# Τεχνικές Εξόρυξης Δεδομένων
# Εαρινό Εξάμηνο 2020-2021
# 1η Άσκηση
### Ατομική ή ομαδική Εργασία (2 Ατόμων)
### Παράδοση: 21/04/2021

Σας ζητείται να επιστρατεύσετε τις ικανότητές σας στην επεξεργασία δεδομένων για να βοηθήσετε σε μία μελέτη που γίνεται για τους τίτλους στη γνωστή πλατφόρμα ψυχαγωγίας Netflix καθώς και να βοηθήσετε στη δημιουργία ενός συστήματος που θα προβάλει προτινόμενους τίτλους ταινιών (recommendation system). Θα χρησιμοποιήσετε python αξιοποιώντας και τις τεχνικές που δείξαμε στα φροντιστήρια. Στο eclass στον σχετικό φάκελο της εργασίας θα βρείτε ένα συμπιεσμένο αρχείο που περιέχει τα αρχεία με τα δεδομένα της άσκησης. Τα δεδομένα που σας δίνονται αποτελούνται από 3 csv αρχεία (*netflix_titles, imdb_movies, imdb_ratings*).
Θα ασχοληθείτε κυρίως με το αρχείο *netflix_titles* το οποίο περιέχει 12 στήλες και αρχικά θα χρειαστεί να μελετησετε αν υπάρχουν missing data. Αποφασίστε πως θα τα χειριστείτε και απαλείψτε τα ή συμπληρώστε τα κατάλληλα. Τα γραφήματα που ζητούνται στα ερωτήματα μπορούν να γίνουν με οποιαδήποτε βιβλιοθήκη της επιλογής σας.

## Ερωτήματα (60 μονάδες)

1. ~~Ποιό είδος υπερτερεί, οι ταινίες ή οι σειρές ;~~ **~~(5 μονάδες)~~**
2. ~~Τα τελευταία χρόνια το~~ netflix ~~επενδύει περισσότερο σε ταινίες ή σε σειρές;~~ **~~(5 μονάδες)~~**
3. ~~Ποιά χώρα έχει το περισσότερο περιεχόμενο;~~ **~~(5 μονάδες)~~**
4. ~~Τι είδους περιεχόμενο έχει κάθε χώρα;~~ **~~(5 μονάδες)~~**
5. ~~Ετοιμάστε γραφήματα που δειχνουν τους ηθοποιούς με τις περισσότερες ταινίες σε κάθε χώρα. Κάντε το ίδιο και για τις σειρές~~ **~~(5 μονάδες)~~**
6. Το netflix υποστηρίζει ότι παρέχει πλούσιο περιεχόμενο για όλες τις ηλικίες. Αληθεύει αυτό; Φτιάξτε ένα γράφημα το οποίο συγκεντρώνει το πλήθος των ταινιων ανάλογα με την προτεινόμενη ηλικία (αναφερόμαστε στη στήλη rating). Τα όρια καθορίζονται σύμφωνα με τον παρακάτω πίνακα. Κάντε το ίδιο γράφημα και για τις σειρές. **(5 μονάδες)**

| Little Kids | Older Kids | Teens | Mature |
|---|---|---|---|
| G, TV-Y, TV-G | PG, TV-Y7, TV-Y7-FV, TV-PG | PG-13, TV-14 | R, NC-17, TV-MA |

7. Αν ένας παραγωγός ήθελε να έχει υψηλή ακροαματικότητα, σκέφτεται ότι θα ήταν ίσως καλύτερα να βγάλει την ταινία του σε μία εποχή που δεν υπάρχει ανταγωνισμός. Κάντε ένα γράφημα με το περιεχόμενο που προστίθεται ανά μήνα για να τον βοηθήσετε να επιλέξει τη σωστή χρονική περίοδο. **(5 μονάδες)**
8. ~~Ετοιμάστε ένα γράφημα που παρουσιάζει συγκεντρωτικά τα είδη του περιεχομένου (αναφερόμαστε στη στήλη listed_in). **(5 μονάδες)**~~
9. ~~Μελετηστε τους σκηνοθέτες ανά χώρα και παρουσιάστε σχετικά γραφήματα. **(5 μονάδες)**~~
10. ~~Μελετήστε τις σειρές και παρουσιάστε ένα γράφημα που τις δείχνει ανάλογα με το αριθμό των seasons. **(5 μονάδες)**~~
11. ~~Αξιοποιήστε τα υπόλοιπα αρχεία της εργασίας για να απαντήσετε στο ερώτημα: Ποιές είναι οι ταινίες με την πιο υψηλή βαθμολογία; Θα χρειαστεί να συννενώσετε κατάλληλα τα αρχεία που σας δίνονται ώστε να κρατήσετε τις ταινίες του netflix για τις οποίες υπάρχουν ratings στο IMBD. **(10 μονάδες)**~~

**Recommendation system (40 μονάδες)**

Σκοπός είναι να εξάγετε χρήσιμη πληροφορία από τα δεδομένα και να προσπαθήσετε να φτιάξετε ένα πρόγραμμα το οποίο θα παράγει προτάσεις (recommendations) για το περιεχόμενο του netflix. Θα αξιοποιήσετε τις στήλες *show_id, title, description*.

