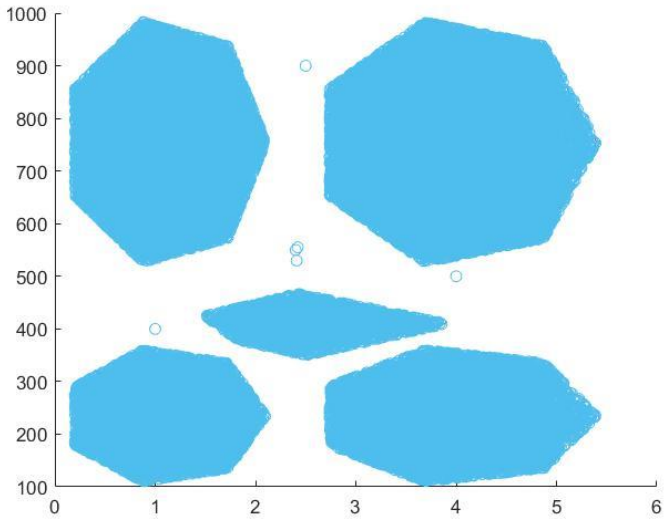
Εργασία για το Μάθημα Αποθήκες και Εξόρυξη Δεδομένων 2023-2024
(2 μονάδες, ομαδική 2 ατόμων)

Γενικές Οδηγίες

Η άσκηση έχει ως βασικό στόχο να βρει συστάδες και ανωμαλίες σε δισδιάστατα σημεία χρησιμοποιώντας τον k-means ως αλγόριθμο στον οποίο θα βασιστεί ο εντοπισμός ανωμαλιών και το Spark ως υπολογιστικό πλαίσιο. Η άσκηση περιλαμβάνει τα εξής βήματα:

1. Λήψη του συνόλου δεδομένων σε μορφή csv, γνωρίζοντας ότι υπάρχουν 5 συστάδες και κάποια outliers (ακρότατα).
2. Καθαρισμός εγγραφών που δεν έχουν και τα δύο πεδία.
3. Μετασχηματιμός 0-1 και στις δύο διαστάσεις.
4. Εφαρμογή του k-means σε περιβάλλον Spark, χρησιμοποιώντας τις έτοιμες συναρτήσεις για τον K-means από το Spark.
5. Εντοπισμός ανωμαλιών βάσει των αποτελεσμάτων του βήματος 4 με κλιμακώσιμο τρόπο.
6. Εντοπισμός συστάδων εκτός Spark ή κριτική αξιολόγηση οδηγιών από chatGPT (αρκεί να κάνετε ένα από τα 2).



Το πρόγραμμα θα δέχεται ως μοναδική παράμετρο εισόδου το όνομα του αρχείου δεδομένων.
Το πρόγραμμα Spark θα εκτυπώνει στην οθόνη:
Α) τον χρόνο εκτέλεσης όλης της main,
Β) τα μη κανονικοποιημένα σημεία που έχουν οριστεί ως ακρότατα.
Για τον εντοπισμό συστάδων εκτός Spark μπορείτε να χρησιμοποιήσετε ό,τι υλοποίηση θέλετε, π.χ., R. Για την κριτική αξιολόγηση οδηγιών chatGPT, καλείστε να πάρετε τις οδηγίες και να κάνετε σχολιασμό αναφορικά με την επιστημονική τους ορθότητα.

Παραδοτέα

Α. Ένα συμπιεσμένο αρχείο με τον πηγαίο και, αν χρησιμοποιήσετε scala, τον εκτελέσιμο κώδικα σε jar χωρίς dependencies που θα ανέβει στο elearning. Η εργασία να αναφέρει οδηγίες χρήσης. Ο πηγαίος κώδικας να είναι όλος σε ένα αρχείο και να έχει εκτενή σχόλια είτε σε ελληνικά είτε στα αγγλικά και όχι σε greeklish. Το όνομα του συμπιεσμένου αρχείου να είναι *AEM1_AEM2.zip.* **Η προθεσμία είναι Δευτέρα 8/1/2024 στις 13:00.**
Β. Μία τεχνική αναφορά σε μορφή παρουσίασης (σε pdf) που να εξηγείτε την προσέγγισή σας (το πολύ 9 διαφάνειες συνολικά).

Διευκρινήσεις

- Οι γραπτές εξετάσεις θα είναι με άριστα το 9. Όσοι δεν παραδώσουν την άσκηση μπορούν να λάβουν κανονικά μέρος στις εξετάσεις και να περάσουν το μάθημα, απλά θα έχουν μέγιστο βαθμό το 9.
- Η άσκηση θα βαθμολογηθεί με άριστα το 2 και θα ελεγχθεί για ορθότητα και κλιμακωσιμότητα σε επιπλέον σύνολα δεδομένων, που θα είναι παρόμοια με αυτό που δίνεται.
- **Για να προσμετρηθεί ο βαθμός της άσκησης, θα πρέπει ο βαθμός στις τελικές εξετάσεις να είναι τουλάχιστον 4 στα 9.**
- **Στη διάλεξη της 9/1/2024 θα γίνει παρουσίαση της κάθε ομάδας καθώς η άσκηση θα εξεταστεί και προφορικά και όσοι την καταθέσουν είναι υποχρεωμένοι να παραστούν την ημερομηνία αυτή, αλλιώς θα ακυρωθεί η εργασία.** Οι παρουσιάσεις θα πρέπει να διαρκούν 8’.
- Ο βαθμός της εργασίας θα ισχύει μέχρι να ξαναδιδαχθεί το μάθημα.