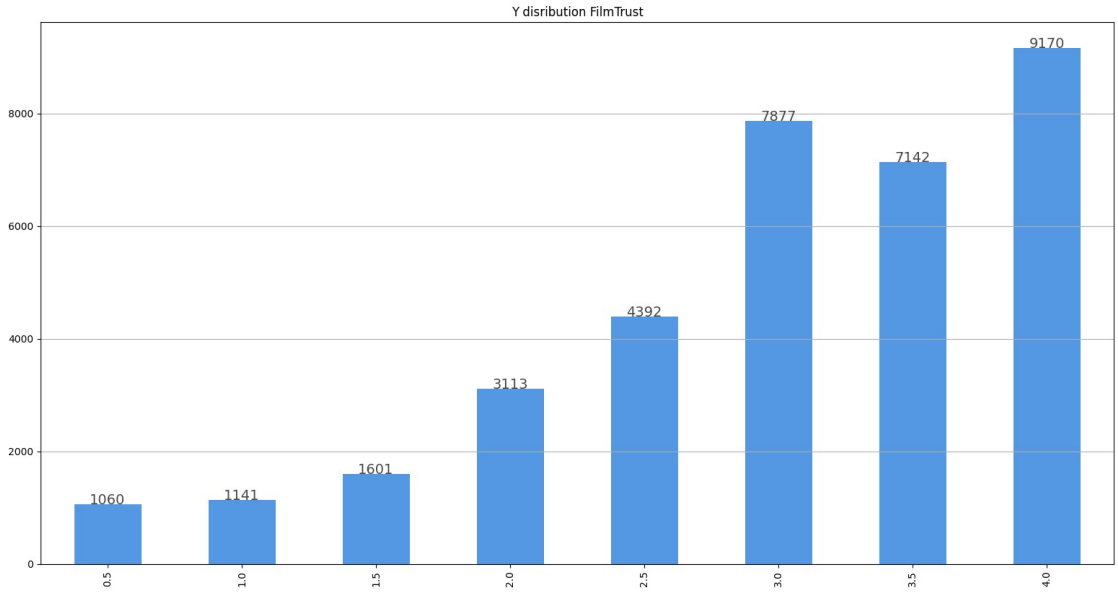


Σχήμα 2.11: Ηλικίες χρηστών που έχουν βαθμολογήσει κάποιο βιβλίο Book-Crossing

Αυτή είναι, κυρίως η διαδικασία για τη συγκεκριμένη βάση. Ακολουθεί η σύνδεση όλων των πληροφοριών, η διαγραφή κάθε γραμμής που δεν περιέχει κάποια τιμή και η διαγραφή όλων των πανομοιότυπων σειρών σε περίπτωση που υπάρχουν. Έτσι καταλήγουμε σε ένα καθαρό αποτέλεσμα που περιέχει σημαντικές πληροφορίες.

2.3.4 Προεπεξεργασία FilmTrust

Η βάση δεδομένων αυτή είχε την πιο απλή διαδικασία από τα προηγούμενα. Ειδικότερα την μόνη πληροφορία που χρησιμοποίησα από την βάση ήταν μόνο οι βαθμολογίες που έδωσαν οι χρήστες σε κάποιο αντικείμενο. Έτσι, αρχικά, το πρώτο πράγμα που πραγματοποίησα είναι ο καθαρισμός της βάσης. Στην συνέχεια, δημιούργησα ένα pivot table και ως γραμμές έχει τους χρήστες και ως στήλες τα αντικείμενα. Η πληροφορία που περιέχει το table αυτό είναι η βαθμολογία που έχει βάλει ο χρήστης σε κάποια από τα αντικείμενα. Αυτή η διαδικασία έγινε γιατί είναι πιο εύκολο να γίνει scale στις βαθμολογίες των αντικειμένων. Τέλος, το σχήμα 2.12 αυτό δείχνει τις βαθμολογίες των χρηστών, όπως γινόταν σε όλα τις προηγούμενες βάσεις δεδομένων.



Σχήμα 2.12: Συνηθέστερες βαθμολογίες χρηστών FilmTrust

Τέλος, δημιουργούμε ένα πλαίσιο δεδομένων που έχει τους χρήστες, το αντικείμενο και την βαθμολογία του αντικειμένου. Έπειτα, γίνεται ο διαχωρισμός σε Train και Test και διαγράφουμε κάθε γραμμή που περιέχει ελλειπή δεδομένα.

2.4 Γενικές πληροφορίες για τα δεδομένα

Η επεξεργασία των δεδομένων είναι μια δύσκολη διαδικασία, ειδικά όταν το δεδομένα έχουν μεγάλο όγκο. Γι αυτό τον λόγω έχω αποθηκεύσει την τελική τους μορφή σε αρχεία, έτσι ώστε να μπορώ να τα φορτώνω κατευθείαν στα μοντέλα. Η διαδικασία γίνεται αυτόματα μέσα από τα προγράμματα. Σε κάθε βάση δεδομένων ο διαχωρισμός ήταν 80% και 20% για το Train και το Test αντίστοιχα.

Ένα μεγάλο πρόβλημα με τις βάσεις δεδομένων είναι οι ελλιπής τιμές [11]. Είναι τιμές που δεν μπορούσα να χρησιμοποιήσω για την εκπαίδευση του νευρωνικού, καθώς θα άλλαζε την συμπεριφορά του παίρνοντας λάθος δεδομένα. Επίσης, δεν μπόρεσα να αξιοποιήσω πληροφορίες, όπως επάγγελμα και την τοποθεσία του χρήστη, γιατί εκεί είναι που υπάρχουν τα περισσότερα ελλιπή στοιχεία(missing values) και αυτό οδηγεί το τελικό αρχείο που θα δημιουργηθεί και θα είναι έτοιμο για την είσοδο του στο μοντέλο να είναι πολύ μικρό σε σχέση με το αρχικό.

Κεφάλαιο 3

Τα μέρη του υβριδικού συστήματος συστάσεων

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναλύσουμε τα μέρη ενός υβριδικού συστήματος συστάσεων, καθώς επίσης θα παρουσιαστούν μερικά από τα στρώματα(layers), τα οποία χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων. Θα αναλύσουμε όλα τα μέρη από τα οποία αποτελείται ένα υβριδικό σύστημα συστάσεων σε θεωρητικό επίπεδο. Πιο συγκεκριμένα θα αναλύσουμε τις πληροφορίες που μπορεί να δεχτεί κάθε μέρος του υβριδικού συστήματος συστάσεων, ώστε να καταλάβουμε την χρησιμότητα κάθε μέρους του υβριδικού συστήματος συστάσεων.

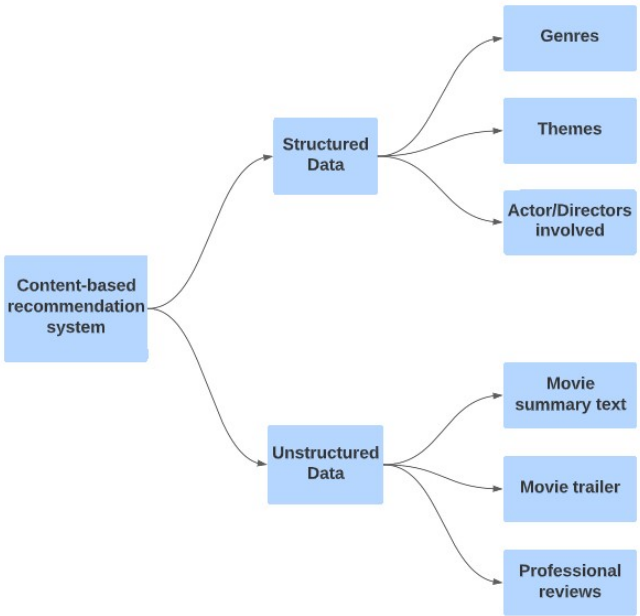
Στην συνέχεια, θα αναλύσουμε κάποια συγκεκριμένα στρώματα(layers) που χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων, όπως στρώσεις ενσωμάτωσης (embedding layers) και LSTM. Οι στρώσεις ενσωμάτωσης αποτελούν καθοριστικό ρόλο στην υλοποίηση των μοντέλων. Επίσης, θα αναφερθούμε στο πως μπορούμε να αξιοποιήσουμε και να παράγουμε αποτέλεσμα από παραπάνω από ένα στρώμα ενσωμάτωσης. Έπειτα, θα αναφερθούμε στα LSTM στρώματα για το πως λειτουργούν, καθώς επίσης και πως επικοινωνούν μεταξύ τους μέσα σε ένα μοντέλο. Τέλος, θα αναφερθούμε για τα αμφίδρομα LSTM στρώματα, τα οποία χρησιμοποιήθηκαν για την υλοποίηση των μοντέλων.

Για την σωστή λειτουργία των μοντέλων δεν θα μπορούσαμε να παραλείψουμε τα dropout στρώματα και το l2 kernel regularization, που είναι υπεύθυνα για την σωστή λειτουργία των μοντέλων καθώς βοηθάνε στην ομαλή εκπαίδευση των μοντέλων αποφεύγοντας την υπερεκπαίδευση του μοντέλου(over-fitting. σ

3.2 Τα μέρη ενός υβριδικού συστήματος συστάσεων

3.2.1 Σύστημα που βασίζεται στο περιεχόμενο

Η τεχνική που βασίζεται στο περιεχόμενο [7], [8] είναι μία από τις πιο συνηθισμένες τεχνικές για τα συστήματα συστάσεων. Τα αντικείμενα τα οποία αρέσουν σε έναν χρήστη αναφέρεται ως ”content”. Τα συστήματα αυτά χρησιμοποιούν πληροφορίες, όπως πληροφορίες σχετικά με χαρακτηριστικά από αντικείμενα που αρέσουν στον χρήστη με σκοπό να μπορέσουν να προτείνουν παρόμοια αντικείμενα στον ίδιο. Ο στόχος των συγκεκριμένων μοντέλων είναι να ταξινομήσει τα προϊόντα με κάποιες συγκεκριμένες λέξεις κλειδιά, να μάθει τα χαρακτηριστικά από αντικείμενα που αρέσει σε κάποιον χρήστη και στην συνέχεια να προτείνει παρόμοια αντικείμενα από την βάση δεδομένων. Η πληροφορίες που αναλύει το συγκεκριμένο είδος μοντέλου μπορεί να είναι δομημένες και μη δομημένες. Συγκεκριμένα κάποιες από τις δομημένες πληροφορίες είναι το είδος και το γενικότερο θέμα της ταινίας ή του βιβλίου. Επιπλέον, οι πληροφορίες σχετικά με προτιμήσεις σε ηθοποιούς, σκηνοθέτες ή συγγραφείς αντίστοιχα. Παρόλα αυτά, υπάρχουν και οι μη δομημένες πληροφορίες, όπως ένα κείμενο περιγραφής της ταινίας ή του βιβλίου, το trailer της ταινίας ή ακόμη και ένα κείμενο με την κριτική από κάποιον επαγγελματία σε μία ταινία ή ένα βιβλίο (Σχήμα 3.1).



Σχήμα 3.1: Οι πληροφορίες που δέχεται ένα μοντέλο βασισμένο στο περιεχόμενο

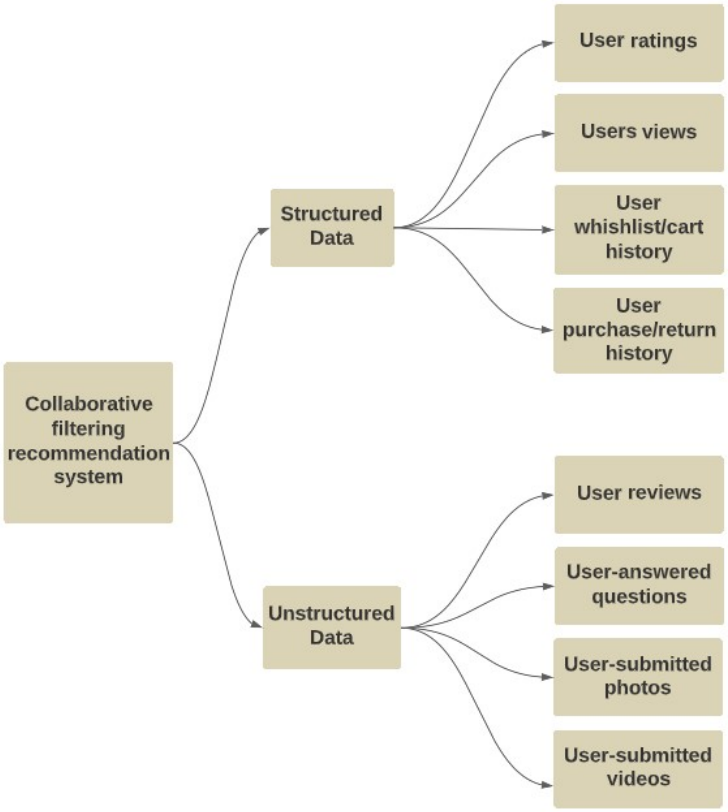
3.2.2 Σύστημα συνεργατικού φιλτραρίσματος

Το συνεργατικό φιλτράρισμα είναι και αυτός ένας αλγόριθμος πρόβλεψης συστάσεων, ο οποίος χρησιμοποιεί αλγορίθμους που φιλτράρουν δεδομένα από τους χρήστες ώστε να κάνει προβλέψεις ξεχωριστά για κάθε χρήστη σε αντικείμενα που τον ενδιαφέρουν. Επίσης, πολλές εταιρίες χρησιμοποιούν το συνεργατικό φιλτράρισμα για να προβάλουν σωστές διαφημίσεις για τους χρήστες στα μέσα κοινωνικής δικτύωσης. Υπάρχουν 2 διαφορετικοί τύποι που χρησιμοποιούνται, κυρίως στα συστήματα συστάσεων, είναι: αυτό που βασίζεται στο γείτονα (neighbor-based) και αυτό που βασίζεται στο αντικείμενο(item-to-item). [9], [7]

Αναλυτικότερα, για το σύστημα που βασίζεται στον γείτονα, συγκρίνονται οι χρήστες με άλλους ενεργούς χρήστες. Ένας χρήστης είναι παρόμοιος με κάποιον άλλο όταν για κάποια αντικείμενα έχουν όμοιες κριτικές. Συγκρίνοντας τα σημεία στα οποία μοιάζουν αυτοί οι χρήστες και βασιζόμενοι σε αυτή την λογική είναι πολύ πιθανό οι κριτικές σε μελλοντικά αντικείμενα να είναι παρόμοιες. Όλα αυτά πραγματοποιούνται για χρήστες που είναι συνήθως καινούριοι. Για τους χρήστες που είναι ενεργοί αρκετό καιρό αναλύεται ο μέσος όρος των κριτικών των χρηστών και από εκεί πραγματοποιείται μία πρόβλεψη και γι'αυτούς τους χρήστες.

