Text

Description automatically generated

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΙΩΑΝΝΗΣ ΚΩΝΣΤΑΝΤΑΚΟΣ

mpked2215

ΓΕΩΡΓΙΟΣ ΛΕΒΑΝΤΗΣ

mpked2216

ΑΝΑΛΥΤΙΚΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Το dataset που επιλέχτηκε ώστε να κάνουμε Προπαρασκευή Δεδομένων, Συσταδοποίηση και Σύστημα Προτεινόμενων Ταινιών σε γλώσσα Python είναι το MovieLens dataset το οποίο περιέχει δεδομένα για 9.742 ταινίες στο σύνολο.

Αποτελείται από 4 αρχεία δεδομένων (links, tags, ratings και movies) ή αλλιώς tables και περιέχονται εξής δεδομένα:

Links: Περιέχει τρείς στήλες με ID των ταινιών. Η πρώτη αναφέρεται στις ταινίες που έχουμε στο δικό μας dataset όπου στο movies.csv μπορούμε να αντιστοιχίσουμε τον τίτλο της κάθε ταινίας με το ID. Οι άλλες δύο στήλες αναφέρονται στους ίδιους τίτλους ταινιών από άλλα datasets (IMDB, MovieDB).

Tags: Περιέχει 3.683 tags τα οποία έχουν γίνει από 58 χρήστες και αναφέρονται σε 1.572 ταινίες (16% του συνόλου των ταινιών), ένα ποσοστό το οποίο μας προδιαθέτει αρνητικά για την αξιοποίηση του συγκεκριμένου αρχείου. Επίσης περιέχει και τα timestamps, δηλαδή η ακριβής ημερομηνία και ώρα που πραγματοποιήθηκαν τα συγκεκριμένα tags με τα οποία όμως δε θα ασχοληθούμε καθόλου σε αυτό το project.

Ratings: Περιέχει 100.836 κριτικές οι οποίες έχουν γίνει από 610 διαφορετικούς χρήστες και αναφέρονται σε 9724 ταινίες. Επομένως έχουμε 18 ταινίες οι οποίες δεν έχουν λάβει κάποια κριτική. Το εύρος των κριτικών είναι από 0.5 – 5.0 (ανά 0.5) και η συντριπτική πλειοψηφία των κριτικών παρουσιάζεται μεταξύ 3 – 5 (81%).

Movies: Υπάρχουν 9.742 τίτλοι ταινιών με τα αντίστοιχα ID τους καθώς και ένα πεδίο με την κατηγορία της ταινίας, η οποία μπορεί να περιέχει και παραπάνω από μία κατηγορία (πχ. *Toy Story (1995) - Adventure| Animation| Children| Comedy| Fantasy).* Υπάρχουν συνολικά 951 διαφορετικοί συνδυασμοί κατηγοριών, 20 μοναδικές κατηγορίες από τις οποίες μια από αυτές είναι η *«no genre listed»* που δηλώνει πως δεν έχει καταχωρηθεί σε κάποια κατηγορία ο τίτλος της ταινίας. Σε αυτή την τελευταία κατηγορία υπάρχουν συνολικά 34 ταινίες, λίγες σε σχέση με το σύνολο.

Αρχικά, εγκαταστήσαμε τις βιβλιοθήκες που χρησιμοποιήσαμε για την υλοποίηση και φορτώσαμε τα αρχεία και τα αποθηκεύσαμε με τη βοήθεια της βιβλιοθήκης pandas σε DataFrames.

Στη συνέχεια, κάναμε περιγραφική στατιστική, χρησιμοποιώντας εργαλεία οπτικοποίησης (seaborn, matplotlib), για να γνωρίσουμε καλύτερα τα δεδομένα μας και προχωρήσαμε στον καθαρισμό τους αφαιρούνται διπλότυπες και κενές τιμές και αφού έτσι διασφαλίσαμε ότι τα δεδομένα είναι ακριβή και έτοιμα για ανάλυση ενώσαμε τα tables ώστε να προχωρήσουμε.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Σειρά είχε ο μετασχηματισμός του συνόλου δεδομένων μας, δηλαδή η διαδικασία τροποποίησης των δεδομένων με κάποιο τρόπο πριν αναλυθούν ή χρησιμοποιηθούν για μηχανική μάθηση. Συγκεκριμένα, αφαιρέσαμε κάποια πεδία και γραμμές, όπως για παράδειγμα ταινίες με λιγότερες από 5 βαθμολογήσεις, δημιουργήσαμε νέα χαρακτηριστικά, ένα από αυτά ήταν να μετασχηματίσουμε ήδη υπάρχοντα πεδία (Adventure|Animation|Children|Comedy|Fantasy σε Adventure, δηλαδή κρατήσαμε μόνο την πρώτη κατηγορία) και να τα κωδικοποιήσουμε ως αριθμητική τιμή με one-hot encoding ώστε να μπορούν να χρησιμοποιηθούν σε αλγόριθμους μηχανικής εκμάθησης.

Chart, scatter chart

Description automatically generated

Η Συσταδοποίηση (Clustering) που ακολούθησε είναι μια τεχνική μηχανικής εκμάθησης χωρίς επίβλεψη που περιλαμβάνει τη διαίρεση ενός συνόλου δεδομένων σε ομάδες (clusters) με βάση την ομοιότητα των δεδομένων σε κάθε ομάδα. Δοκιμάσαμε τρία διαφορετικά μοντέλα χρησιμοποιώντας τους αλγορίθμους k-means και DBSCAN σε κάθε ένα και καταλήξαμε στο ότι καλύτερη και ευδιάκριτη συσταδοποίηση πραγματοποιείται με την εφαρμογή του k-means για πέντε clusters στο μοντέλο που λαμβάνει υπόψη τον Μέσο Όρο Βαθμολογίας (AVGrate) κάθε ταινίας, τον Αριθμό Βαθμολογιών κάθε ταινίας (CNTrate) και το είδος της ταινίας το οποίο έχουμε κωδικοποιήσει.

