

**Αναγνώριση Προτύπων**

*Εργασία 2018-2019*

Π16036 – Ιωαννίδης Παναγιώτης

Π16112 – Παραβάντης Αθανάσιος

Περιεχόμενα

[1 Εισαγωγή 2](#_Toc876245)

[2 Ερώτημα 1 3](#_Toc876246)

[3 Ερώτημα 2 4](#_Toc876247)

[3.1 Ζητούμενο 4](#_Toc876248)

[3.2 Τεκμηρίωση 4](#_Toc876249)

[3.2.1 Βασικό ακολουθιακό αλγοριθμικό σχήμα (BSAS) 4](#_Toc876250)

[3.2.2 Εκτίμηση αριθμού ομάδων 4](#_Toc876251)

[3.2.3 Εκτέλεση 4](#_Toc876252)

# Εισαγωγή

Η παρούσα εργασία για το μάθημα της Αναγνώρισης Προτύπων βασίζεται στο σύνολο δεδομένων [MovieLens 100K Dataset](https://grouplens.org/datasets/movielens/100k/) που είναι διαθέσιμο στην ιστοσελίδα του Grouplens. Για τη επίλυση όλων των ερωτημάτων της εργασίας, χρησιμοποιήσαμε τη Python με την έκδοση 3.7.2, καθώς και αρκετές βιβλιοθήκες ανάλυσης δεδομένων και μηχανικής μάθησης.

Η λίστα με τις βιβλιοθήκες της Python που χρησιμοποιούνται στην εργασία είναι οι εξής:

* [Pandas](https://pandas.pydata.org/)  
  Βιβλιοθήκη που παρέχει δομές και εργαλεία για την ανάλυση δεδομένων.
* [Scikit Learn](https://scikit-learn.org/)  
  Βιβλιοθήκη μηχανικής μάθησης αλλά και ανάλυσης δεδομένων.
* [Scipy](https://www.scipy.org/)  
  Σύνολο βιβλιοθηκών που παρέχει βασικές δομές για τη κάλυψη μαθηματικών υπολογισμών.
* [Numpy](https://www.numpy.org/)  
  Βιβλιοθήκη για επιστημονικούς υπολογισμούς, κυρίως πινάκων.
* [Matplotlib](https://matplotlib.org/)  
  Βιβλιοθήκη για την προβολή δεδομένων με γραφικές παραστάσεις.

Δόθηκε απάντηση για όλα τα ερωτήματα της εργασίας.

# Ερώτημα 1

Αφού διαβάσαμε το αρχείο README και κατανοήσαμε τη δομή των δεδομένων αποφασίσαμε να δημιουργήσουμε τη κλάση **MovieLensData** στο αρχείο **movielens\_data.py** ώστε να διαχειριζόμαστε τα αρχεία αυτά και να λαμβάνουμε, κάθε φορά, τα δεδομένα που μας είναι απαραίτητα.

Συγκεκριμένα, γίνεται η εξής επεξεργασία και φόρτωση δεδομένων:

* Σαρώνονται όλα τα περιεχόμενα του αρχείου u.data με τη βοήθεια της συνάρτηση read\_csv του pandas και αποθηκεύονται στη μνήμη οι βαθμολογίες χρηστών που είναι μεγαλύτερες από 4.
* Σαρώνονται τα περιεχόμενα του αρχείου u.item ώστε να τοποθετήσουμε στη μνήμη πληροφορίες για κάθε ταινία.
* Ενοποίηση των βαθμολογιών μαζί με τις πληροφορίες για κάθε ταινία, με βάση το user id των δεδομένων.
* Εύρεση του αθροίσματος ταινιών ανά κατηγορία στις οποίες ο χρήστης έχει βάλει βαθμό.
* Κανονικοποίηση των δεδομένων με βάση τον μέσο όρο του αθροίσματος ταινιών ανά κατηγορία για κάθε στήλη.