Για το σύστημα που βασίζεται στο αντικείμενο, συγκεκριμένα η διαδικασία δημιουργεί έναν πίνακα με τις προτιμήσεις των χρηστών σε αντικείμενα που έχει βαθμολογήσει. Στην συνέχεια, η διαδικασία κάνει μία πρόβλεψη σε αντικείμενα που δεν έχει βαθμολογήσει ο χρήστης, η οποία βασίζεται στις πραγματικές προτιμήσεις του χρήστη. Η διαδικασία αυτή γίνεται για όλα τα αντικείμενα που βρίσκονται σε αυτόν τον πίνακα.

Τέλος, κάποια συστήματα συνεργατικού φιλτραρίσματος χωρίζονται σε συστήματα που βασίζονται στην μνήμη(memory-based) και κάποια άλλα σε συστήματα που βασίζονται στο μοντέλο(model-based). Το πρώτο απλά συγκρίνει ομοιότητες μεταξύ χρηστών και αντικειμένων, ενώ το δεύτερο χρησιμοποιεί μηχανική μάθηση για να συγκρίνει ανόμοια αντικείμενα (Σχήμα 3.2).



Σχήμα 3.2: Οι πληροφορίες που δέχεται ένα μοντέλο συνεργατικού φιλτραρίσματος

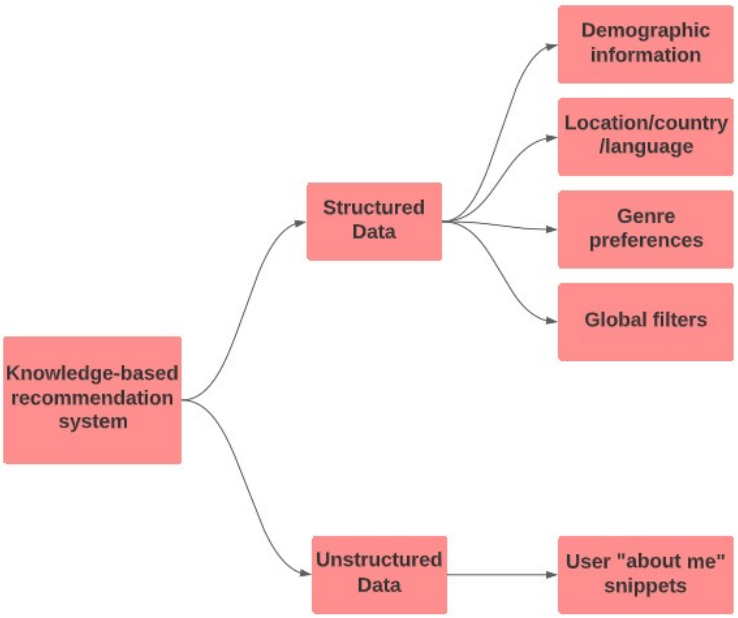
3.2.3 Σύστημα βασισμένο στην γνώση(Knowledge-based system)

Άλλο ένα σύστημα συστάσεων είναι το σύστημα που βασίζεται στην γνώση [15], [16], [8], το οποίο είναι πολύ σημαντικό σε κάποιες συγκεκριμένες περιπτώσεις αφού ο χρήστης δεν χρειάζεται να έχει δώσει σχεδόν καμία πληροφορία στο σύστημα προηγουμένως. Ένα σύστημα συστάσεων βασίζεται στην γνώση, όταν οι προτάσεις που πραγματοποιεί το σύστημα δεν βασίζονται στις προτιμήσεις του χρήστη ή στις βαθμολογίες που έχει δώσει ο χρήστης στα αντικείμενα, αλλά σε συγκεκριμένες προτάσεις που δίνει ο χρήστης κατά την είσοδό του στο σύστημα. Οι πληροφορίες που δίνει είναι μία σειρά από χαρακτηριστικά του αντικειμένου που ενδιαφέρεται, ώστε το σύστημα να σχηματίσει μία γρήγορη εικόνα για τον χρήστη. Εκτός από διάφορα φίλτρα στην ουσία που μπορεί να δώσει ο χρήστης στο σύστημα, επίσης μπορεί να δώσει δημογραφικές πληροφορίες. Επιπλέον, οι πληροφορίες μπορεί να δώσε μπορούν να αφορούν σχετικά με την τοποθεσία στην οποία κατοικεί ο χρήστης, την χώρα ακόμη και την γλώσσα που μιλάει.

Ένα απλό παράδειγμα είναι αν κάποιος ψάχνει για κάποιο σπίτι, ο χρήστης πρέπει να δώσει μία εικόνα για το πως θέλει να μοιάζει το σπίτι που θέλει να αγοράσει, δηλαδή ποιο είναι το εύρος της τιμής, πόσα υπνοδωμάτια θα έχει το σπίτι και πόσο μεγάλο θέλει να είναι. Εκτός από αυτά τα χαρακτηριστικά ο χρήστης μπορεί να δώσει πιο γενικές πληροφορίες όπως αν θέλει να είναι πιο σύγχρονο ή αν θέλει να είναι ευρύχωρο.

Αυτό μπορεί να μας οδηγήσει στο ερώτημα γιατί αυτή η διαδικασία θεωρείται ένα σύστημα συστάσεων αφού μπορεί απλά να γίνει σε μία ιστοσελίδα με συγκεκριμένα φίλτρα.

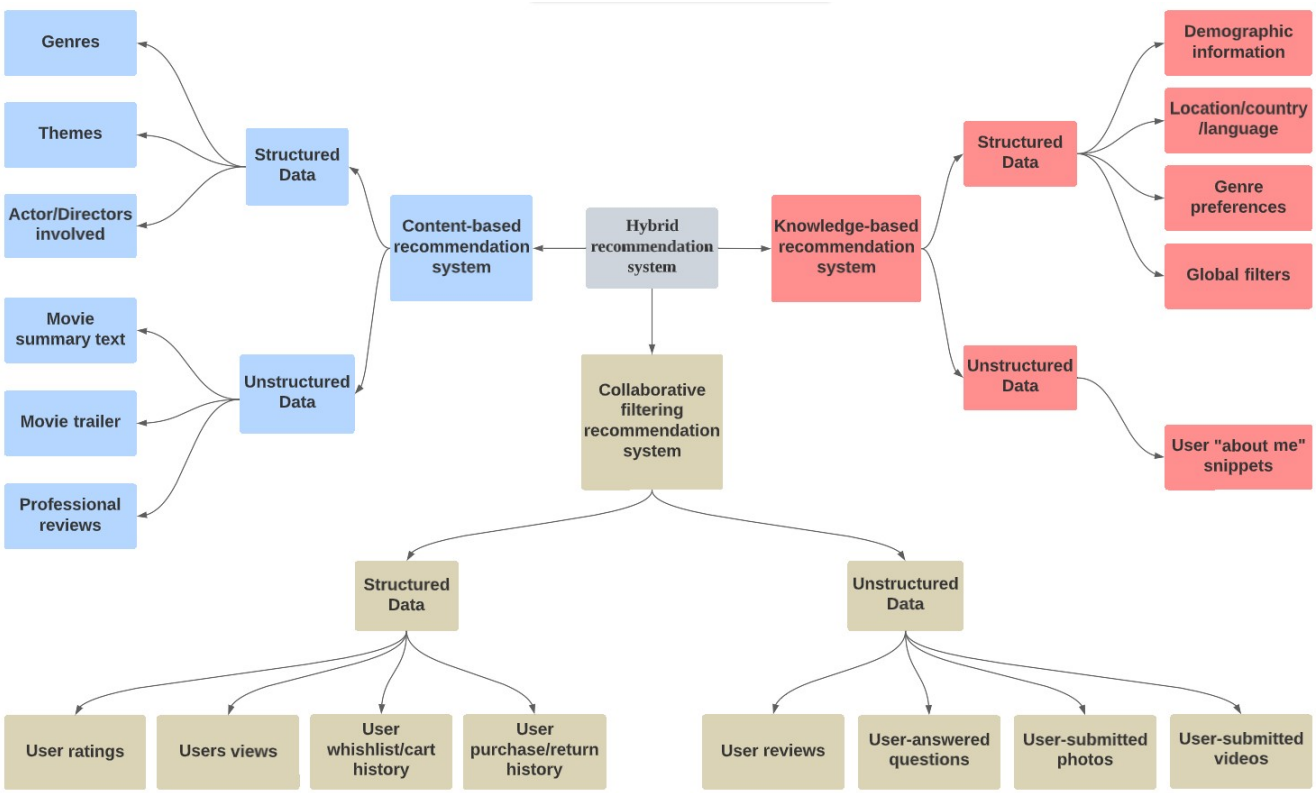
Η απάντηση σε αυτό το ερώτημα είναι ότι ναι μπορεί κάποιος να το κάνει απλά χρησιμοποιώντας φίλτρα απλά σε ένα σύστημα συστάσεων η είσοδος που δίνει ο χρήστης είναι μοναδική και μπορεί να βάλει ο ίδιος ότι θέλει. Αυτό μπορεί να οδηγήσει στην δημιουργία περίπλοκων φίλτρων με αποτέλεσμα απλές εφαρμογές που χρησιμοποιούν φίλτρα σε μία βάση δεδομένων να μην έχουν κάποια επιλογή να επιστρέψουν. Τέλος, ένα σύστημα συστάσεων που βασίζεται στην γνώση έχει την δυνατότητα να διαβάζει ξεχωριστά και πιο προσωπικά κάθε χρήστη και τα αποτελέσματα να είναι μοναδικά για κάθε χρήστη. Κατά την διάρκεια του χρόνου μπορείς να ζητάς, δηλαδή κάποια ανατροφοδότηση από τον χρήστη και το σύστημα μόνου του να δίνει βαρύτητα σε συγκεκριμένες πληροφορίες που έχει δώσει ο χρήστης (Σχήμα 3.3).



Σχήμα 3.3: Οι πληροφορίες που δέχεται ένα μοντέλο που βασίζεται στην γνώση

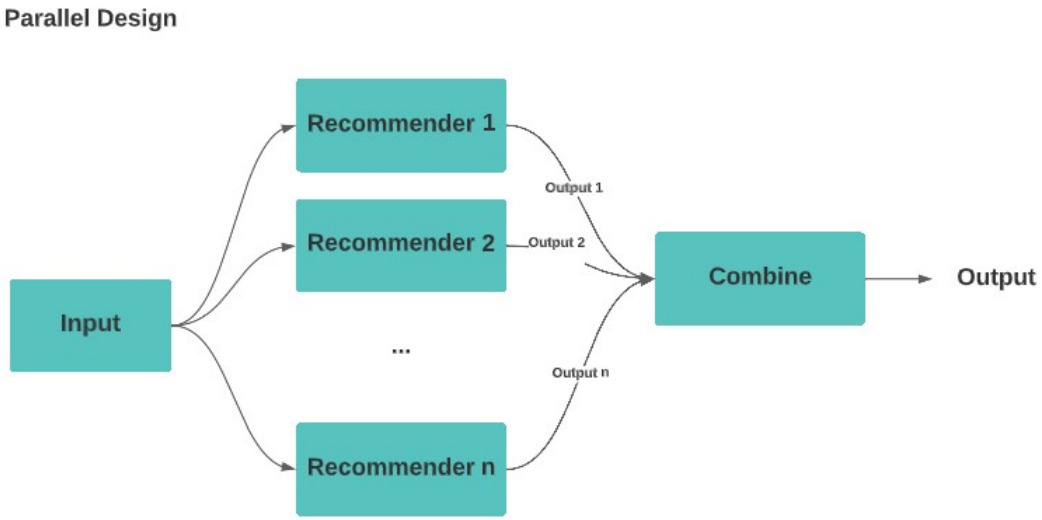
3.2.4 Υβριδικό σύστημα συστάσεων

Το υβριδικό σύστημα συστάσεων [7], [10] είναι το σύστημα που επέλεξα να δημιουργήσω και για τις 4 βάσεις δεδομένων σε αυτή την εργασία. Οι αναφορές που έγιναν για τα άλλα συστήματα πραγματοποιήθηκαν γιατί το υβριδικό σύστημα συστάσεων είναι ένας συνδυασμός των 3 αυτών κυρίως συστημάτων. Με τον συνδυασμό των παραπάνω μοντέλων δημιουργούμε ένα πιο ισχυρό και πιο γενικό σύστημα, το οποίο μπορεί να δέχεται και να συνδυάζει μεγάλο αριθμό διαφορετικών πληροφοριών (Σχήμα 3.4).



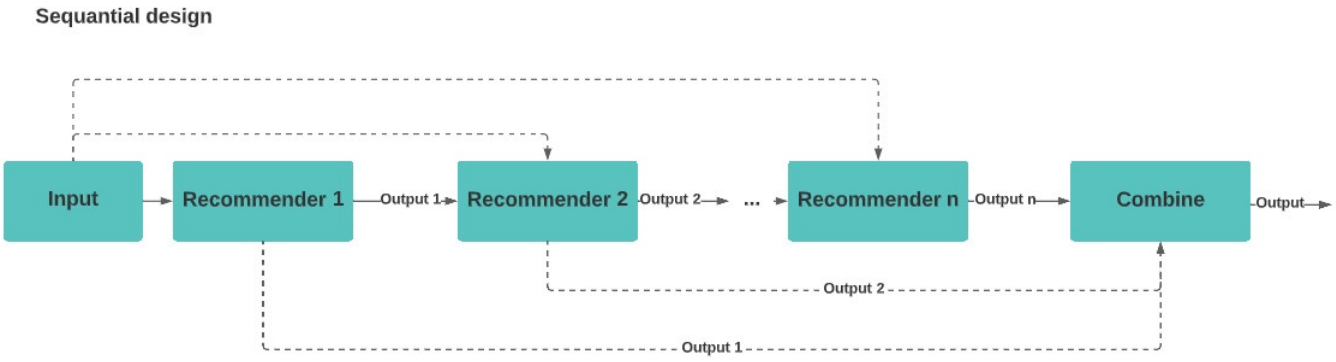
Σχήμα 3.4: Οι πληροφορίες που δέχεται ένα υβριδικό σύστημα συστάσεων μοντέλο

Τα υβριδικά συστήματα συστάσεων έχουν 2 επικρατέστερα συστήματα, παράλληλα και σε σειρά. Το παράλληλο, το οποίο πραγματοποίησα και εγώ, η πληροφορία εισόδου δίνεται σε πολλά συστήματα συστάσεων και το καθένα δημιουργεί ένα αποτέλεσμα. Τα αποτελέσματα των συστημάτων συστάσεων συνδυάζονται και δημιουργούν μία έξοδο (Σχήμα 3.5).



