ΠΑΠΑΔΟΠΟΥΛΟΣ ΝΙΚΟΛΑΟΣ 4140

ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
ΔΕΥΤΕΡΟ ΣΕΤ ΑΣΚΗΣΕΩΝ

ΑΝΑΦΟΡΑ ΤΡΙΤΗΣ ΑΣΚΗΣΗΣ

Σημείωση: στο notebook, σε κάποια φάση εκτέλεσης, παρουσιάστηκε πρόβλημα μνήμης που με ανάγκασε να ξαναρχίσω το kernel και να χάσω τις μεταβλητές, να ξανατρέξω ό,τι ήταν απαραίτητο για τα blocks που προηγουμένως δεν μπορούσα να εκτελέσω λόγω ανεπάρκειας μνήμης, κρατώντας όμως και τα outputs από blocks που δεν κράτησα τις μεταβλητές τους για εξοικονόμηση μνήμης. Έτσι, παρόλο που όλα τα blocks έχουν outputs, μερικά δεν είναι εκτελεσμένα στην παρούσα φάση του notebook.

ΕΡΩΤΗΜΑ 1

K-means clustering: Φαίνεται από τις κορυφαίες λέξεις του κάθε cluster ότι το πρώτο αντιστοιχεί στα spas, το δεύτερο ταιριάζει περισσότερο σε shopping και το τρίτο σε εστιατόρια. Από το confusion matrix και το precision matrix συμπεραίνει κανείς ότι στο cluster 0 κατέληξαν μόνο επιχειρήσεις beauty & spas, δηλαδή καμία δεν κατέληξε λανθασμένα εκεί, όμως δεν κατατάχθηκαν όλες οι επιχειρήσεις beauty & spas εκεί. Μερικές κατέληξαν λανθασμένα στο δεύτερο cluster (shopping). Οι περισσότερες επιχειρήσεις shopping κατατάχθηκαν σωστά, όπως και τα εστιατόρια. Παρόλα αυτά, στο cluster του shopping έπεσαν γενικά πολλές ξένες επιχειρήσεις.

Agglomerative clustering with complete linkage: Φαίνεται από τις κορυφαίες λέξεις του κάθε cluster ότι το πρώτο αντιστοιχεί στα spas κυρίως (αν και βρίσκονται λέξεις όπως food, drinks που προϊδεάζουν ότι κατατάχθηκαν και bars σε αυτό), το δεύτερο ταιριάζει περισσότερο σε shopping με ορισμένα spas και το τρίτο σε εστιατόρια/bars. Αυτά τα συμπεράσματα επιβεβαιώνονται και από το confusion matrix και τις μετρικές, καθώς βλέπουμε στη σειρά του confusion matrix για το 1ο cluster υψηλές τιμές και στην πρώτη (spas) και στην τρίτη (bars) στήλη. Μία σύγχυση παρατηρείται και στο δεύτερο cluster. Το τρίτο είναι αρκετά ακριβές, όπως φαίνεται και από το precision score του. Στον recall score πίνακα αξίζει να σημειωθεί ότι έχουμε υψηλή τιμή (0.91179713) για το score του cluster του shopping, που σημαίνει ότι τα περισσότερα shopping businesses κατατάχθηκαν σωστά.

Agglomerative clustering with single linkage: Φαίνεται πως το cluster 1 με αυτή τη μέθοδο κατέληξε να είναι cluster kano & kayak και το cluster 2 ηλεκτρονικά, ενώ το cluster 0 είναι μία μίξη από bars και spas. Απτη στιγμή που δεν έχουμε true labels για kano & kayak και ηλεκτρονικά, είναι αναμενόμενο το confusion matrix να έχει μηδενικές τιμές στις θέσεις αυτών των κατηγοριών.

Agglomerative clustering with ward linkage: Εδώ φαίνεται πως προέκυψαν δύο clusters για beauty & spas, με λιγάκι διαφορετικές κορυφαίες λέξεις. Έτσι, αναγκαστικά κάποιο από τα δύο clusters δε θα ταιριάξει καλά με το true label του και αυτό γίνεται ξεκάθαρο στο confusion matrix,

καθώς και από το χαμηλό recall score για beauty & spas (0.25181598), που σημαίνει ότι δεν έκανα την ιδανική επιλογή για το true label μεταξύ των δύο clusters, καθώς οι περισσότερες από αυτές τις επιχειρήσεις πήγαν στα άλλα clusters. Απ΄τα recall scores αξιοσημείωτη είναι η μονάδα στη δεύτερη στήλη, που σημαίνει ότι όλα τα bars μαζεύτηκαν στο σωστό cluster.