Η παραπάνω προεπεξεργασία και φόρτωση δεδομένων είναι πολύτιμη επειδή χρησιμοποιείται ως είσοδο για τα επόμενα ερωτήματα της εργασίας. Οποιαδήποτε σάρωση και αναπαράσταση δεδομένων γίνεται με τη βοήθεια των δομών της βιβλιοθήκης pandas.

# Ερώτημα 2

## Ζητούμενο

Εφαρμόστε το βασικό σχήμα ακολουθιακής ομαδοποίησης για να εκτιμήσετε το  
πλήθος των ομάδων των χρηστών, ως προς τις προτιμήσεις τους.

## Τεκμηρίωση

### Βασικό ακολουθιακό αλγοριθμικό σχήμα (BSAS)

Στο ερώτημα αυτό, υλοποιήσαμε το βασικό ακολουθιακό αλγοριθμικό σχήμα (BSAS) όπως και τον αλγόριθμο εκτίμησης αριθμού ομάδων.

Στην περίπτωση μας ο αριθμός των ομάδων δεν είναι γνωστός. Το ζητούμενο μας είναι να βρούμε τον αριθμό αυτόν. Εφόσον δεν γνωρίζουμε τον αριθμό των ομάδων, η παράμετρος η οποία εμπλέκεται στον αλγόριθμο και η οποία ορίζεται από τον χρήστη, είναι το κατώφλι ανομοιότητας Θ.

Κάθε νέο διάνυσμα εξετάζεται από τον αλγόριθμο και καταχωρείται είτε σε μια από τις υπάρχουσες ομάδες είτε σε μια νέα ομάδα που δημιουργείται, ανάλογα με την απόσταση από τις άλλες υπάρχουσες ομάδες.

Για τον υπολογισμό της απόστασης θα χρησιμοποιήσουμε το τετράγωνο της ευκλείδειας απόστασης.

Η διάταξη με την οποία παρουσιάζονται τα διανύσματα στον BSAS παίζει σημαντικό ρόλο στη διαδικασία ομαδοποίησης, καθώς μπορεί να οδηγήσουν σε διαφορετικές ομαδοποιήσεις. Επίσης σημαντικός παράγοντας είναι και το Θ.

Κάθε ομάδα αντιπροσωπεύεται από ένα διάνυσμα. Ο αντιπρόσωπος τις ομάδας μετρά από κάθε προσθήκη ενός νέου διανύσματος στην ομάδα θα πρέπει να ενημερώνεται.

### Εκτίμηση αριθμού ομάδων

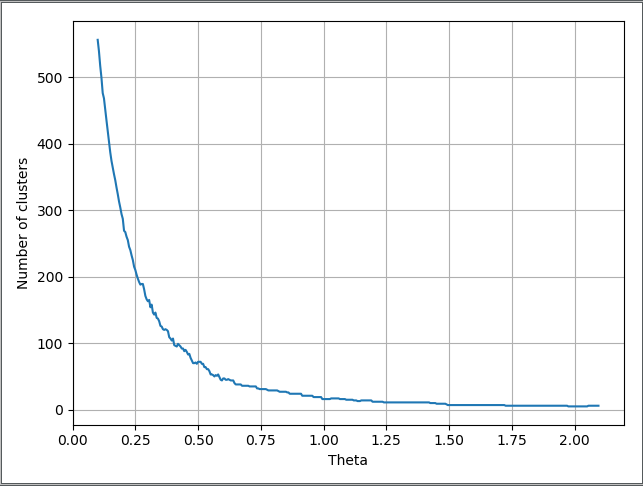
Για να μπορέσουμε να εκτιμήσουμε τον αριθμό των ομάδων, αρκεί να εκτελέσουμε επαναληπτικά τον αλγόριθμο BSAS, για Θ από a έως b με βήμα c. Όσο περισσότερες φόρες εκτελέσουμε τον αλγόριθμο BSAS τόσο μεγαλύτερη θα είναι και η ακρίβεια των αποτελεσμάτων.