Σχήμα 3.5: Υβριδικό σύστημα συστάσεων παράλληλα

Από την άλλη, τα συστήματα συστάσεων που συνδέονται σε σειρά η πληροφορία εισόδου δίνεται σε ένα σύστημα συστάσεων και η έξοδος του περνάει ως είσοδος στο επόμενο σύστημα συστάσεων με την σειρά (Σχήμα 3.6).

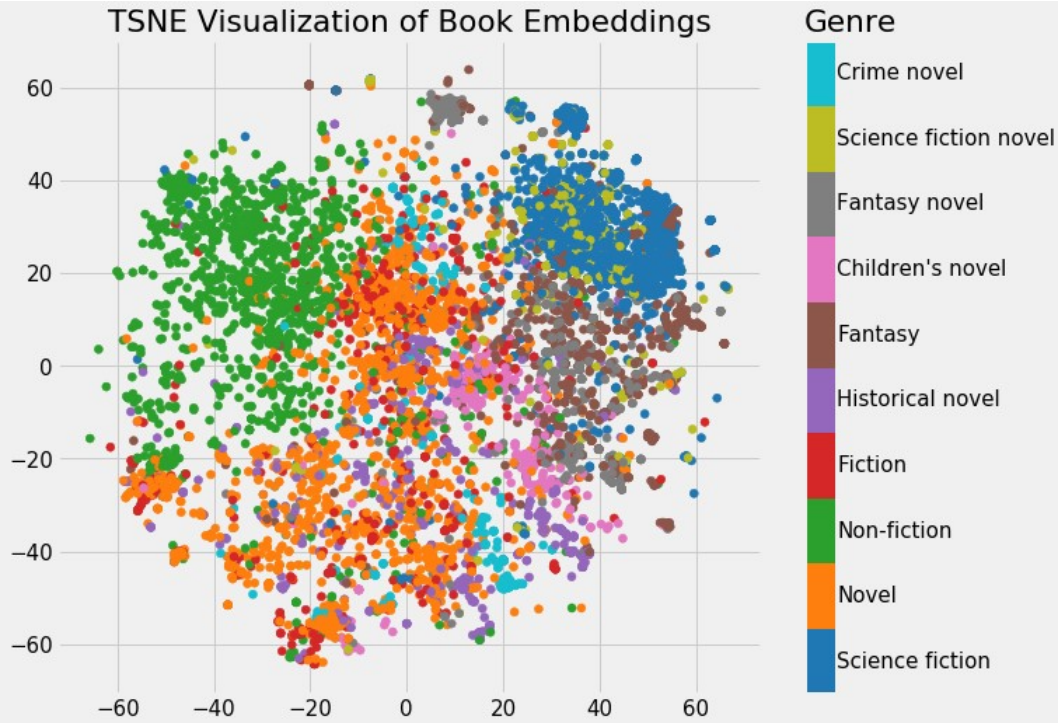


Σχήμα 3.6: Υβριδικό σύστημα συστάσεων σε σειρά

3.2.5 Στρώσεις ενσωμάτωσης

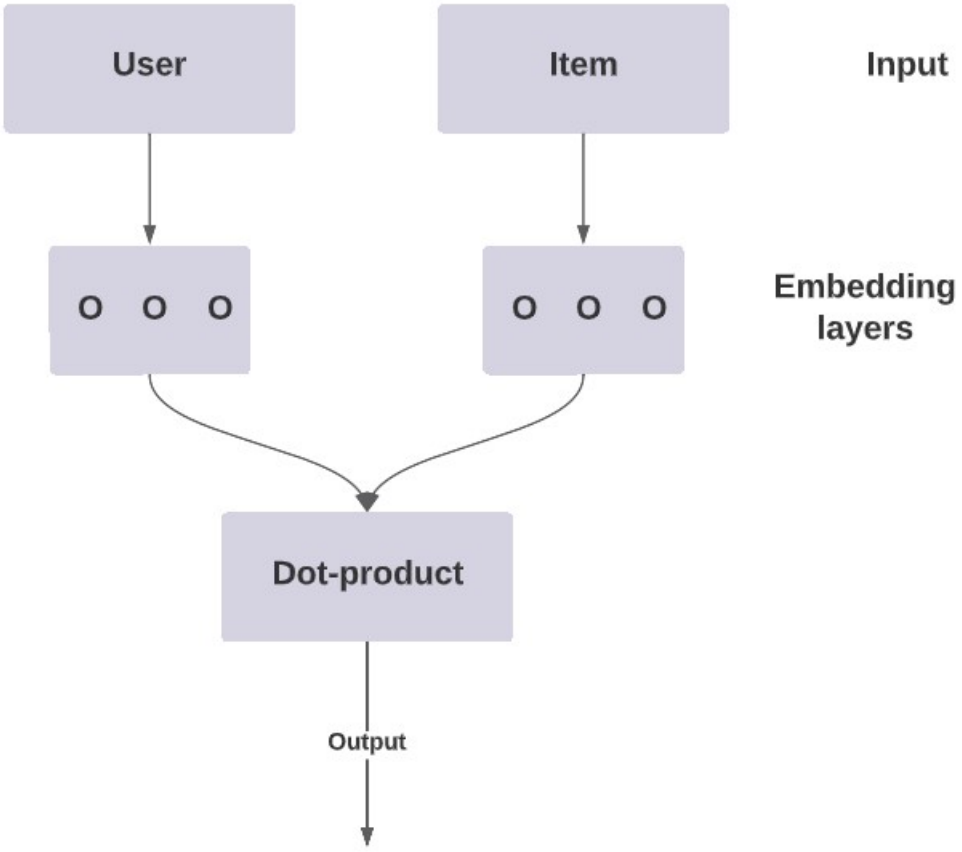
Τα στρώματα ενσωμάτωσης(embedding layers) είναι πολύ σημαντικά στρώματα, τα οποία χρησιμοποιούνται κυρίως για την επεξεργασία φυσικής γλώσσας(Natural Language Processing). Το τελευταίο είναι ένα μέρος από την επιστήμη των υπολογιστών και πιο συγκεκριμένα με την τεχνιτή νοημοσύνη. Ειδικότερα, είναι μία διαδικασία που δίνει την ικανότητα στους υπολογιστές να καταλάβουν κείμενο, ακόμη και τις λέξεις όταν μιλάει κάποιος σε κάποιο βίντεο με τον ίδιο τρόπο, όπως οι άνθρωποι μπορούν.

Τα στρώματα ενσωμάτωσεις αναπαριστούν τις διακριτές και κατηγοριοποιημένες μεταβλητές. Σε αντίθεση με μία μέθοδο κωδικοποίησης όπως η κωδικοποίηση one-hot, τα στρώματα ενσωμάτωσης έχουν την δυνατότητα να μαθαίνουν από μόνα τους κατά την διάρκεια την εκπαίδευσης. Αυτό σημαίνει πως μπορούν να καταλάβουν πιο εύκολα τις ομοιότητες 2 αντικειμένων και να τις τοποθετήσει το ένα πιο κοντά στο άλλο στο χώρο ενσωμάτωσης(embedding space) (Σχήμα 3.7).



Σχήμα 3.7: Embedding space map

Τα στρώματα ενσωμάτωσης μπορούν να χρησιμοποιηθούν και στα συστήματα συστάσεων. Τα περισσότερα συστήματα συστάσεων βασίζονται στους χρήστες και στα αντικείμενα (ταινίες, βιβλία, παιχνίδια κ.α.). Έτσι, οι λέξεις για τα στρώματα ενσωμάτωσης είναι οι χρήστες και τα αντικείμενα, δηλαδή δημιουργούνται 2 στρώματα ενσωμάτωσης (Σχήμα 3.8).



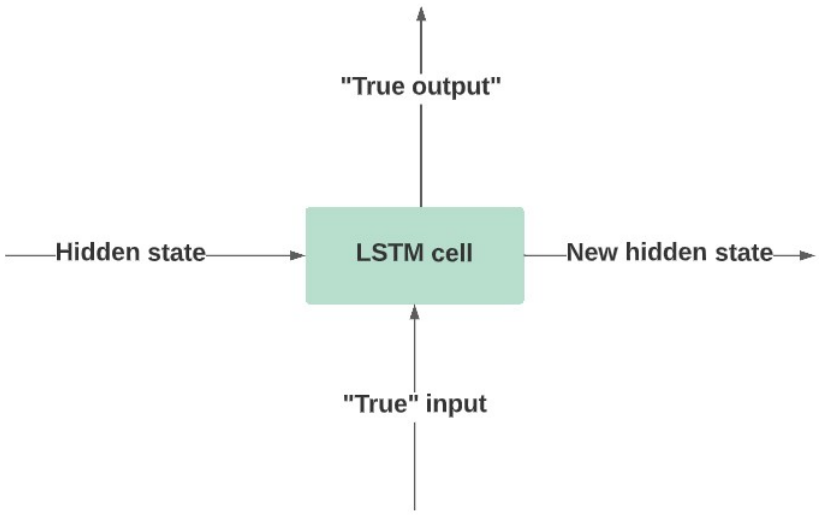
Σχήμα 3.8: Dot-product από 2 embedding layers

Η παραπάνω διαδικασία πραγματοποιείται σε όλα τα μοντέλα στην εργασία αυτή. Για να πάρουμε το τελικό αποτέλεσμα υπολογίζουμε το dot-product που δίνει την τελική τιμή ανάμεσα στα στρώματα ενσωμάτωσης του χρήστη και στα στρώματα ενσωμάτωσης των αντικειμένων. Το τελικό αυτό αποτέλεσμα δείχνει την αλληλεπίδραση του χρήστη με το αντικείμενο αυτό.

3.2.6 Μακροπρόθεσμη προσωρινή μνήμη

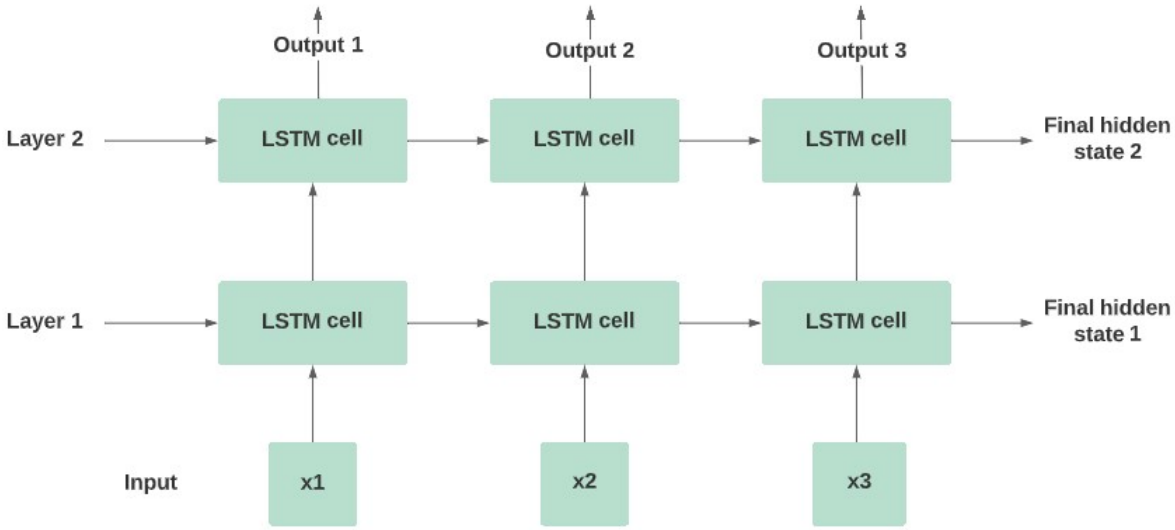
Στα περισσότερα νευρωνικά δίκτυα όλοι οι είσοδοι και όλοι οι έξοδοι είναι ανεξάρτητοι μεταξύ τους. Σε συγκεκριμένες περιπτώσεις, όπως αν θέλεις να προβλέψεις την επόμενη λέξη από μία πρόταση, είναι αναγκαίο το μοντέλο να θυμάται τις προηγούμενες λέξεις, ώστε η πρόβλεψη να είναι σχετική με τις προηγούμενες λέξεις. Επίσης, ένα ακόμη παράδειγμα είναι πως αν εσύ έχεις παρακολουθήσει μία ταινία μέχρι ένα σημείο και την σταματάς και πρέπει να προβλέψεις το τέλος της, εξαρτάται σε εσένα κατά πόσο έχεις παρακολουθήσει την ταινία και κατά πόσο έχεις παρακολουθήσει τα κύρια σημεία της ταινίας. Ακριβώς έτσι δουλεύει το RNN. Το RNN θυμάται όλες τις πληροφορίες που έχουν περάσει από το μοντέλο. Ωστόσο, υπάρχουν τα LSTM(long short-term memory) που στην ουσία είναι μία εξελιγμένη μορφή των RNN και έχουν ένα ιδιαίτερο χαρακτηριστικό να θυμούνται εκτός απ'όλες τις πληροφορίες που έχουν περάσει απο το μοντέλο αλλά και μία σειρά απο τις προηγούμενες εισόδους για να γίνουν ακόμη πιο αποδοτικά.

Για να καταλάβουμε την λειτουργία των bidirectional LSTM πρέπει πρώτα να καταλάβουμε πως λειτουργούν τα απλά LSTM. Ένα LSTM δίκτυο αποτελείται από LSTM κελιά(cells) (Σχήμα 3.9). Τα LSTM κελιά πραγματοποιούν αρκετές διαδικασίες. Παρόλα αυτά, για να μπορέσουμε να καταλάβουμε την λειτουργία του, θα τα θεωρήσουμε ως μεμονωμένες και πλήρως υπολογιστικές μονάδες. Ας πούμε ότι ένα LSTM κελί παίρνει 2 εισόδους. Η μία είσοδος είναι το "true input" και η άλλη είσοδος είναι το "hidden input" (προηγούμενο "hidden state"). Έτσι, παράγονται 2 έξοδοι, το "true output" και το καινούριο "hidden state".