Agglomerative clustering with average linkage: Το linkage με τα χειρότερα αποτελέσματα. Αυτό γιατί δεν έχουμε κατάλληλα true labels για το δεύτερο και τρίτο cluster, όπου από τις κορυφαίες λέξεις τους, το δεύτερο φαίνεται να ομαδοποιεί καταστήματα με τσιγάρα και συναφή προϊόντα, ενώ το τρίτο φαίνεται να έχει συγκεντρώσει επιχειρήσεις με σπαθιά και παρόμοια όπλα. Προφανώς το γράφημα του confusion matrix έχει δύο κενές σχεδόν σειρές με κουτιά. Καταλαβαίνουμε επίσης ότι στο πρώτο cluster έπεσαν υπερβολικά πολλές επιχειρήσεις που δεν ανήκουν σε αυτό.

ΕΡΩΤΗΜΑ 2

Από το silhouette plot φαίνεται πως το k = 4 είναι ιδανική για τον αλγόριθμο k means στη δικιά μας εφαρμογή, εφόσον επιτυγχάνει το υψηλότερο silhouette score από τα k στο διάστημα [0,10], που σημαίνει ότι οι επιχειρήσεις που πέφτουν στο ίδιο cluster θα είναι πολύ παρόμοιες μεταξύ τους. Επίσης το σημείο αυτό είναι και τοπικό μέγιστο του plot, άρα αν αυξήσουμε (η μειώσουμε το k) θα πάρουμε χειρότερα αποτελέσματα. Τα συμπεράσματα αυτά επιβεβαιώνονται και από το elbow plot που έκανα, αφού στο k = 4 φαίνεται να συναντάμε το elbow point (υπάρχει μεγάλη πτώση error πριν αυτό το σημείο, ενώ το error από κει και ύστερα μειώνεται με πολύ μικρότερους ρυθμούς.

Δοκιμάζω, λοιπόν, τον αλγόριθμο για k = 4. Φαίνεται ότι στο πρώτο cluster, με βάση τις κορυφαίες λέξεις, πέφτουν τα spas. Το δεύτερο cluster αφορά ξεκάθαρα κομμωτήρια/ barber shops (που με k = 3 μπορούμε να υποθέσουμε ότι θα έπεφτα στο cluster με τα spas, δημιουργώντας ανομοιότητα). Το τρίτο cluster αφορά shopping και το τελευταίο εστιατόρια και bars. Δεν προχώρησα σε confusion matrix και μετρικές, αφού δεν είχα true labels για barber shops, οπότε δεν έχει νόημα ένας μη τετραγωνικός confusion matrix.

ΕΡΩΤΗΜΑ 3

Τυπώνοντας τα counters των κατηγοριών των επιχειρήσεων οι οποίες ανήκουν στο cluster Beauty & Spa, αλλά πάνε στο cluster που αντιστοιχεί στο Shopping, φαίνεται ξεκάθαρο γιατί συμβαίνει αυτό. 74 από τις 341 επιχειρήσεις που εμφανίζουν αυτό το φαινόμενο έχουν ως δευτερεύουσα κατηγορία το shopping, που σημαίνει ότι είναι πιθανό να πέσουν στο άλλο cluster λόγω του λεξιλογίου των κριτικών. Επίσης, εμφανίζονται κατηγορίες όπως Fashion : 9, Drugstores : 8, Jewelry : 7 κλπ οι οποίες πιθανόν να αφορούν επιχειρήσεις της μορφής καταστημάτων.

Τέλος, εκμεταλλεύομαι τα reviews αυτών των επιχειρήσεων ώστε να δω ποιες λέξεις εμφανίζονται συχνότερα σε αυτές και να καταλάβω γιατί κατατάχθηκαν σε λάθος cluster. Terms όπως customer : 1081, deal : 636, price : 1154, quality : 556 κλπ που φαίνεται πως έχουν πολλές εμφανίσεις στα reviews αυτών των επιχειρήσεων παραπέμπουν και σε εμπειρίες shopping. Γι’ αυτό βοηθούν την κατάταξη των επιχειρήσεων στο cluster του shopping.