### Εκτέλεση

Ο αλγόριθμος BSAS και αλγόριθμος εκτίμησης αριθμού ομάδων βρίσκονται στο αρχείο bsas.py. Ως ελάχιστο Θ δίνουμε την τιμή 0.1. ως μέγιστο Θ την τιμή 2.1 και ως βήμα την τιμή 0.005. Τα δεδομένα που δίνονται ως είσοδος στο αλγόριθμο είναι το άθροισμα των κατηγοριών των ταινιών με βαθμολογία από 4 και πάνω. Στη συνέχεια κανονικοποιούμε τα διανύσματα που προκύπτουν ώστε οι τιμές να είναι μεταξύ 0 και 1.

Αρκεί να εκτελέσουμε το αρχείο bsas.py και στο αρχείο result.txt που θα δημιουργηθεί θα δούμε τον αριθμό των ομάδων που δημιουργούνται για κάθε Θ.

Μετά την εκτέλεση του αλγόριθμου προκύπτει η ακόλουθη γραφική απεικόνιση του αριθμού των ομάδων ως προς Θ, εικόνα 3.1.



Εικόνα 3.1

Με βάση τις παραμέτρους που έχουμε θέσει ο αριθμός των ομάδων είναι n = 7.

# Ερώτημα 3

## Ζητούμενο

Στο 3ο ερώτημα ζητείται η εφαρμογή του αλγόριθμου Κ-Μέσων και του αλγορίθμου ιεραρχικής ομαδοποίησης, πάνω στο dataset της εργασίας.

## Αλγόριθμος Κ-Μέσων

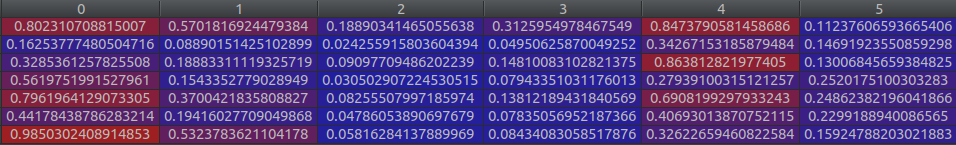
Ο αλγόριθμος Κ-Μέσων υλοποιείται από τη βιβλιοθήκη SciKit Learn, χρησιμοποιώντας τη κλάση [KMeans](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html). Το σύνολο δεδομένων, η προεπεξεργασία και η κανονικοποίησή τους γίνεται με τη βοήθεια της κλάσης MovieLensData.

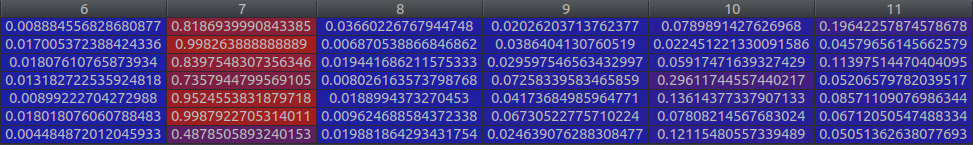
Επιλογές παραμέτρων για τον αλγόριθμο:

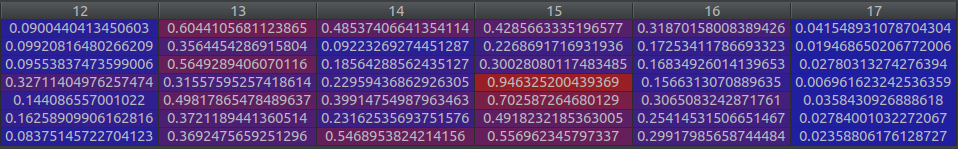
* Ο αριθμός των ομάδων και των κεντροειδών επιλέξαμε να είναι 7.
* Ο αριθμός των επαναλήψεων επιλέξαμε να είναι 100.