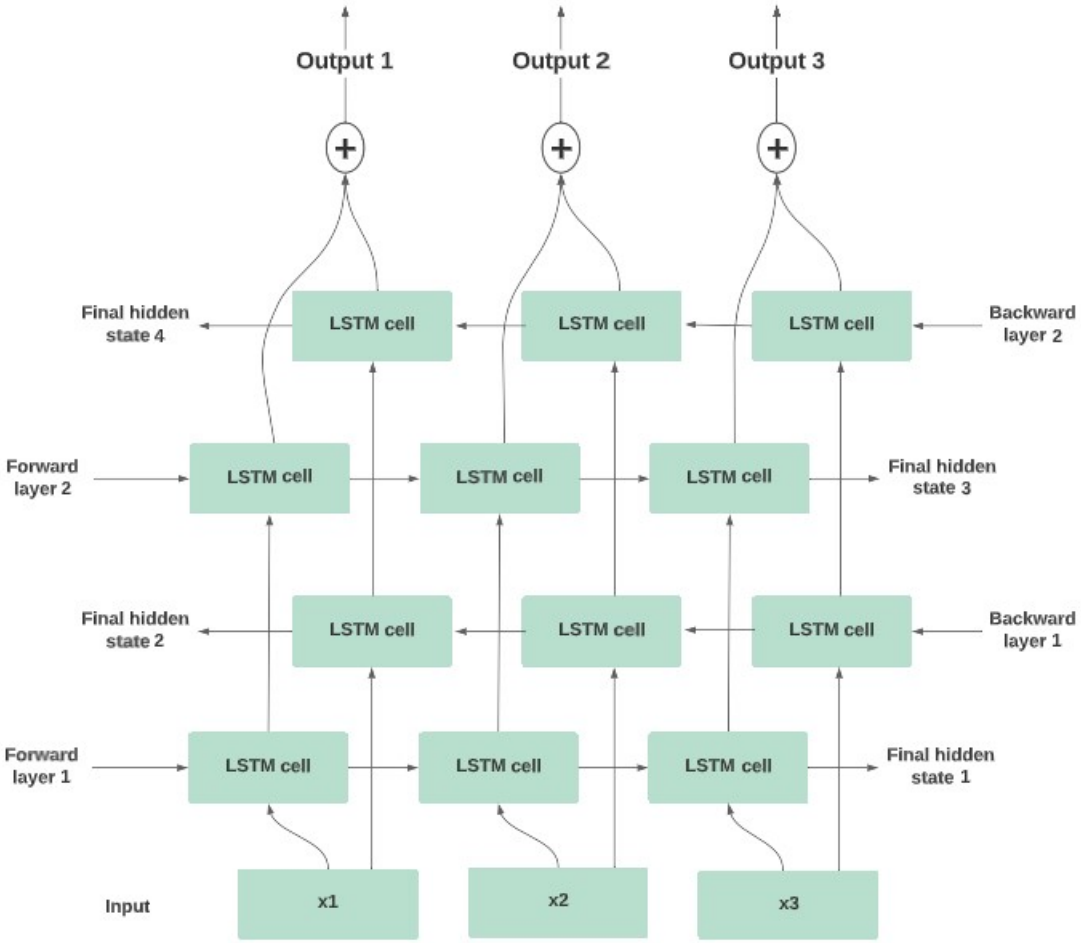
Σχήμα 3.9: LSTM cell

Οι διαφορές ανάμεσα στην ”True” και στην ”hidden” είσοδο είναι ότι η έξοδος hidden πηγαίνει στην κατεύθυνση της σείρας των LSTM κελιών, ενώ η έξοδος ”True” συνεχίζει πιο βαθιά στο δίκτυο. Έτσι, αν υπάρχουν 2 και παραπάνω LSTM στρώματα η έξοδος ”True” είναι η είσοδος στο επόμενο LSTM στρώμα (Σχήμα 3.10).



Σχήμα 3.10: LSTM cells communication

Για τα αμφίδρομα LSTM στρώματα (bidirectional LSTM layers) είναι ακριβώς όπως τα κανονικά LSTM στρώματα με την διαφορά ότι η επικοινωνία των LSTM κελιών να είναι αμφίδρομη. Υπάρχει επικοινωνία και από τα αριστερά προς τα δεξιά (κανονικά LSTM) και υπάρχει επικοινωνία και απο δεξιά προς τα αριστερά. Ωστόσο, η ”True” έξοδος πάνε προς την ίδια κατεύθυνση. Οι 2 κατευθύνσεις αυτές λειτουργούν τελείως ανεξάρτητα, μέχρι το τελευταίο στρώμα, όπου οι τελικοί έξοδοι ενώνονται (Σχήμα 3.11).



Σχήμα 3.11: Bidirectional LSTM cells communication

Κεφάλαιο 4

Υλοποίηση Μοντέλων

4.1 Εισαγωγή

Η υλοποίηση των μοντέλων αποτελεί ένα από τα κύρια μέρη για την εκπόνηση της διπλωματικής εργασίας. Έπρεπε να υλοποιηθούν 4 διαφορετικά μοντέλα για κάθε βάση δεδομένων, αφού κάθε βάση είναι μοναδική και περιέχει τα δικά του χαρακτηριστικά. Έτσι, έπρεπε να υλοποιηθεί διαφορετικό μοντέλο για κάθε βάση, η οποία θα πληρούσε όλα τα χαρακτηριστικά για να μπορέσει να είναι υβριδικό μοντέλο.

Στο προηγούμενο κεφάλαιο 3 αναφέραμε όλες σχεδόν τις μορφές των συστημάτων συστάσεων. Έτσι, στο κεφάλαιο αυτό θα πραγματοποιηθεί ανάλυση για κάθε μέρος όλων των μοντέλων αναλυτικά. Έτσι, θα κατανοήσετε πως λειτουργεί ένα υβριδικό σύστημα συστάσεων, καθώς επίσης στο τέλος θα δείτε την τελική μορφή κάθε μοντέλου.

4.2 Υλοποίηση μοντέλων

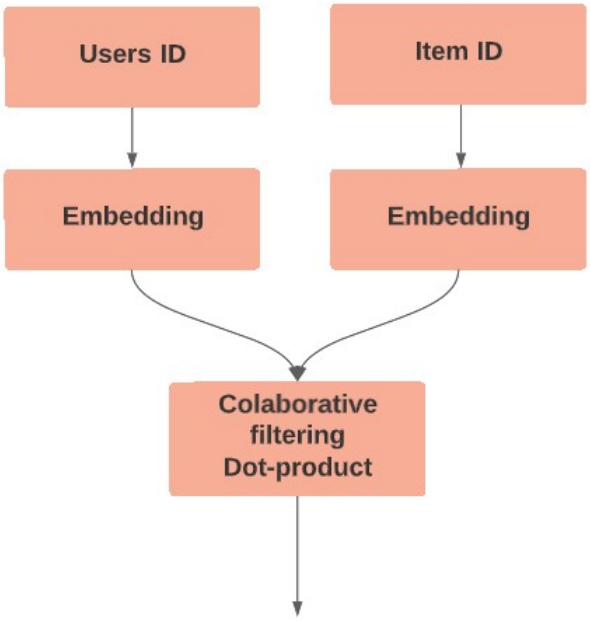
4.2.1 Μοντέλο για το Movielens100k

Για το μοντέλο αυτό έχουμε 4 εισόδους. Η πρώτη και η δεύτερη είσοδος παίρνουν το ID του χρήστη και το ID της ταινίας αντίστοιχα. Η τρίτη είσοδος παίρνει ως είσοδο τα είδη της ταινίας και η τελευταία είσοδος παίρνει τα υπόλοιπα δεδομένα που ετοιμάσαμε, δηλαδή πότε έγινε η αξιολόγηση ποια μέρα και ποια ώρα. Αυτές είναι οι είσοδοι στο μοντέλο και μπορείτε να τις δείτε στο σχήμα 4.1.

Στην συνέχεια, θα μιλήσουμε για τις 2 πρώτες εισόδους του μοντέλου. Η πληροφορία

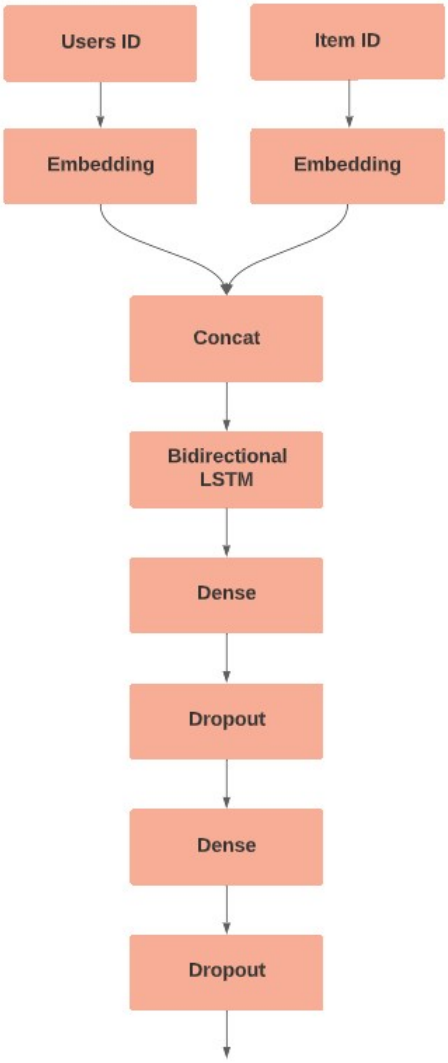
Σχήμα 4.1: Οι είσοδοι στο μοντέλο Movielens100k

αμέσως μετα περνάει από τις στρώσεις ενσωμάτωσης και για τις 2 είσοδοι. Έτσι, δημιουργείται ένας ενσωματωμένος χώρος για τους χρήστες και ένας για τις ταινίες. Το πρώτο μέρος του μοντέλου είναι μία υλοποίηση συνεργατικού συστήματος, όπου οι 2 στρώσεις ενσωμάτωσης αποτελούν είσοδο σε ένα Dot στρώμα. Το πρώτο μέρος του μοντέλου φαίνεται στο σχήμα 4.2.



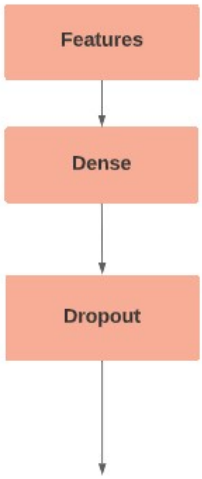
Σχήμα 4.2: Dot-product για το μοντέλο Movielens100k

Έπειτα, το επόμενο μέρος του μοντέλου είναι ένα σύστημα συνεργατικού φιλτραρίσματος που στην συνέχεια περνάει από ένα νευρωνικό δίκτυο. Σε αυτή την περίπτωση οι έξοδοι από τις στρώσεις ενσωμάτωσης δεν περνάν από το dot στρώμα, απλά ενώνονται και αποτελούν την είσοδο στο νευρωνικό δίκτυο. Το νευρωνικό δίκτυο αποτελείται από ένα αμφίδρομο LSTM [17], [13] στρώμα με συνάρτηση ενεργοποίησης την relu. Στην συνέχεια, ακολουθεί ένα dense στρώμα, το οποίο έχει και αυτό συνάρτηση ενεργοποίησης την relu και,επίσης, έχει και kernel regularizer τον l2 [18]. Μετά το dense στρώμα έχει σειρά ένα dropout στρώμα [19] με 0.2% ποσοστό. Τέλος υπάρχουν ξανά σε σειρά τα 2 τελευταία στρώματα. Το μέρος του μοντέλου το βλέπουμε στο σχήμα 4.3.



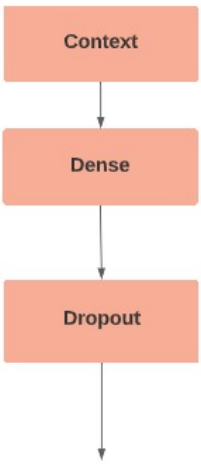
Σχήμα 4.3: Collaborative filtering και νευρωνικό δίκτυο του Movielens100k μοντέλου

Τα 2 τελευταία μέρη του μοντέλου είναι αρκετά απλά σε σχέση με τα προηγούμενα. Το πρώτο παίρνει ως είσοδο τα είδη της ταινίας. Για την συγκεκριμένη βάση δεδομένων τα είδη είναι 20 για κάθε ταινία. Το μέρος αυτό του μοντέλου αποτελείται από ένα dense στρώμα, το οποίο έχει και αυτό l2 kernel regularizer, όπως φαίνεται στο σχήμα 4.3.



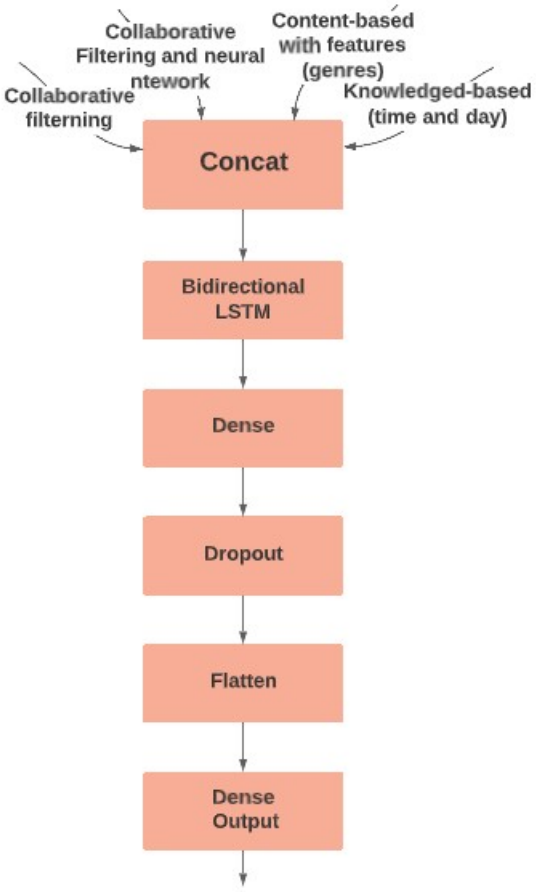
Σχήμα 4.4: Το content-based μέρος του Movielens100k μοντέλου

Τέλος, ακριβώς η ίδια διαδικασία πραγματοποιείται στο μέρος του μοντέλου που βασίζεται στην γνώση. Αποτελείται από ένα dense και από ένα dropout στρώμα με ακριβώς τα ίδια χαρακτηριστικά με το προηγούμενο μέρος του μοντέλου, οπως φαίνεται στο σχήμα 4.5.



