Εργασία 10 – Clustering

Περιγραφή:

Στόχος της εργασίας είναι η διερεύνηση μεθόδων ομαδοποίησης (Clustering) σε δεδομένα μουσικών κομματιών. Για την εργασία αυτή, θα χρησιμοποιήσετε το dataset του Spotify, το οποίο περιέχει 125 είδη μουσικών κομματιών, (**genres**). Τα κομμάτια θα πρέπει να ομαδοποιηθούν με βάση τα χαρακτηριστικά τους (21 χαρακτηριστικά) σε έναν ικανοποιητικό αριθμό ομάδων. **Το αποτέλεσμα της Εργασίας θα είναι ένα Recommendation System για μουσικά κομμάτια που θα μπορείτε να χρησιμοποιήσετε!**

Ερωτήματα:

1. Επεξεργαστείτε κάθε dataframe ως εξής:
    a. Αφαιρέστε τα χαρακτηριστικά (Number, Track Id, Artists, Album Name, Track Name) από τα και κρατήστε τα σε ένα ξεχωριστό dataframe, καθώς θα τα χρειαστείτε παρακάτω.
    b. Εφαρμόστε One-Hot Encoding στα Genres.
2. Κανονικοποιήστε τα χαρακτηριστικά κατάλληλα.
3. Εφαρμόστε τον αλγόριθμο K-Means για $k = \{2, 3, 5, 7, 10, 15, 20\}$. Μπορείτε να συμβουλευτείτε το documentation του K-Means για να πειράξετε όποια άλλη παράμετρο θέλετε: K-Means. Για κάθε k, να υπολογιστεί η μετρική SSE και έπειτα να δημιουργηθεί Line-Plot με το k.
4. Επιλέξτε το k που ομαδοποιεί καλύτερα τα δεδομένα, σύμφωνα με το Elbow Method και υπολογίστε τα clusters των μουσικών κομματιών (*labels_*).
5. Για κάθε cluster, υπολογίστε το ποσοστό του κυρίαρχου είδους (genre) των κομματιών και δημιουργήστε αντίστοιχο ραβδόγραμμα.
6. Επιλέξτε 1 κομμάτι που σας αρέσει. Αναφέρετε τον τίτλο, τον καλλιτέχνη και το cluster στο οποίο ανήκει. Στη συνέχεια, να προτείνετε τα Top-3 παρόμοια τραγούδια, υπολογίζοντας την Ευκλείδεια απόσταση του κομματιού με τα υπόλοιπα κομμάτια του Cluster που ανήκει και επιλέξτε τα 3 με τη μικρότερη απόσταση (Euclidean Distance). Να αναφέρετε ποια κομμάτια προτείνονται, καθώς και αν (κατά τη δική σας κρίση) ταιριάζουν με το κομμάτι που επιλέξατε.
7. Επαναλάβετε το ερώτημα 6, χρησιμοποιώντας το Cosine Distance (1 – Cosine Similarity) ως μετρική απόστασης και ελέγξτε αν οι συστάσεις είναι καλύτερες.
8. Επαναλάβετε τα ερωτήματα 4, 5, 6, 7, 8, 10 χρησιμοποιώντας όμως τη μέθοδο elbow για την επιλογή των k ομάδων.
9. Να εφαρμόσετε τον dbscan (ή μια γρηγορότερη παραλλαγή του, όπως τον hdbscan) για την ομαδοποίηση των δεδομένων. Συγκρίνετε τον αριθμό των προτεινόμενων cluster με τα ερωτήματα 4 και 8.
10. **Προαιρετικό (Bonus)**: Να εφαρμόσετε τον dbscan++, που είναι μια βελτιωμένη παραλλαγή του dbscan (https://github.com/jenniferjang/dbscanpp) και να συγκρίνετε τα αποτελέσματα με αυτά του ερωτήματος 9.