Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 17 Νοεμβρίου 11:59 μ.μ. Κάνετε turn-in τον κώδικα σας και τα αποτελέσματα σας, και παραδώστε τις υπόλοιπες ερωτήσεις είτε ηλεκτρονικά, είτε σε χαρτί. Η αναφορά σας θα πρέπει να είναι γραμμένη ηλεκτρονικά. Για καθυστερημένες υποβολές ισχύει η πολιτική στην σελίδα του μαθήματος. Λεπτομέρειες για το turn-in, και για το πώς να γράφετε αναφορές είναι στη σελίδα Ασκήσεις του μαθήματος.

Ερώτηση 1 (Reservoir Sampling)

Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο Reservoir Sampling ώστε να κάνει δειγματοληψία $\boldsymbol{K}$ αντικειμένων από ένα ρεύμα N αντικειμένων ομοιόμορφα τυχαία, ώστε το κάθε αντικείμενο να έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.

1. Περιγράψετε τον αλγόριθμο που διαλέγει ένα ομοιόμορφο δείγμα K αντικειμένων από ένα ρεύμα N αντικειμένων. Ο αλγόριθμος σας θα πρέπει να δουλεύει με ένα μόνο πέρασμα στα δεδομένα διαβάζοντας τα αντικείμενα ένα-ένα, χωρίς προηγούμενη γνώση του μεγέθους του ρεύματος, και να χρησιμοποιεί $O(K)$ μνήμη (υποθέστε ότι το μέγεθος του κάθε αντικειμένου είναι σταθερό). Η περιγραφή του αλγορίθμου **δεν** πρέπει να είναι σε κώδικα ή ψευδοκώδικα, αλλά να εξηγεί τη λογική του αλγορίθμου στα Ελληνικά με απλό τρόπο.
2. Αποδείξτε ότι ο αλγόριθμος σας παράγει ένα ομοιόμορφα τυχαίο δείγμα, δηλαδή, για κάθε $i, 1 \leq i \leq N$, το i -οστό στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.
3. Γράψτε ένα πρόγραμμα **σε Python** που υλοποιεί τον αλγόριθμο σας. Το πρόγραμμα σας θα πρέπει να παράγει ένα δείγμα με K τυχαίες γραμμές από ένα αρχείο κειμένου. Θα πρέπει να μπορούμε να τρέξουμε το πρόγραμμα από την γραμμή εντολών, θα παίρνει σαν όρισμα εντολής την τιμή του K, θα διαβάζει γραμμές από το standard input και θα εκτυπώνει το δείγμα στο standard output. Για παράδειγμα, η παρακάτω εντολή θα πρέπει να τυπώνει στην οθόνη ένα τυχαίο δείγμα 10 γραμμών από το αρχείο input.txt:
   “`sample.py 10 < input.txt`”.

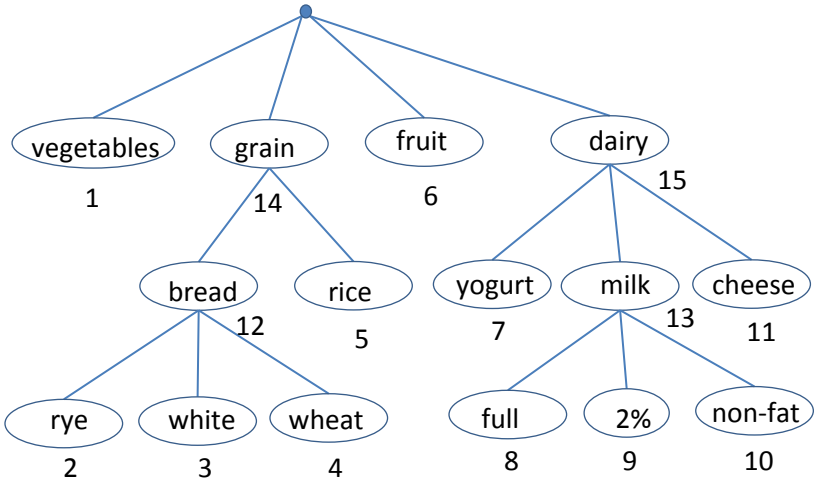
Ερώτηση 2

Στη σελίδα Ασκήσεις του μαθήματος σας δίνεται το αρχείο “data.csv”. Το αρχείο έχει τρεις στήλες χωρισμένες με κόμμα, με ονόματα A, B, C, και 1000 γραμμές. Οι τιμές των B και C είναι συνάρτηση αυτών της A. Ο στόχος σας είναι να βρείτε την συνάρτηση μεταξύ των στηλών B, C και της στήλης A. Για να καταλάβετε την σχέση δημιουργήστε γραφήματα των B και C ως προς το A, όπως είδαμε την τάξη. Παραδώσετε ένα Iron Python Notebook το οποίο θα περιέχει τον κώδικα για την επεξεργασία των δεδομένων, τα γραφήματα και τους υπολογισμούς που κάνατε, καθώς και μία αναφορά με τα συμπεράσματα σας.