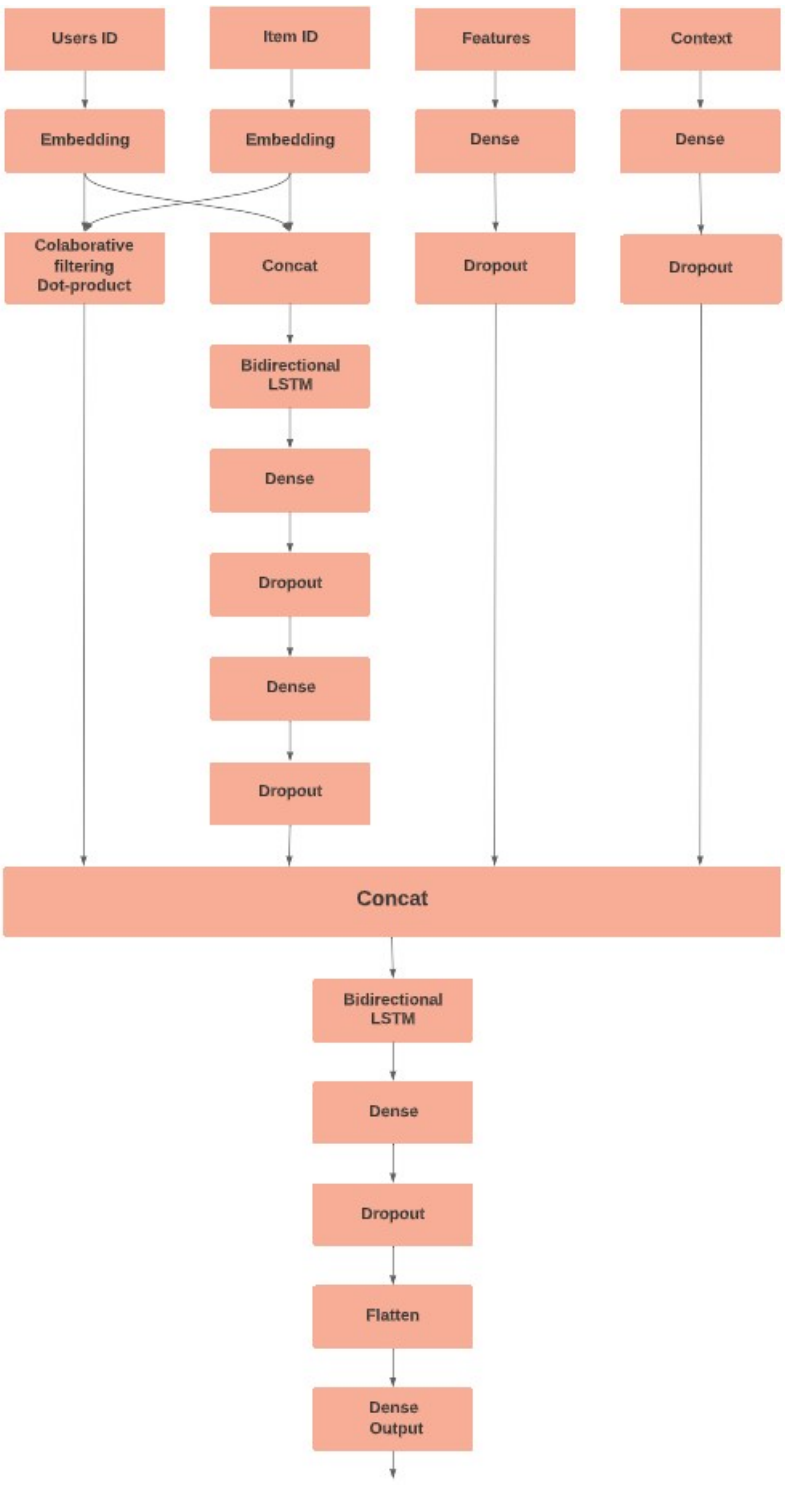
Σχήμα 4.5: Το knowledged-based μέρος του Movielens100k μοντέλου

Αυτά είναι τα μέρη του μοντέλου. Ωστόσο, οι έξοδοι αυτών αποτελούν την είσοδο στο τελευταίο κομμάτι του μοντέλου. Όλες οι πληροφορίες από τα μέρη αυτά ενώνονται και περνάνε ως είσοδο σε ένα τελευταίο νευρωνικό δίκτυο. Το δίκτυο αυτό αποτελείται από ένα αμφίδρομο LSTM στρώμα, 2 dense στρώματα που ανάμεσα έχουν dropout στρώματα, ακριβώς όπως στο νευρωνικό δίκτυο στο 2 μέρος του μοντέλου 4.3. Τέλος, πριν την τελική έξοδο χρησιμοποιούμε ένα flatten στρώμα. Έτσι, δημιουργήσαμε ένα υβριδικό μοντέλο συστάσεων, το οποίο δέχεται διαφόρων ειδών πληροφορίες 4.6.



Σχήμα 4.6: Το τελικό μέρος του Movielens100k μοντέλου

Το τελικό Hybrid μοντέλο με όλα τα μέρη του φαίνεται στο σχήμα 4.7.

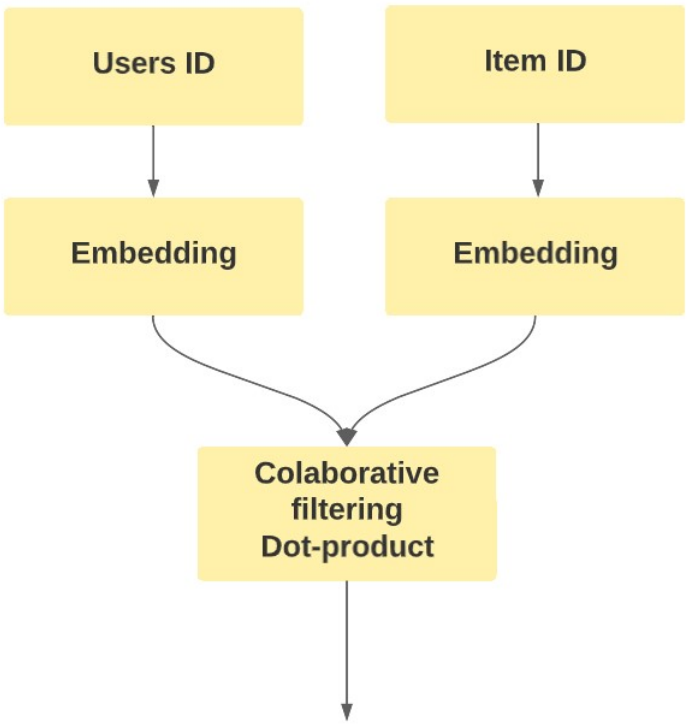


Σχήμα 4.7: Το hybrid μοντέλο για το movielens100k μοντέλο

4.2.2 Μοντέλο για το Movielens1m

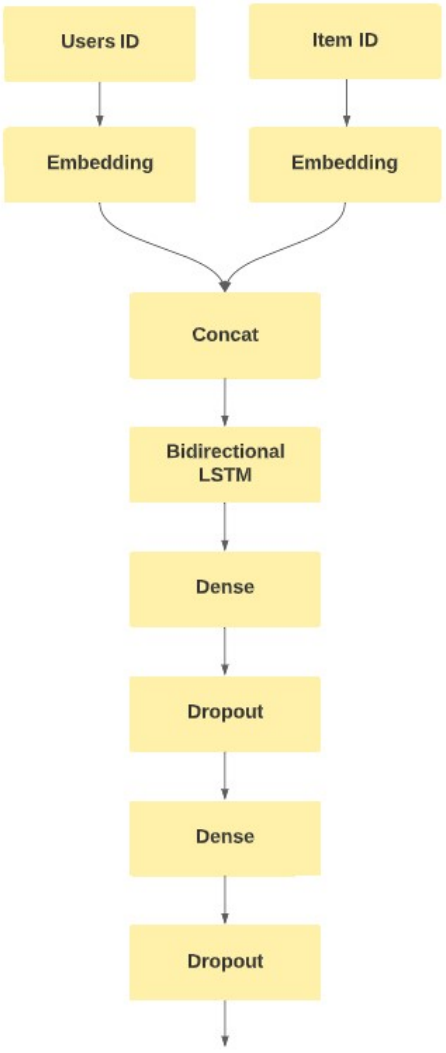
Το Movielens1m είναι ένα σύνολο δεδομένων που είναι παρόμοιο με το Movielens100k. Η μόνη διαφορά τους είναι ότι υπάρχουν περισσότερες εγγραφές στις βαθμολογίσεις ταινιών. Γι'αυτό το λόγο τα μοντέλα είναι ίδια. Ως είσοδο στο μοντέλο έχουμε το users ID, το item ID, τα είδη των ταινιών και τις επιπλέον πληροφορίες που έχουμε δημιουργήσει από την προεπεξεργασία των δεδομένων, όπως φαίνεται στο σχήμα 4.8.

Σχήμα 4.8: Οι είσοδος στο Movielens1m μοντέλο.



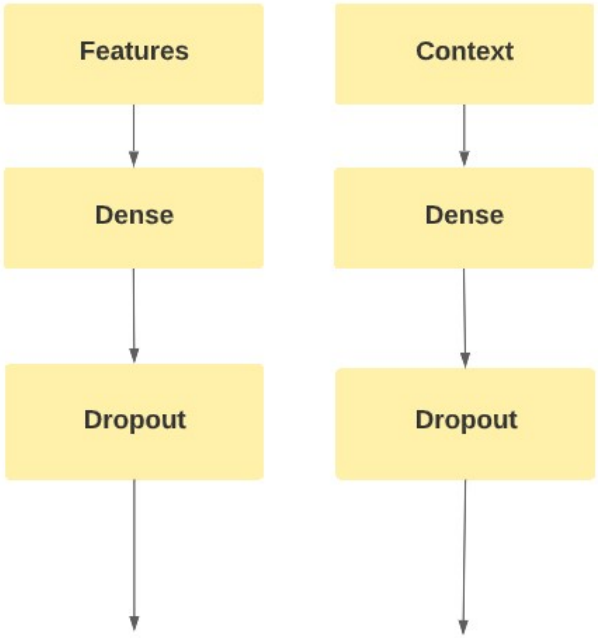
Σχήμα 4.9: Dot-product για το μοντέλο Movielens1m

Το μοντέλο αποτελείται από ακριβώς ίδια μέρη. Στην αρχή, έχουμε τα users ID και τα items ID, τα οποία περνάνε στις στρώσεις ενσωμάτωσης και δημιουργούν το Dot-product, όπως φαίνεται στο σχήμα 4.9 Στο σχήμα 4.10 παρατηρούμε την δομή του δεύτερου μέρους που αποτελείται, επίσης, από τις στρώσεις ενσωμάτωσης, τα οποία αποτελούν είσοδο σε ένα νευρωνικό δίκτυο. Τέλος, για τις 2 τελευταίες εισόδους υπάρχουν απλά σε δομή μοντέλα, όπως και στο Movielens100k μοντέλο. Την δομή του δικτύου την βλέπουμε στο σχήμα 4.11. Όλα αυτά τα μέρη συνδέονται και αποτελούν είσοδο σε ένα νευρωνικό δίκτυο για να δημιουργηθεί το τελικό αποτέλεσμα του μοντέλου.

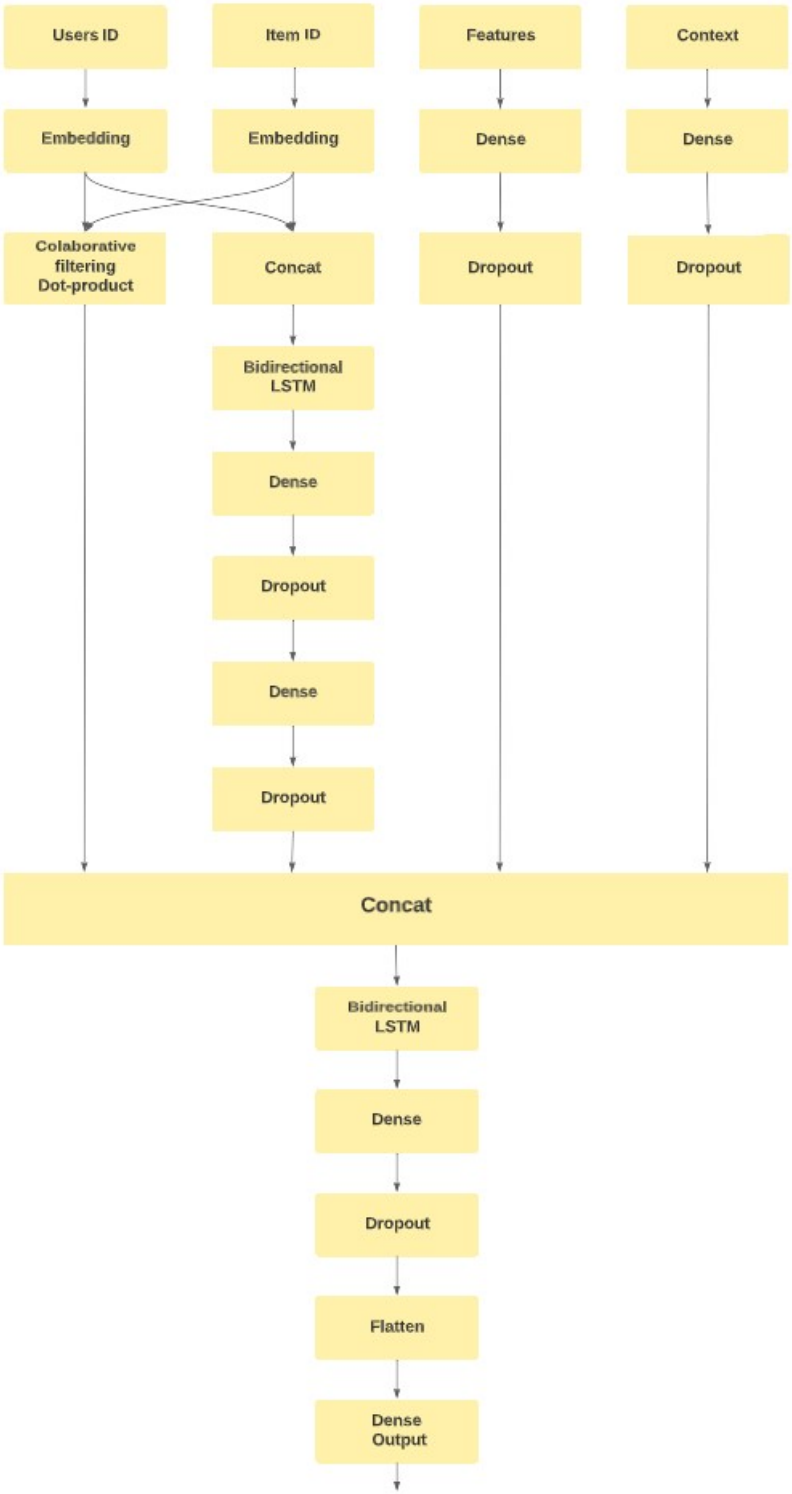


Σχήμα 4.10: Collaborative filtering και νευρωνικό δίκτυο του Movielens1m μοντέλου

Οι περισσότεροι παράμετροι είναι ίδιοι με το movielens100k μοντέλο. Υπάρχουν ίδιοι παράμετροι στα dropout στρώματα, ίδιοι παράμετροι για το l2 kernel regulizer. Οι κύρια διαφορά τους είναι στο χρόνο εκπαίδευσης τον 2 αυτών μοντέλων. Το πρώτο, επειδή η βάση δεδομένων περιέχει λιγότερα δεδομένα θέλει λιγότερο χρόνο για να εκπαιδευτεί και λιγότερες επαναλήψεις. Το τελικό μοντέλο για το movielens1m φαίνεται στο σχήμα 4.12.