Τα δεδομένα που εισάγονται στον αλγόριθμο είναι 17 διαστάσεων και αναπαριστούν το άθροισμα των βαθμολογιών ανά κατηγορία για κάθε χρήστη. Οι διαστάσεις αυτές προκύπτουν από τις 17 κατηγορίες ταινιών στο σύνολο δεδομένων που μας παρέχεται. Εκτελώντας το αρχείο kmeans.py λαμβάνουμε ενδεικτικά αποτελέσματα για τα κεντροειδή καθώς και την ομαδοποίηση που προκύπτει.

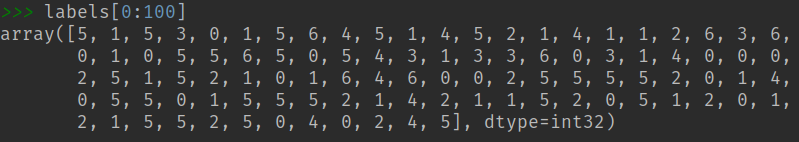
Παρακάτω εμφανίζονται οι συντεταγμένες των 7 κεντροειδών με (x1, x2, …, x17):







Επιπλέον, παρατηρούμε την ομαδοποίηση που γίνεται για μερικά στοιχεία του πίνακα:



Ο αριθμός που προκύπτει από το 0 έως το 6 βασίζεται στην αρχική επιλογή των 7 ομάδων που ορίσαμε ως παράμετρο στη κλάση της βιβλιοθήκης. Τα αποτελέσματα εμφανίζονται έπειτα από 100 επαναλήψεις του αλγορίθμου.

