Ερώτηση 1

Στην ερώτηση αυτή θα εξασκηθείτε με αλγορίθμους για ομαδοποίηση (clustering). Θα χρησιμοποιήσετε και πάλι τα δεδομένα για τις επιχειρήσεις στην Philadelphia, και συγκεκριμένα επιχειρήσεις που είναι εστιατόρια. Για την ομαδοποίηση θα χρησιμοποιήσετε το κείμενο των reviews για τις επιχειρήσεις. Για τη αξιολόγηση θα ερευνήσουμε αν τα clusters που βρίσκουμε συμφωνούν με την κατηγορία της επιχείρησης.

Σας δίνεται το αρχείο philly_businesses.csv στη σελίδα ασκήσεων του μαθήματος με τις επιχειρήσεις στην Philadelphia. Από το αρχείο αυτό κρατήστε μόνο τις επιχειρήσεις που εμφανίζεται το String Restaurant μέσα στο πεδίο categories, και επίσης ένα από τα παρακάτω Strings: Japanese, Italian, Burgers. Αφαιρέστε τις επιχειρήσεις που εμφανίζονται σε παραπάνω από μία από τις τρεις παραπάνω κατηγορίες. Θα πρέπει να μείνετε με 951 επιχειρήσεις. Κάθε επιχείρηση έχει πλέον ένα μοναδικό category από τα Japanese, Italian, Burgers.

Τραβήξτε όλα τα reviews που έχουν γίνει γι αυτές τις επιχειρήσεις, και συνενώστε τα σε ένα μεγάλο String. Χρησιμοποιήστε τον tf-idf vectorizer και πάρετε ένα διάνυσμα για κάθε επιχείρηση (αφαιρέστε τα stop-words και σας προτείνεται να βάλετε κάποια κατώφλια στον αριθμό των features, στο document frequency των λέξεων, και στον αριθμό των εμφανίσεων μιας λέξης). Αυτά είναι τα δεδομένα που θέλουμε να κάνουμε cluster.

Θα κάνετε τα εξής βήματα

1. Εφαρμόστε τον k-means αλγόριθμο και τον agglomerative clustering αλγόριθμο για όλες τις διαφορετικές επιλογές της παραμέτρου linkage, για αριθμό clusters $k = 3$. Εκτυπώστε τον πίνακα σύγχυσης μεταξύ των cluster labels και των categories. Τι παρατηρείτε για την απόδοση των αλγορίθμων?
2. Για τον k-means αλγόριθμο, πάρετε τα κέντρα των clusters και για κάθε κέντρο τυπώστε τις 10 λέξεις με τις μεγαλύτερες τιμές. Τι παρατηρείτε και τι συμπέρασμα βγάζετε?
3. Για τον k-means αλγόριθμο και τον καλύτερο agglomerative αλγόριθμο από το Βήμα 1, υπολογίστε το precision, recall, F1-measure για κάθε cluster και συνολικά. Τι παρατηρείτε?
4. Για τον k-means αλγόριθμο, δοκιμάστε τιμές του k από 2 εώς 10 και δημιουργείστε το συνδυασμένο διάγραμμα με το SSE και το Silhouette Coefficient για να αποφασίσετε ποιος είναι ο “σωστός” αριθμός από clusters. Σχολιάστε το γράφημα και την απόφαση σας.
   Στη συνέχεια δημιουργείστε τον πίνακα σύγχυσης για την τιμή του k που επιλέξατε και τυπώστε πάλι τις 10 πιο σημαντικές λέξεις από τα κέντρα των clusters. Τι παρατηρείτε?

Παραδώσετε ένα notebook με τον κώδικα σας και την αναφορά με τον σχολιασμό και ανάλυση των αποτελεσμάτων σας.

Υποδείξεις:

- Μπορεί να σας φανεί χρήσιμη η μέθοδος agg για pandas groupby
- Μπορείτε να χρησιμοποιήσετε την συνάρτηση που φτιάξαμε στην τάξη για να αντιστοιχίσετε τα cluster με τα category labels.