Σχήμα 4.11: Το content-based και το knowledged-based μέρη του Movielens1m μοντέλου

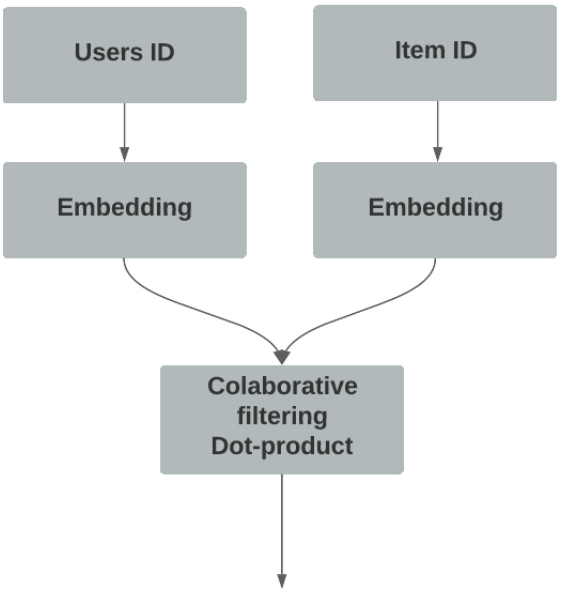


Σχήμα 4.12: Το hybrid μοντέλο για το movielens1m μοντέλο

4.2.3 Μοντέλο για το Book-Crossing

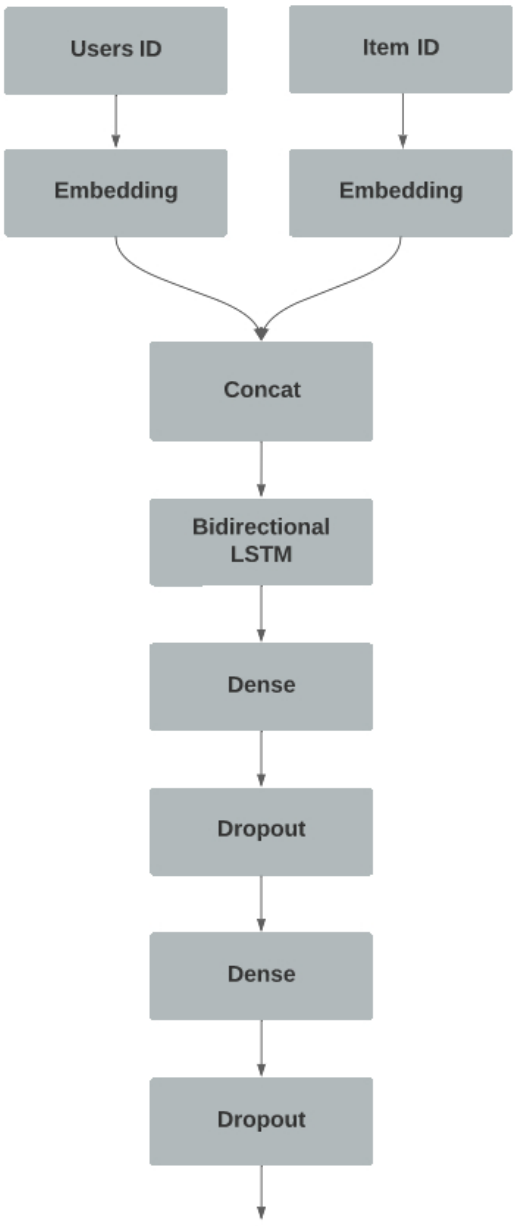
Το μοντέλο για το Book-Crossing έχει κάποιες σημαντικές διαφορές σε σχέση με τα 2 προηγούμενα μοντέλα. Κάποιες από τις δυνατότητες που μας δίνουν τα δεδομένα είναι αρκετά σημαντικές, όπως παραδείγματος χάρη υπάρχουν πολλοί χρήστες, πολλές αξιολογήσεις βιβλίων κ.α. Ωστόσο, δεν υπάρχει πληροφορία για το είδος του βιβλίου. Οι πληροφορίες που έχουμε είναι για τον τίτλο του βιβλίου, πότε κυκλοφόρησε ή ακόμη και ποιος είναι ο εκδότης. Δεν υπάρχει, όμως η πληροφορία στην οποία αναφέρεται το είδος, δηλαδή η κατηγορία του βιβλίου. Έτσι, το κομμάτι ενός υβριδικού συστήματος συστάσεων, το οποίο βασίζεται στο περιεχόμενο σε αυτό το μοντέλο δεν υπάρχει. Για το μοντέλο αυτό, λοιπόν, έχουμε 3 εισόδους. Οι 2 πρώτες είναι είναι το ID των χρηστών και το ID των βιβλίων. Η τρίτη είσοδος έχει την ηλικία του χρήστη, η οποία χωρίζεται σε 3 μέρη. Αν ο χρήστης είναι κάτω των 18, αν ο χρήστης είναι άνω των 65 και αν ο χρήστης βρίσκεται ενδιάμεσα στις ηλικίες αυτές. Έτσι, στο σχήμα 4.13 έχουμε τις εισόδους στο μοντέλο αυτό.

Σχήμα 4.13: Είσοδοι για το μοντέλο Book-crossing

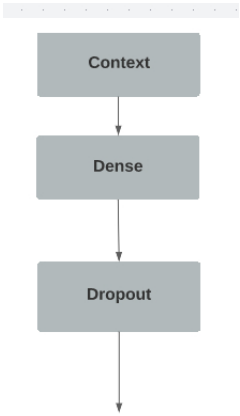


Σχήμα 4.14: Dot-product για το μοντέλο Book-crossing

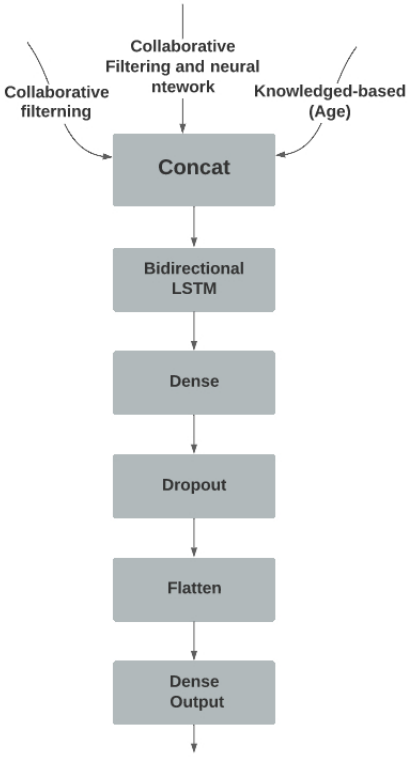
Στην συνέχεια, όπως βλέπουμε στο σχήμα 4.14 και στο σχήμα 4.15 ακολουθούμε την ίδια ακριβώς διαδικασία όπως και στα προηγούμενα μοντέλα. Έχουμε 2 στρώσεις ενσωμάτωσης, το ένα με τα ID των χρηστών και το άλλο με τα ID των βιβλίων. Στο σχήμα 4.14 δημιουργείται το Dot-product, ενώ στο σχήμα 4.15 τα στρώματα ενσωμάτωσης καταλήγουν σε ένα νευρωνικό δίκτυο.



Σχήμα 4.15: Collaborative filtering και νευρωνικό δίκτυο του Book-crossing μοντέλου

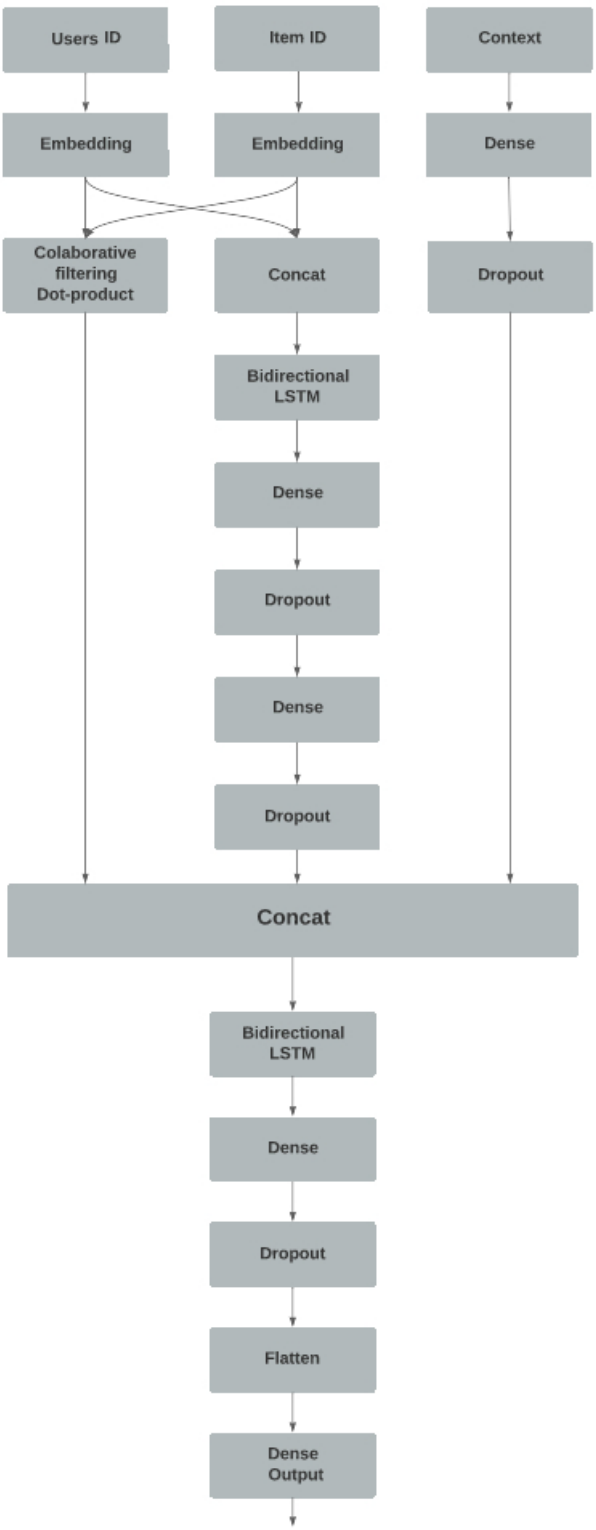


Σχήμα 4.16: Το knowledged-based μέρος του Book-crossing μοντέλου



Σχήμα 4.17: Το τελικό μέρος του Book-crossing μοντέλου

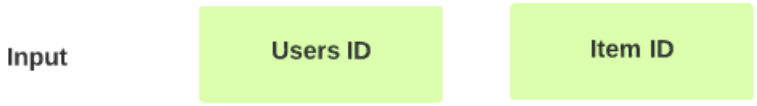
Το τρίτο κομμάτι του μοντέλου που δέχεται ως είσοδο την ηλικία των χρηστών αποτελείται από απλά dense και dropout στρώματα, όπως φαίνεται στο σχήμα 4.16. Τέλος, όλα τα παραπάνω μέρη ενώνονται και αποτελούν την είσοδο σε ένα νευρωνικό δίκτυο. Στο σχήμα 4.17 παρακάτω φαίνεται το τελικό κομμάτι του μοντέλου. Ενώνοντας όλα τα μέρη του μοντέλου φτάνουμε στο τελικό μοντέλο, το οποίο φαίνεται στο σχήμα 4.18. Όλες τα ποσοστά και τα δεδομένα για το dropout στρώματα και για τον l2 kernel regularizer είναι ίδιες με τα προηγούμενα μοντέλα.



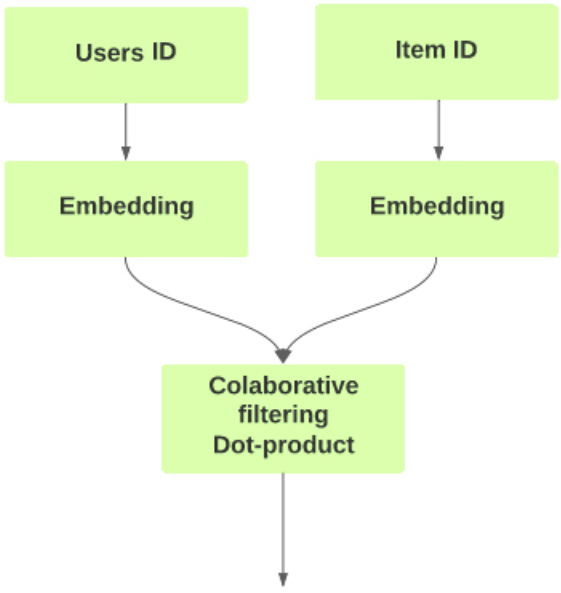
Σχήμα 4.18: Το hybrid μοντέλο για το Book-crossing dataset

