Programming Project 2

Κανόνες Συσχέτισης:

Η μέθοδος ‘association_rules’ από την ενότητα mlxtend.frequent_patterns χρησιμοποιείται για την εύρεση κανόνων συσχέτισης στα δεδομένα κειμένου. Η συνάρτηση έχει διάφορες παραμέτρους:

1. ‘frequent_item_sets’: είναι το αποτέλεσμα του αλγόριθμου apriori που εφαρμόζεται στον πίνακα top-k λέξεων κάθε εγγράφου, όπου κάθε έγγραφο αντιμετωπίζεται ως μια συναλλαγή και κάθε λέξη ως ένα στοιχείο. Το ελάχιστο όριο υποστήριξης για τα συχνά στοιχειοσύνολα ορίζεται σε 0,5, δηλαδή ένα σύνολο αντικειμένων πρέπει να εμφανίζεται σε τουλάχιστον 50% των συναλλαγών για να θεωρηθεί συχνό. Η παράμετρος ‘use_colnames=True’, καθορίζει ότι τα ονόματα στοιχείων στο πλαίσιο δεδομένων θα πρέπει να χρησιμοποιούνται ως ονόματα στηλών αντί για τους προεπιλεγμένους ακέραιους δείκτες.
2. ‘metric’: η μετρική που χρησιμοποιείται για την αξιολόγηση της ποιότητας των κανόνων που βρέθηκαν. Οι τιμές που μπορεί να πάρει είναι ‘support’, ‘confidence’, ‘lift’, ‘leverage’ και ‘conviction’. Στον κώδικά μου χρησιμοποιώ την εμπιστοσύνη ως μετρική αξιολόγησης.
3. ‘min_threshold’: ορίζει το ελάχιστο όριο της μεταβλητής ‘metric’ για τους κανόνες συσχέτισης. Η εμπιστοσύνη είναι ένα μέτρο της αξιοπιστίας του κανόνα συσχέτισης και ορίζεται ως το ποσοστό των συναλλαγών στις οποίες υπάρχει το antecedent (το μέρος "if" του κανόνα) που περιέχουν επίσης το consequent (το μέρος "then" του κανόνα). Στην προκειμένη περίπτωση έχω ορίσει το κατώφλι της εμπιστοσύνης σε 0,8. Αυτό σημαίνει ότι ένας κανόνας συσχέτισης πρέπει να έχει τουλάχιστον 80% εμπιστοσύνη για να θεωρηθεί έγκυρος.
4. Η μέθοδος ‘itertuples’ μετατρέπει το προκύπτον πλαίσιο δεδομένων pandas σε έναν επαναλήπτη namedtuples, όπου κάθε πλειάδα αντιπροσωπεύει έναν κανόνα συσχέτισης. Ο βρόχος for-loop στην συνέχεια επαναλαμβάνει αυτές τις πλειάδες και αποσυμπιέζει τις τιμές για τα antecedents, consequents, support, confidence και lift σε μεταβλητές antecedents, consequents, support, conf και lift, αντίστοιχα.

Οι τιμές 0,5 και 0,8 που χρησιμοποίησα είναι απλώς ενδεικτικές τιμές και χρησιμοποιούνται για να μην εμφανιστεί υπερβολικά υψηλός αριθμός κανόνων συσχέτισης.

Αποτελέσματα κανόνα για min_support = 0.5, min_confidence = 0.8:
organ -> line (Confidence: 0.9838192216487655, Lift: 1.5127657887651715)
line -> organ (Confidence: 0.9585485186191899, Lift: 1.5127657887651715)
line -> subject (Confidence: 0.9847784724109812, Lift: 1.5281557587241585)

Σοφιανοπούλου Αποστολία – Μαρία
iis21118

subject -> line (Confidence: 0.9938280071320805, Lift: 1.5281557587241585)
organ -> subject (Confidence: 0.978658111312596, Lift: 1.5186583282664532)
subject -> organ (Confidence: 0.9622822658071595, Lift: 1.5186583282664532)
organ, line -> subject (Confidence: 0.9910676307954063, Lift: 1.5379151247866172)
organ, subject -> line (Confidence: 0.9962941847206386, Lift: 1.5319478670738387)
line, subject -> organ (Confidence: 0.9646701628484683, Lift: 1.5224268688056313)
organ -> line, subject (Confidence: 0.9750313851304228, Lift: 1.5224268688056313)
line -> organ, subject (Confidence: 0.9499864093503669, Lift: 1.5319478670738385)
subject -> organ, line (Confidence: 0.9587162254834728, Lift: 1.5379151247866172)

Κατηγοριοποίηση Κειμένου:

Εφαρμόζω πάλι τη μέθοδο ‘association_rules’ από την ενότητα mlxtend.frequent_patterns για την εύρεση κανόνων συσχέτισης, ωστόσο τώρα έχω κρατήσει μόνο τις 6 πρώτες κατηγορίες και έχω προσθέσει στις λέξεις των εγγράφων και την κατηγορία του κάθε εγγράφου. Για να πάρω κανόνες συσχέτισης, οι οποίοι έχουν στο δεξί τους μέρος την κατηγορία του εγγράφου, εφαρμόζω την συνθήκη «if len(conseq_words) == 1 and conseq_words[0] in categories». Ακόμα, ορίζω min_support=0,4 και min_confidence=0,5 προκειμένου να εμφανιστεί επιθυμητός αριθμός κανόνων.