1. Θα δημιουργήσετε 2 διαφορετικά representations για κάθε ταινία, βασιζόμενοι στις περιγραφές που περιέχονται στις παραπάνω στήλες:
    a. Δημιουργήστε το **boolean BoW** πίνακα των unigrams και των bigrams από τη κειμενική αναπαράσταση των ταινιών. (χρησιμοποιήστε την παράμετρο ngram_range του CountVectorizer).
    b. Δημιουργήστε τον **TF-IDF** (Term Frequency - Inverse Document Frequency) πίνακα των unigrams και των bigrams από τη κειμενική αναπαράσταση των ταινιών.

Πειραματιστείτε με τις διάφορες παραμέτρους των vectorizers π.χ. max_df, min_df, max_features, stopwords, για να πάρετε χρήσιμες αναπαραστάσεις.

2. Χρησιμοποιώντας τις παραπάνω αναπαραστάσεις των ταινιών υπολογίστε την ομοιότητα μεταξύ τους.
    a. Για την αναπαράσταση από το 1.α, θα αξιοποιήσετε το **Jaccard/Tanimoto coefficient** για να υπολογίσετε την ομοιότητα 2 ταινιών δοθέντων των feature vectors τους.
    b. Για την αναπαράσταση από το 1.β, θα αξιοποιήσετε το **cosine similarity**.

Διατρέξτε τους πίνακες από το 1 και υπολογίστε τις ομοιότητες όλων των ταινιών μεταξύ τους (και με τους δύο τρόπους 2.α, 2.β). Έπειτα αποθηκεύστε για κάθε ταινία τις 100 πιο όμοιες με αυτή σε ένα python dictionary.

3. Πρόβλεψη (1): Φτιάξτε μία συνάρτηση η οποία παίρνει σαν είσοδο ένα τίτλο, ένα ακέραιο αριθμό N και τον τρόπο εύρεσης της ομοιότητας (boolean ή tf-idf σύμφωνα με τα 1,2) και επιστρέφει τους N πιο όμοιους τίτλους.

```
get_similar_movies1('Title_of_the_movie', N=10, method='boolean')
```

Παρουσιάστε ενδεικτικά αποτελέσματα από την παραπάνω συνάρτηση και σχολιάστε για τις ποιοτικές διαφορές με τους δύο τρόπους υπολογισμού της ομοιότητας των ταινιών.4.

4. Πρόβλεψη (2): Φτιάξτε μία συνάρτηση η οποία παίρνει σαν είσοδο μια ακολουθία από λέξεις (π.χ. Περιγραφή μιας ταινίας), ένα ακέραιο αριθμό N και τον τρόπο εύρεσης της ομοιότητας και επιστρέφει τους N πιο όμοιους τίτλους με βάση αυτή τη περιγραφή. Το συγκεκριμένο σύστημα θα μετασχηματίζει τη περιγραφή της εισόδου σε ένα feature vector, ανάλογα με την επιλεγμένη μεθοδολογία, και χρησιμοποιώντας τους κατάλληλους πίνακες και μετρικά από τα 1,2 θα επιστρέφει τις πιο σχετικές ταινίες.

```
get_similar_movies2('War between America and Vietnam', N=10,

method='tf-idf')
```

Παρουσιάστε κάποια καλά ενδεικτικά αποτελέσματα από την παραπάνω συνάρτηση και αναφέρετε πιθανούς λόγους αστοχίας αυτών.

## Παραδοτέο:

Η εργασία μπορεί να εκπονηθεί ατομικά ή σε ομάδες 2 ατόμων. Θα ανεβάσετε στο eclass ένα φάκελο της μορφής sdixxxx.zip (όπου sdi το ΑΜ ενός εκ των ατόμων της ομάδας) ο οποίος θα περιέχει μόνο τον κώδικά σας σε μορφή Ipython notebook (**προσοχή**: δεν χρειάζεται να ανεβάσετε τα αρχεία *.csv).
Το notebook πρέπει να έχει "τρέξει" ώστε να φαίνονται τα αποτελέσματα της εργασίας σας. Το notebook αποτελεί και την ολοκληρωμένη αναφορά για την εργασία σας (**δεν θα παραδώσετε τίποτα σε doc, pdf**) , σχεδιάστε το με προσοχή, να θυμάστε να γράψετε μία περιγραφή σε κάθε βήμα για το τι κάνει ο κώδικάς σας σε κάθε κελί.