[[ 3.25928802e+00 1.00344130e+01 7.69230769e-02 2.34817814e-02

2.79352227e-02 2.87044534e-01 6.03238866e-02 1.70040486e-02

2.18218623e-01 3.23886640e-03 1.21457490e-03 4.12955466e-02

2.02429150e-03 5.26315789e-03 2.83400810e-03 2.02429150e-03

8.50202429e-03 4.04858300e-04 2.42914980e-03]

[ 3.74771551e+00 1.14328244e+02 1.45038168e-01 1.52671756e-02

3.05343511e-02 2.13740458e-01 8.39694656e-02 -1.73472348e-18

1.14503817e-01 8.67361738e-19 -4.33680869e-19 1.52671756e-02

4.33680869e-19 7.63358779e-03 1.08420217e-18 -8.67361738e-19

3.46944695e-18 0.00000000e+00 0.00000000e+00]

[ 3.56764111e+00 6.42422145e+01 8.30449827e-02 3.80622837e-02

1.73010381e-02 2.80276817e-01 6.22837370e-02 3.46020761e-03

1.45328720e-01 6.92041522e-03 -1.30104261e-18 6.92041522e-03

3.46020761e-03 6.92041522e-03 2.38524478e-18 3.46020761e-03

8.67361738e-18 1.08420217e-19 8.67361738e-19]

[ 3.46199718e+00 3.28678720e+01 9.73574409e-02 2.92072323e-02

1.94714882e-02 2.68428373e-01 4.86787204e-02 4.17246175e-03

1.80806676e-01 2.78164117e-03 2.78164117e-03 2.36439499e-02

-5.85469173e-18 2.78164117e-03 -4.98732999e-18 5.56328234e-03

6.95410292e-03 -3.79470760e-19 2.78164117e-03]

[ 3.98070610e+00 2.07675000e+02 2.25000000e-01 -3.46944695e-18

-3.46944695e-18 7.50000000e-02 1.00000000e-01 -1.73472348e-18

7.50000000e-02 4.33680869e-19 2.16840434e-19 1.38777878e-17

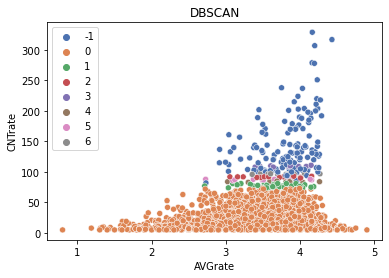
2.16840434e-19 5.00000000e-02 -8.67361738e-19 4.33680869e-19

2.50000000e-02 -1.62630326e-19 -1.30104261e-18]]

Chart, scatter chart

Description automatically generated

Αν και κάναμε αρκετές δοκιμές στον DBSCAN βλέπουμε ότι ο k-means λειτουργεί καλύτερα. Ο DBSCAN δημιουργεί μία μεγάλη κλάση με τα δεδομένα που έχουν μικρό αριθμό βαθμολογιών, ενώ αν μειώσουμε τον απαιτούμενο αριθμό samples μιας κλάσης τότε τείνει να κάνει πολλές αχρείαστες κλάσεις με μικρές διαφορές μεταξύ τους.



Ύστερα από έρευνα σε αντίστοιχες προσεγγίσεις στο διαδίκτυο καταλήξαμε στο γεγονός ότι το σύνολο δεδομένων μας είναι αρκετά μικρό και ποιοτικά και ποσοτικά, δηλαδή χρειαζόμαστε περισσότερα στατιστικά σημαντικά πεδία όπως η ηλικία και το φύλο για τους users και ιδανικά θα θέλαμε περισσότερα πεδία για το movies dataset όπως βραβεία, διάρκεια και αριθμό ηθοποιών όπως επίσης και μεγαλύτερο σύνολο δεδομένων. Επομένως προχωρήσαμε στην εξαγωγή ενός απλού Recommendation System αξιοποιώντας τα δεδομένα μας.

Συγκεκριμένα φτιάξαμε ένα σύστημα το οποίο αρχικοποιεί ένα τίτλο μίας ταινίας και με βάση τον αριθμό των βαθμολογιών και τον μέσο όρο κριτικών της, προτείνει στον χρήστη 15 ταινίες με παρόμοια χαρακτηριστικά.

title

Forrest Gump (1994) 1.000000 329

Bio-Dome (1996) 0.723238 31

Thor: The Dark World (2013) 0.715809 21

Fatal Attraction (1987) 0.701856 36

X-Men: Days of Future Past (2014) 0.682284 30

National Lampoon's Vacation (1983) 0.677043 26

Easy A (2010) 0.670081 27

Bridesmaids (2011) 0.663176 21

Cabin in the Woods, The (2012) 0.653015 22

Patch Adams (1998) 0.652302 27

Mr. Holland's Opus (1995) 0.652144 80

Pirates of the Caribbean: On Stranger Tides (2011) 0.645193 24

To Wong Foo, Thanks for Everything! Julie Newma... 0.641189 22

Waterboy, The (1998) 0.632631 28

Grumpy Old Men (1993) 0.630896 29