4.2.4 Μοντέλο για το Film-Trust

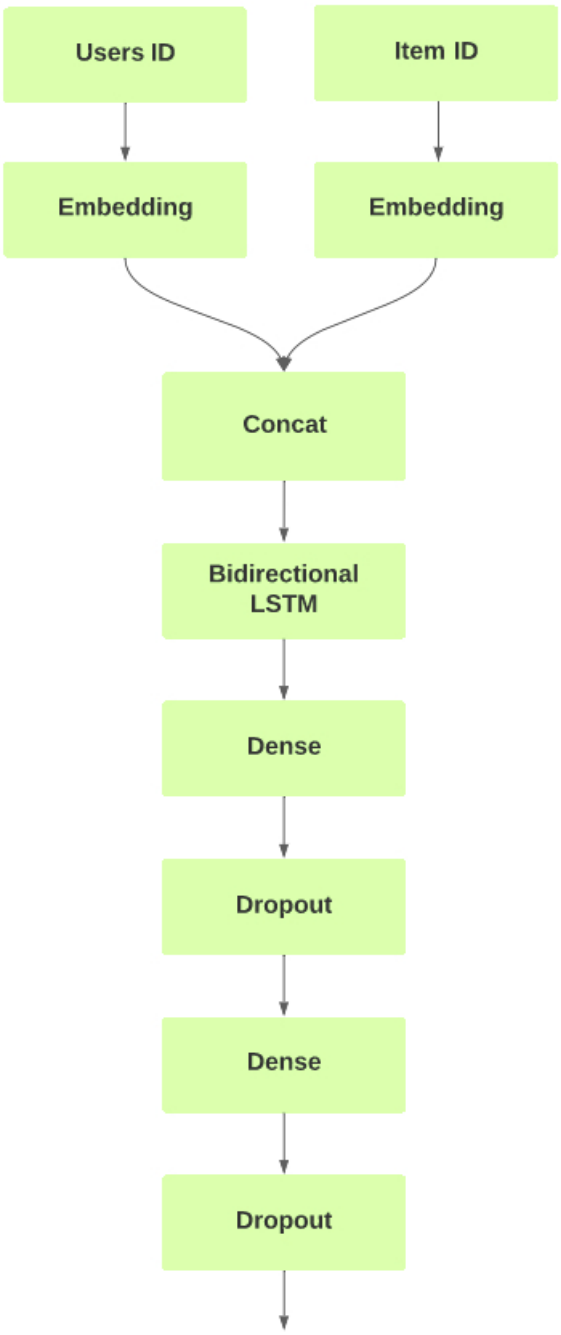
Το τελευταίο μοντέλο είναι το πιο απλό μοντέλο από τα 4, καθώς δεν περιέχει πολλές πληροφορίες ,ώστε να μπορέσουν να αξιοποιηθούν. Έτσι, ως είσοδο στο μοντέλο αυτό έχουμε μόνο τα ID των χρηστών και των ταινιών, οι οποίες πληροφορίες συνεχίζουν σε στρώσεις ενσωμάτωσης. Το ένα δημιουργεί το Dot-product και το άλλο συνεχίζει σε ένα νευρωνικό δίκτυο. Αυτά είναι τα κομμάτια αυτού το μοντέλου. Όπως φαίνεται στο σχήμα 4.19 μας δείχνει την είσοδο του μοντέλου, στο σχήμα 4.20 βλέπουμε το απλό κομμάτι του συνεργατικού φιλτραρίσματος και τέλος στο σχήμα 4.21 φαίνεται το κομμάτι του συνεργατικού φιλτραρίσματος σε συνδυασμό με νευρωνικό δίκτυο.



Σχήμα 4.19: Οι είσοδο για το Film-Trust μοντέλο

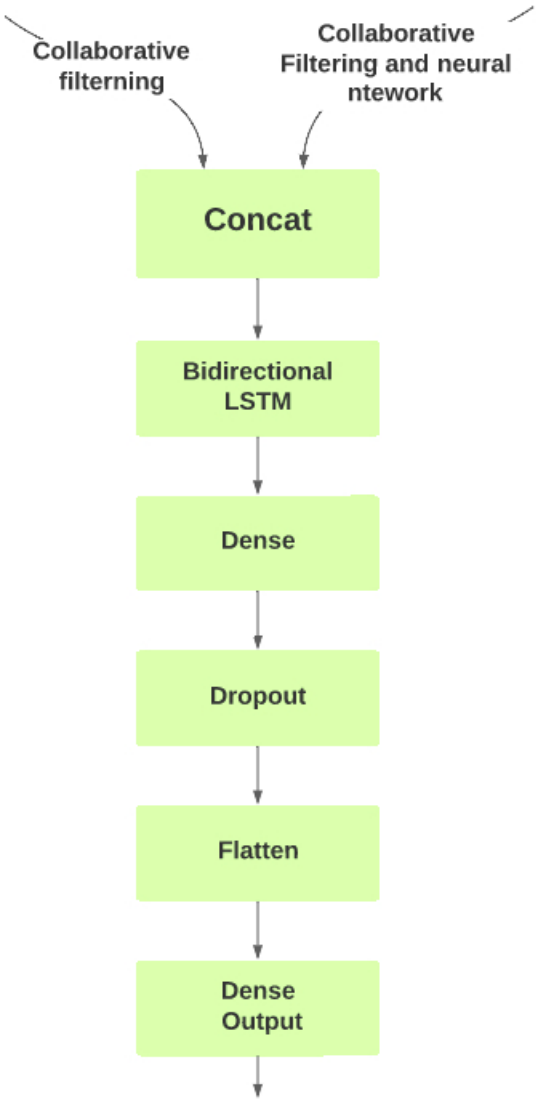


Σχήμα 4.20: Dot-product για το μοντέλο Film-Trust



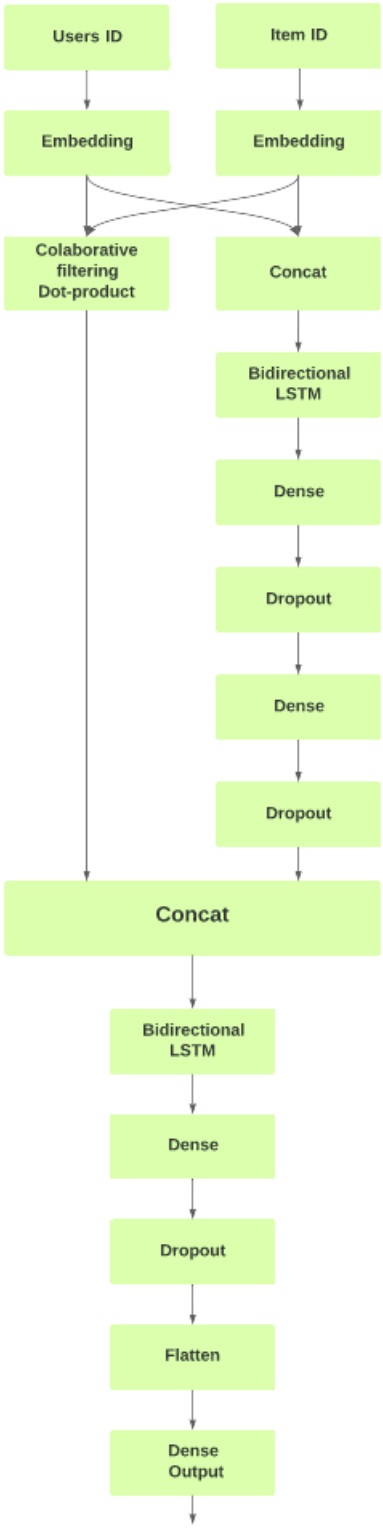
Σχήμα 4.21: Collaborative filtering και νευρωνικό δίκτυο του Film-Trust μοντέλου

Τέλος, τα 2 κομμάτια αυτά ενώνονται και αποτελούν την είσοδο για το τελευταίο μέρος του μοντέλου που είναι το ίδιο νευρωνικό μοντέλο με τα προηγούμενα μοντέλα, όπως φαίνεται στο σχήμα 4.21. Το τελικό μοντέλο φαίνεται στο σχήμα 4.21.



Σχήμα 4.22: Dot-product για το μοντέλο Film-Trust

Όπως και σε όλα τα μοντέλα έτσι και σε αυτό όλες οι παράμετροι για τα dropout στρώματα και το l2 kernel regularizer είναι οι ίδιες.



Σχήμα 4.23: Το hybrid μοντέλο για το Film-Trust dataset

Κεφάλαιο 5

Τρόποι αξιολόγησης των μοντέλων

5.1 Εισαγωγή

Η αξιολόγηση του συστήματος είναι σημαντικό να πραγματοποιηθεί με σωστό τρόπο, δηλαδή με κατάλληλους τρόπους ανάλογα με το σύστημα. Η αξιολόγηση των μοντέλων πραγματοποιείται με 2 τρόπους. Ένας τρόπος είναι χρησιμοποιώντας συναρτήσεις σφάλματος. Ο τρόπος αυτός είναι αρκετά εύκολος και γρήγορος τρόπος και είναι ένας καλός τρόπος για να πάρεις μια πρώτη εικόνα για το σύστημα. Ο δεύτερος τρόπος αξιολόγησης είναι το ποσοστό επιτυχίας του μοντέλου. Υπάρχουν διάφοροι τρόποι μέτρησης της επιτυχίας του μοντέλου. Για τα συγκεκριμένα μοντέλα χρησιμοποιήθηκαν 4 συναρτήσεις σφάλματος και 1 που αφορά το ποσοστό επιτυχίας των προβλέψεων. Συγκεκριμένα, χρησιμοποιήθηκαν το MSE, RMSE, MAE, R-squarred [20] και για τον υπολογισμό του ποσοστού επιτυχίας πραγματοποιείται μια συγκεκριμένη διαδικασία που υπολογίζει το ποσοστό επιτυχίας του συνόλου.

5.1.1 Μέσο Τετραγωνικό Λάθος

Είναι από τις πιο απλές και από τις πιο συνηθισμένες συναρτήσεις για να υπολογίσεις το σφάλμα ενός μοντέλου. Η μαθηματική έκφραση της Μέσο Τετραγωνικό Λάθος (Mean Squared Error, MSE) [21] συνάρτησης είναι αυτή:

$$MSE = \sum_{i=1}^{D} (x_i - y_i)^2$$

Η συνάρτηση στην ουσία είναι απλή γιατί είναι ο μέσος όρος του τετραγώνου της διαφοράς ανάμεσα στην τιμή που πρόβλεψε το σύστημα και στην πραγματική τιμή. Το κυριότερο όφελος από την συνάρτηση είναι πως με την συνάρτηση αυτή το μοντέλο καταφέρνει να χειρίζεται εύκολα τις ακραίες λανθασμένες προβλέψεις χάρη στο τετράγωνο της συνάρτησης καθώς δίνει περισσότερη βάση σε αυτές τις περιπτώσεις. Τέλος, το MSE είναι πάντα θετικό και όσο μικρότερη είναι η τιμή του τόσο το καλύτερο.

5.1.2 Μέση τετραγωνική ρίζα σφάλματος

Είναι αρκετά συνηθισμένη συνάρτηση, χρησιμοποιείται κυρίως για ποσοτικά δεδομένα και για 'time series' μοντέλα. [22] Ο τύπος της συνάρτησης είναι αυτός:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{n}}$$

Αν παρατηρήσουμε το τύπο από μαθηματικής πλευράς, αν αφαιρέσουμε την διαίρεση με το n παρατηρούμε πως είναι ο τύπος της ευκλείδειας απόστασης μεταξύ 2 διανυσμάτων. Έτσι, η μέση τετραγωνική ρίζα σφάλματος (Root Mean Square Error, RMSE) αποτελεί μία κανονικοποίηση ανάμεσα στις τιμές που έχει προβλέψει το μοντέλο με τις πραγματικές τιμές. Η συνάρτηση αυτή δίνει περισσότερη βάση σε μεγάλα σφάλματα, το RMSE αποτελεί το μέσο σφάλμα του μοντέλου και όσο μικρότερες είναι οι τιμές τόσο καλύτερα λειτουργεί το μοντέλο.

5.1.3 Μέσο απόλυτο σφάλμα

Το MAE [23] μπορεί στην όψη να μοιάζει αρκετά με το MSE, όμως έχουν αντίθετες ιδιότητες. Για τον υπολογισμό του σφάλματος αυτού παίρνουμε την απόλυτη τιμή της διαφοράς της τιμής που προέβλεψε το μοντέλο με την πραγματική τιμή και το διαιρούμε με το σύνολο των δεδομένων. Η συνάρτηση αυτή παράγει πάντα θετικό αποτέλεσμα και όσο μικρότερη είναι η τιμή του τόσο καλύτερα δουλεύει το μοντέλο, όπως και με την MSE. Η μαθηματική έκφραση του μέσου απόλυτου σφάλματος (Mean Absolute Error, MAE) συνάρτησης είναι:

$$MAE = \frac{1}{n} \sum_{i=1}^{D} |x_i - y_i|$$

Τέλος, το κυριότερο χαρακτηριστικό της συνάρτησης αυτής είναι πως ακόμη και να πραγματοποιηθεί μία ακραία λανθασμένη πρόβλεψη δεν θα δώσει τόσο βάση σε αυτό το αποτέλεσμα και θα προσπαθήσει να δει αν το μοντέλο έχει μία πιο γενική εικόνα για το μοντέλο και μας δείχνει αν το μοντέλο συμπεριφέρεται σωστά σε διαφορετικές καταστάσεις και όχι σε κάτι πολύ συγκεκριμένο.

5.1.4 Συντελεστής προσδιορισμού

Η συνάρτηση αυτή είναι λίγο διαφορετική από τις προηγούμενες. [24] Το αποτέλεσμα της συνάρτησης μας δείχνει κατά πόσο καλά κατανοεί τα δεδομένα το μοντέλο. Ο μαθηματικός τύπος είναι:

$$R_{squarred} = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}$$

$$R_{squarred} = 1 - \frac{\sum_i^N (y_i - \hat{y}_i)^2}{\sum_i^N (y_i - \bar{y})^2}$$

Το Sum of Squares Regression αντιπροσωπεύει τη συνολική διακύμανση όλων των προβλεπόμενων τιμών που βρίσκονται στη γραμμή ή το επίπεδο παλινδρόμησης από τη μέση τιμή όλων των τιμών των μεταβλητών απόκρισης. Το sum of squares total αντιπροσωπεύει τη συνολική διακύμανση των πραγματικών τιμών από τη μέση τιμή όλων των τιμών των μεταβλητών απόκρισης. Η τιμή του συντελεστή προσδιορισμού (R-squarred) χρησιμοποιείται για τη μέτρηση του 'goodness of fit or best-fit line'. Τέλος, σε αντίθεση με τις άλλες συναρτήσεις η τιμή της συνάρτησης όσο μεγαλύτερη είναι τόσο καλύτερο για το μοντέλο.