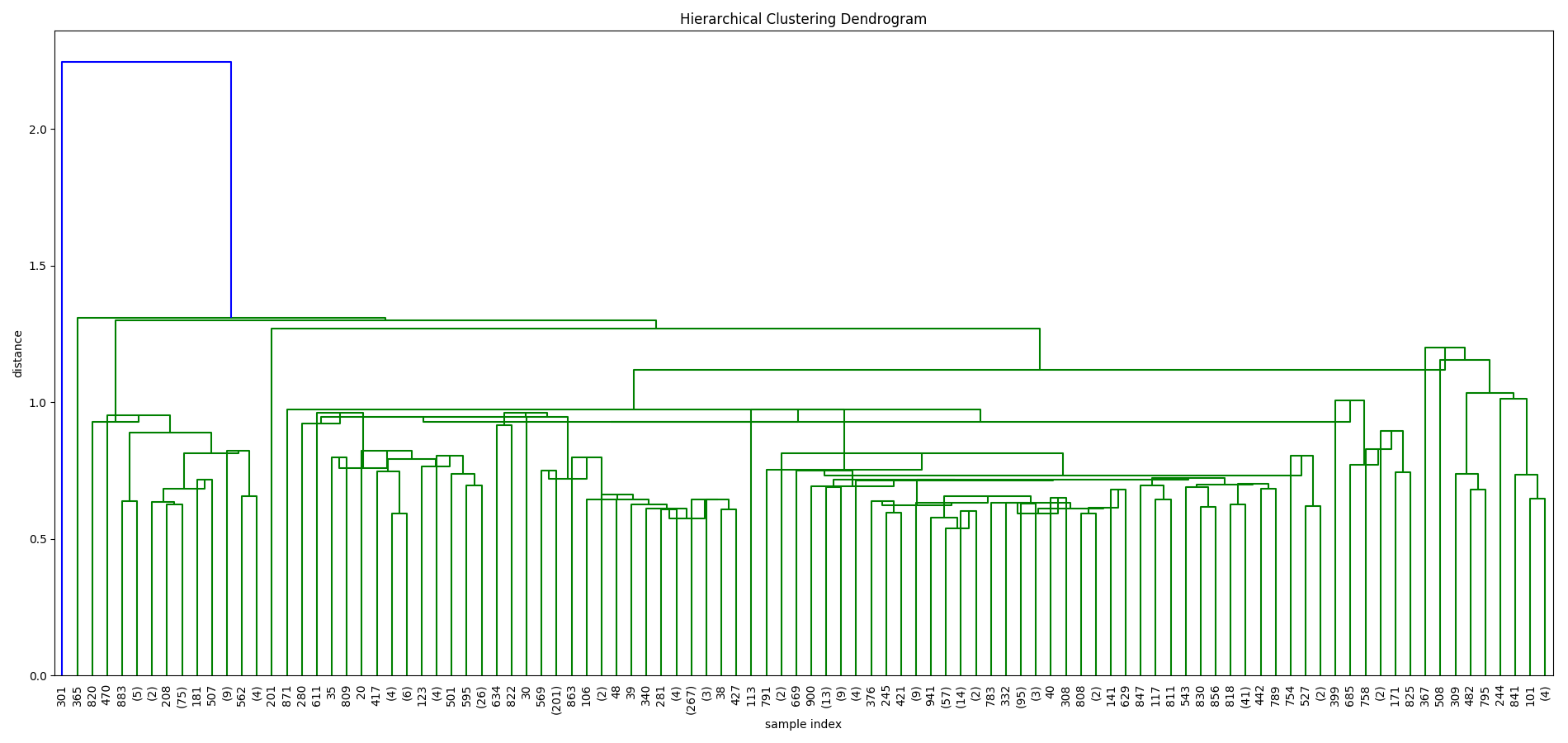
## Αλγόριθμος Ιεραρχικής Ομαδοποίησης

Ο αλγόριθμος ιεραρχικής ομαδοποίησης υλοποιείται από τη βιβλιοθήκη Scipy, χρησιμοποιώντας τη συνάρτηση [linkage](https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.cluster.hierarchy.linkage.html). Το σύνολο δεδομένων, η προεπεξεργασία και η κανονικοποίησή τους γίνεται με τη βοήθεια της κλάσης MovieLensData.

Όπως και προηγουμένως, τα δεδομένα μας είναι σε 17 διαστάσεις λόγω του πλήθους των κατηγοριών για τις ταινίες. Οι παράμετροι του αλγορίθμου που καθορίσαμε είναι οι εξής:

* Επιλέγουμε τη μέθοδο ‘median’ για την ενοποίηση των κεντροειδών. Αυτό σημαίνει ότι ο αλγόριθμος υπολογίζει τη μέση τιμή μεταξύ των κεντροειδών δυο ομάδων.
* Επιλέγουμε την ευκλείδεια απόσταση για τον καθορισμό του δείκτη ομοιότητας μεταξύ των δεδομένων μας.

Με την εκτέλεση του αρχείου hierarchical.py παίρνουμε το ακόλουθο δενδρόγραμμα που αναπαριστά τη διαδικασία του αλγορίθμου ενοποιώντας τις ομάδες:



## Συμπέρασμα

Συμπεραίνοντας, έχοντας υπ’ όψη τους δυο αλγόριθμους ομαδοποίησης Κ-Μέσων και Ιεραρχικής Ομαδοποίησης παρατηρούμε ότι ενώ χρησιμοποιούνται για τον ίδιο σκοπό, το αποτέλεσμα είναι αρκετά διαφορετικής σημασίας.

Στον αλγόριθμο Κ-Μέσων το αποτέλεσμα των επαναλήψεων είναι η ομαδοποίηση όλων των δεδομένων σε μια από τις κατηγορίες που ορίζουμε. Όταν όμως μελετάμε την ιεραρχική ομαδοποίηση, παρατηρούμε ότι το δενδρόγραμμα παρέχει μια παραπάνω πληροφορία ιεραρχίας μεταξύ των δεδομένων. Αυτό το στοιχείο μπορεί να χρησιμοποιηθεί για ταξινόμηση δεδομένων βιολογικής σημασίας, όμως στη περίπτωσή μας το σύνολο δεδομένων είναι αρκετά διαφορετικό.