Μαρίνος Μάριος (Dai17147) Recommender Systems in Python.

# Σχολιασμός Κώδικα.

Όσον αφορά τον κώδικα, χρησιμοποιήσα Python και την βιβλιοθήκη pandas συγκεκριμένα. Πέρα από την προεπεξεργασία που αναφέρεται, έκανα και κάποια πράματα ακόμη. Αρχικά, αφαίρεσα τις στήλες 'imageUrlS', 'imageUrlM', 'imageUrlL' μιας και δεν χρειάζονται πουθενά. Έπειτα, κοίταξα αν οι βαθμολογίες που υπάρχουν στο books\_ratings csv υπάρχουν και τα αντίστοιχα τους βιβλία στο books.csv οπότε όποιες βαθμολογίες δεν έχουν το αντίστοχο βιβλίο τις αφαίρεσα από το ratings dataframe. Το ίδιο έκανα και για τους χρήστες αλλά εκεί όλοι αντιστοιχούσαν. Έπειτα σε δύο συγκεκριμένα βιβλία είχε year = 'DK Publishing Inc' ως λάθος και βλέποντας τις αντίστοιχες εγγραφές παρατήρησα οτι ήταν όλα μία στήλη πιο δεξιά (δηλαδή στην στήλη bookAuthor είχε την τιμή του bookTitle κ.ο.κ.) από όσο έπρεπε άρα διόρθσα τις αντίστοιχες γραμμές. Το ίδιο ίσχυε για μία γραμμή με year = Gallimard και έκανα τα ίδια. Συνεχίζοντας, διόρθωσα τις τιμές των years που υπάρχουν και είναι άκυρες, δηλαδή τις τιμές πάνω από το 2005 μιας και το dataset δημιουργήθηκε τότε και το θεωρούμε ως άνω όριο ενώ και τα βιβλιά που έχουν year = 0 και τα αντικατέστησα με τον μέσο όλων των υπόλοιπων years. Έπειτα κοίταξα για ποιές γραμμές δεν έχουν τιμή στο bookAuthor και βρήκα οτι ήταν μία μόνο οπότε έψαξα ποιός είναι ο συγγραφέας και το αντικατέστησα. Για την εξαγωγή keywords από τα bookTitles χρησιμοποιήσα την βιβλιοθήκη rake\_nltk και πιο συγκεκριμένα το Rake.

# Αποτελέσματα πειράματος.

Για τα 10 αρχεία των 5 χρηστών, η ομοιότητα μεταξύ των λιστών των 4 από τους 5 χρήστες είναι 100% μιας και είναι **ακριβώς ίδιες** και επομένως και η golden standard λίστα αποτελεσμάτων είναι η ίδια άρα και για το ερώτημα 3 μόνο για τον χρήστη 192705 θα υπολογίσουμε.Μόνο για τον χρήστη με userID 192705 δεν ισχύει αυτό άρα θα υπολογίσουμε τις ομοιότητες. Βλέποντας το excel θα πάρουμε αρχικά το 1ο αποτέλεσμα της 1ης λίστας με το 1ο της άλλης λίστας και θα βρούμε την ομοιότητα, έπειτα το 1ο **και**  το 2ο της 1ης λίστας με το 1ο και το 2ο της άλλης λίστας και θα βρούμε την ομοιότητα κ.ο.κ μέχρι να έχουμε και τα 10 με τα άλλα 10.

* 1 με 1 : Ομοιότητα = 1/1 = 1
* 1,2 με 1,2 : Ομοιότητα = 2/2 = 1
* 1,2,3 με 1,2,3 : Ομοιότητα = 2/3 = 0.67
* 1,2,3,4 με 1,2,3,4 : Ομοιότητα = 2/4 = 0.5
* 1,2,3,4,5 με 1,2,3,4,5 : Ομοιότητα = 2/5 = 0.4
* ενώ σε όλα τα παρακάτω από την 6η μέχρι την 10η ταινία είναι ίδιες άρα έχουμε 3/6 = 0.5, 4/7 = 0.57, 5/8=0.625, 6/9 = 0/67, 7/10 = 0.7.

M.O. = (1+1+0.67+0.5+0.4+0.5+0.57+0.625+0.67+0.7) / 10 = 6.635 / 10 = 0.663 ή 66.3% ομοιότητα μεταξύ των δύο λιστών. Όσον αφορά το ερώτημα 3 παρατηρούμε οτι και οι δύο λίστες προτείνουν τα ίδια βιβλία άρα δεν δεν μπορούμε να φτιάξουμε την νέα λίστα βάσει το πλήθος εμφανίσεων και θα κοιτάξουμε την ομοιότητα με το προφιλ του χρήστη.

# Σχολιασμός και άλλες προσεγγίσεις.