5.1.5 Ποσοστό επιτυχίας

Με την μέτρηση των σωστών αποτελεσμάτων του μοντελου μπορείς να πάρεις μία γενική εικόνα για το πως συμπεριφέρεται το μοντέλο. Ένας τελευταίος τρόπος αξιολόγησης του μοντέλου είναι ο υπολογισμός των σωστών επιλογών ανάλογα με των χρήστη. Έτσι, η απλή αυτή μέτρηση έγινε με τον τύπο:

$$Accuracy = \frac{\text{Total number of correct predictions}}{\text{Total number of predictions}}$$

Έτσι, για να υπολογίσουμε το αποτέλεσμα της συνάρτησης αυτής, αφού τελειώσουμε με την εκπαίδευση του μοντέλου τότε πραγματοποιείται πρόβλεψη για κάθε χρήστη από την αντίστοιχη βάση δεδομένων. Με αυτόν τον τρόπο υπολογίζουμε πόσες από τις προβλέψεις ήταν σωστές και πόσες ήταν σύνολο και υπολογίζουμε το ποσοστό αυτό. Ένα από τα θετικά χαρακτηριστικά του ποσοστού είναι πως έτσι παίρνεις μία πραγματική εικόνα για την συμπεριφορά του μοντέλου σου. Ένα από τα αρνητικά χαρακτηριστικά που έχουν αυτές οι μορφές αξιολόγησης του μοντέλου είναι πως το αποτέλεσμα του μοντέλου είναι μικροί αριθμοί με αποτέλεσμα πολλές φορές οι διαφορά της λάθος πρόβλεψης να είναι πολύ μικρή με την σωστή πρόβλεψη. Αυτό έχει ως αποτέλεσμα το μοντέλο να προτείνει κάτι που είναι παρόμοιο με αυτό που είχε πραγματικά επιλέξει ο χρήστης για πάρα πολύ μικρή διαφορά.

5.1.6 Συμπεράσματα για την αξιολόγηση των μοντέλων

Αυτοί είναι οι τρόποι για να αξιολογήσουμε όλα τα μοντέλα με όλα τις βάσεις δεδομένων. Στην συνέχεια, θα παρουσιάσουμε τα τελικά αποτελέσματα των μοντέλων με όλους τους τρόπους αξιολόγησης. Έπειτα, θα πραγματοποιηθεί σύγκριση με τα μοντέλα [12] και των δικών μου μοντέλων. Τα μοντέλα [12] έχουν ακριβώς τους ίδιους τρόπους αξιολόγησης και έτσι μπορούμε εύκολα να πραγματοποιήσουμε την σύγκριση.

Κεφάλαιο 6

Αξιολόγηση μοντέλων και σύγκριση με το DNNRec: A novel deep learning based hybrid recommender system

6.1 Εισαγωγή

Σε αυτή την ενότητα, έπειτα από την παρουσίαση της επεξεργασίας των δεδομένων και από την παρουσίαση των μοντέλων, θα δείξουμε τα αποτελέσματα των μετρήσεων που πραγματοποιήθηκαν στα μοντέλα. Οι τρόποι αξιολόγησης των μοντέλων είναι αυτοί που δείξαμε στο κεφάλαιο 4. Αρχικά, θα παρουσιαστούν τα αποτελέσματα για όλα τα μοντέλα και, στην συνέχεια θα γίνει η σύγκριση με τα μοντέλα από το DNNRec: A novel deep learning based hybrid recommender system [12].

6.2 Αποτελέσματα για το Movielens100k μοντέλο

Αρχικά, για το Movilens100k μοντέλο οι τελικές μετρήσεις είναι για το MSE είναι 0.0409, για το RMSE είναι 0.0412, για το MAE είναι 0.1655, για το R-squared είναι 0.2934 και το ποσοστό επιτυχίας είναι 0.2019. Οι μετρήσεις αυτές υπολογίστηκαν έπειτα από 10 epochs με batch size 128 και learning rate 5e-3, όπως φαίνεται και στον πίνακα. 6.1

6.3 Αποτελέσματα για το Movielens1m μοντέλο

Στην συνέχεια, για το Movilens1m μοντέλο οι τελικές μετρήσεις, όπως φαίνεται και στον πίνακα 6.1, είναι για το MSE είναι 0.1110, για το RMSE είναι 0.1114, για το MAE είναι 0.2698, για το R-squared είναι 0.4107 και το ποσοστό επιτυχίας είναι 0.4247. Οι μετρήσεις αυτές υπολογίστηκαν έπειτα από 10 epochs με batch size 128 και learning rate 5e-3, όπως και στο Movielens100k μοντέλο.

Model	MSE	RMSE	MAE	R-squared	Accuracy
HRS Movielens100k	0.0409	0.0412	0.1655	0.2934	0.2019
HRS Movielens1m	0.1110	0.1114	0.2698	0.4107	0.4247
HRS Book-Crossing	0.0346	0.0351	0.1447	0.2327	0.2047
HRS Film-Trust	0.2783	0.2793	0.5075	0.5469	0.3577

Πίνακας 6.1: MSE, RMSE, MAE, R-squared και accuracy στα ML100K, Book-Crossing, ML1M και στο Film-Trust datasets.

6.4 Αποτελέσματα για το Book-Crossing μοντέλο

Έπειτα, για το Book-Crossing μοντέλο οι τελικές μετρήσεις, όπως φαίνεται και στον πίνακα 6.1, είναι για το MSE είναι 0.0346, για το RMSE είναι 0.0351, για το MAE είναι 0.1447, για το R-squared είναι 0.2327 και το ποσοστό επιτυχίας είναι 0.4247. Οι μετρήσεις αυτές υπολογίστηκαν με τα ίδια χαρακτηριστικά όπως και στα προηγούμενα μοντέλα.

6.5 Αποτελέσματα για το Film-Trust μοντέλο

Τέλος, για το Film-Trust μοντέλο οι τελικές μετρήσεις είναι για το MSE είναι 0.2783, για το RMSE είναι 0.2793, για το MAE είναι 0.5075, για το R-squared είναι 0.5469 και το ποσοστό επιτυχίας είναι 0.4247, όπως φαίνεται και στον πίνακα 6.1. Οι μετρήσεις αυτές υπολογίστηκαν με τα 50 epochs, με batch size 64 και με learning rate 5e-3.

6.6 Σύγκριση με το DNNRec: A novel deep learning based hybrid recommender system

Οι αξιολογίσεις πραγματοποιήθηκαν ακριβώς με τους ίδιους τρόπους σε όλα τα μοντέλα, γι αυτό το λόγο είναι εφικτό να πραγματοποιηθεί σύγκριση μεταξύ των μοντέλων τις ίδιας βάσης δεδομένων. Τα αποτελέσματα φαίνονται στον πίνακα 6.2

Dataset	Model	MSE	RMSE	MAE	R-squared
Mavielens100k	My HRS	0.0409	0.0412	0.1655	0.2934
	DNNRec	0.747	0.864	0.666	0.338
Movielens1m	My HRS	0.1110	0.1114	0.2698	0.4107
	DNNRec	0.766	0.875	0.688	0.417
Book-Crossing	My HRS	0.0346	0.0351	0.1447	0.2327
	DNNRec	2.759	1.661	1.280	0.169
Film-Trust	My HRS	0.2783	0.2793	0.5075	0.5469
	DNNRec	0.649	0.805	0.626	0.225

Πίνακας 6.2: Σύγκριση όλων των μοντέλων με τα αντίστοιχα datasets.

Παρατηρούμε πως σχεδόν σε όλες τις μετρήσεις τα μοντέλα αυτής της εργασίας έχουν καλύτερα αποτελέσματα. Αυτό είναι λογικό, αφού η υλοποίηση των μοντέλων και η επεξεργασία των δεδομένων βασίστηκαν στην λογική του DNNRec: A novel deep learning based hybrid recommender system και με κάποιες επιπλέον τεχνικές και με επιπλέον πληροφορία από τα ίδια τα δεδομένα καταφέραμε να πάρουμε καλύτερα αποτελέσματα.

Κεφάλαιο 7

Συμπεράσματα

7.1 Σύνοψη και συμπεράσματα

Στην διπλωματική αυτή καταφέραμε να δημιουργήσουμε 4 διαφορετικά υβριδικά συστήματα συστάσεων για 4 διαφορετικές βάσεις δεδομένων. Εξηγήσαμε πως πραγματοποιήθηκε η επεξεργασία των δεδομένων και ποιες δυσκολίες αντιμετωπίσαμε σε όλα τις βάσεις δεδομένων, ώστε να είναι έτοιμα για την είσοδό τους στα μοντέλα. Επίσης, εξηγήσαμε αναλυτικά την υλοποίηση κάθε μοντέλου και παρουσιάσαμε αναλυτικά κάθε μέρος τους. Τα αποτελέσματα που πήραμε ήταν κυρίως θετικά, καθώς εκτός από τις τελικές μετρήσεις αξιολόγησης που πήραμε, τα μοντέλα δείξαμε πως δουλεύουν και δίνουν σωστά αποτελέσματα έπειτα από κάποια πειράματα που πραγματοποιήθηκαν. Παρόλα αυτά, υπάρχουν κάποια αρνητικά σημεία, όπως η δυσκολία να αποκτήσεις και να δημιουργήσεις μία ιδανική βάση δεδομένων, ώστε τα μοντέλα να εκπαιδεύονται καλύτερα. Για να το πετύχεις αυτό πρέπει όσο είναι δυνατόν το σύστημα να αντλεί όσο γίνεται περισσότερη πληροφορία απ'όλους τους χρήστες. Αυτό θα βοηθήσει, επίσης ακόμη περισσότερο τα μοντέλα αφού τα υβριδικά συστήματα συστάσεων είναι σχεδιασμένα να δέχονται μεγάλου όγκο πληροφορίας. Τέλος, ένα δύσκολο κομμάτι της διπλωματικής αυτής είναι πως η επεξεργασία των δεδομένων και η εκπαίδευση των μοντέλων απαιτούν μεγάλη υπολογιστική ισχύ και αυτό έχει ως αποτέλεσμα να κάνει όλη την διαδικασία λίγο πιο δύσκολη.

7.2 Μελλοντικές επεκτάσεις

Σε αυτή την ενότητα θα αναφερθούν κάποιες από τις σημαντικές επεκτάσεις που θα μπορούσαν να πραγματοποιηθούν. Αρχικά, ένα από τα σημαντικότερα κομμάτια που θα μπορούσε να υλοποιηθεί είναι η αξιοποίηση όλων το δεδομένων που δίνουν οι βάσεις δεδομένων. Ένα παράδειγμα είναι πώς θα μπορούσε να αξιοποιηθούν πληροφορίες για τον χρήστη, όπως επάγγελμα, περιοχή κατοικίας και διάφορες τέτοιες πληροφορίες. Σημαντική επίσης θα ήταν η δημιουργία βάσεων δεδομένων με όλες τις απαραίτητες πληροφορίες και χωρίς ελλιπή στοιχεία, ώστε να μπορούν να εκμεταλλευτούν πλήρως. Επίσης, αφού το σύστημα λειτουργούσε σε πραγματικό χρόνο θα ήταν ωραία προσθήκη, η αυτόματη ανανέωση των βάσεων δεδομένων, κάθε φορά που ο χρήστης αλληλεπιδρά με την εφαρμογή. Με αυτό το τρόπο θα μπορέσουμε να δημιουργήσουμε μία βάση που θα πληροί όλες τις προϋποθέσεις. Τέλος, αφού πραγματοποιηθούν οι 2 προηγούμενες μελλοντικές προσθήκες θα πρέπει να πραγματοποιείται και η επανεκπαίδευση του μοντέλου με τα καινούρια δεδομένα. Αυτό θα μπορούσε να πραγματοποιηθεί όταν το 30% των χρηστών στην εφαρμογή έχει αλληλεπιδράσει με καινούρια αντικείμενα, τα οποία έχουν καταγραφεί στη βάση δεδομένων.

Βιβλιογραφία

[1] C. A. GOMEZ-URIBE and N. HUNT, "The netflix recommender system: Algorithms, business value and innovation," *ACM Transactions on Management Information Systems*, Dec. 2015.

[2] E. S. Paul Covington Jay Adams, "Deep neural networks for youtube recommendations,"

[3] *Pycharm*, `https://www.jetbrains.com/pycharm/`.

[4] *Cuda*, `https://en.wikipedia.org/wiki/CUDA`.

[5] *Cudnn*, `https://developer.nvidia.com/cudnn`.

[6] *Tensorflow*, `https://www.tensorflow.org/api_docs`.

[7] P. B. Thorat, R. M. Goudar, and S. Barve, "Survey on collaborative filtering, content-based filtering and hybrid recommendation system," *Computer Engineering MIT Academy of Engineering Pune India*, Jan. 2015.

[8] C. C. Aggarwal, *Recommender Systems: The Textbook*.

[9] N. Good, J. B. Schafer, J. A. Konstan, *et al.*, "Combining collaborative filtering with personal agents for better recommendations," *GroupLens Research Project Department of Computer Science and Engineering University of Minnesota*,

[10] Y. Afoudi, M. Lazaar, and M. A. Achhab, "Hybrid recommendation system combined content-based filtering and collaborative prediction using artificial neural network," *ENSIAS, Mohammed V University in Rabat, Morocco ENSA, Abdelmalek Essaadi University in Tetuan, Morocco*, Dec. 2021.

[11] X. N. Lam, T. Vu, T. D. Le, and A. D. Duong, "Addressing cold-start problem in recommendation systems," *ACM Transactions on Management Information Systems*, Jan. 2008.

[12] B. B. Kiran R Pradeep Kumar, "Dnnrec: A novel deep learning based hybrid recommender system," *Information Technology and Systems, Indian Institute of Management Lucknow, India, Indian Institute of Management, Raipur*, 2019.

[13] J. Wang, L. Zhu, T. Dai, and Y. Wang, "Deep memory network with bi-lstm for personalized context-aware citation recommendation," *School of Software Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China*, May 2020.

[14] C. Z. 1, J. You, X. Wen, and andXiaowu Li, "Deep bi-lstm networks for sequential recommendation," *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650504, China, Computer Technology Application Key Lab of Yunnan Province, Kunming 650504, China*, 2020.

[15] R. L. Rosa, G. M. Schwartz, W. V. Ruggiero, and D.́o. Z. R.́ıguez, "A knowledge-based recommendation system that includes sentiment analysis and deep learning," *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, Apr. 2019.

[16] Q. Shambour, "A deep learning based algorithm for multi-criteria recommender systems," *Department of Software Engineering, Faculty of Information Technology, Al-Ahliyya Amman University, Amman, Jordan*, Oct. 2020.

[17] H. Daneshvar and R. Ravanmehr, "A social hybrid recommendation system using lstm and cnn," *Department of Computer Engineering, CentralTehran Branch, Islamic Azad University, Tehran,Iran*, Nov. 2021.

[18] *Layer weight regularizers*, `https://keras.io/api/layers/regularizers/`.

[19] *Dropout layer*, `https://keras.io/api/layers/regularization_layers/dropout/`.

[20] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, *Recommender Systems Handbook*. 2010.

[21] *Mean squared error (mse)*, `https://statisticsbyjim.com/regression/mean-squared-error-mse/`.

[22] *Root mean square error (rmse)*, `https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/`.

[23] *Mean absolute error (mae)*, https://stephenallwright.com/good-mae-score/.

[24] *R-squared*, https://corporatefinanceinstitute.com/resources/knowledge/other/r-squared/.