Ερώτηση 3

Στην παρακάτω εικόνα σας δίνεται ένα απλό ταξινομικό δέντρο τροφών. Κάθε φύλλο αντιστοιχεί σε ένα απλό στοιχείο. Κάθε εσωτερικός κόμβος αναπαριστά μια κατηγορία, ή αλλιώς ένα στοιχείο υψηλότερου επιπέδου. Κάθε κόμβος αντιστοιχεί σε ένα αριθμό που φαίνεται στο δέντρο. Απαντήστε τα παρακάτω ερωτήματα.

1. Έστω ότι $X = \{x_1, x_2, \dots, x_k\}$ είναι ένα συχνό στοιχειοσύνολο. Ας υποθέσουμε ότι αντικαθιστούμε κάποιο x_i με κάποιον πρόγονο του στο ταξινομικό δέντρο (εφόσον έχει), ώστε να προκύψει ένα νέο στοιχειοσύνολο X'. Ποια είναι η σχέση μεταξύ της υποστήριξης του X και του X'?
2. Θεωρήστε την βάση δεδομένων από απλά στοιχεία που φαίνεται στον παρακάτω πίνακα. Χρησιμοποιείστε το κατώφλι minsup = 7/8. Βρείτε όλα τα συχνά στοιχειοσύνολα που αποτελούνται από στοιχεία υψηλότερου επιπέδου στην ταξινομία (όχι φύλλα). Λάβετε υπόψη ότι αν ένα απλό στοιχείο εμφανίζεται σε μια συναλλαγή, τότε υποθέτουμε ότι και όλοι οι πρόγονοι του εμφανίζονται στην συναλλαγή.
3. Με βάση την μεθοδολογία που χρησιμοποιήσατε στο 2 προτείνετε ένα αλγόριθμο για την εύρεση συχνών στοιχειοσυνόλων που περιλαμβάνουν στοιχεία υψηλού επιπέδου για δεδομένα οργανωμένα σε ταξινομία. Ο αλγόριθμος σας θα πρέπει να είναι αποτελεσματικός και να μην κοιτάει όλα τα πιθανά στοιχειοσύνολα.



tid	items
1	2 3 6 7
2	1 3 4 6 8 11
3	3 9 11
4	1 5 6 7
5	1 3 8 10 11
6	3 5 7 9 11
7	4 6 8 10 11
8	1 3 5 8 11

Ερώτηση 4

Για την ερώτηση αυτή θα πρέπει να χρησιμοποιήσετε τα δεδομένα από το Yelp Academic Challenge dataset. Μπορείτε να τα βρείτε εδώ. Δεν μας ενδιαφέρει το κομμάτι με τις εικόνες. Σας προτείνετε να κατεβάσετε τα δεδομένα σε JSON μορφή για πιο εύκολη επεξεργασία.

Για την ερώτηση θα επικεντρωθούμε στα δεδομένα για την πόλη του Τορόντο. Ο στόχος μας είναι να καταλάβουμε τον χαρακτήρα των διαφορετικών γειτονιών του Τορόντο, χρησιμοποιώντας τα δεδομένα του Yelp. Από το αρχείο business.json, κρατήστε τις επιχειρήσεις στην πόλη του Τορόντο, για τις οποίες έχουμε πληροφορία για την γειτονιά στην οποία βρίσκονται και για τις κατηγορίες στις οποίες ανήκουν. Κρατήσετε τις γειτονιές για τις οποίες έχουμε τουλάχιστον 100 επιχειρήσεις.

Στη συνέχεια, χρησιμοποιώντας αυτά τα δεδομένα, προτείνετε μια μεθοδολογία για να βρείτε για κάθε περιοχή τις 10 κατηγορίες που την περιγράφουν καλύτερα. Λάβετε υπόψη σας ότι οι κατηγορίες θα πρέπει να είναι αρκετά συχνές στη γειτονιά και πιο συχνές απ ότι σε άλλες γειτονιές (π.χ., η κατηγορία Food δεν είναι ενδιαφέρουσα μιας και εμφανίζεται παντού με την ίδια περίπου συχνότητα – αλλά δεν μας ενδιαφέρουν και πάρα πολύ σπάνιες κατηγορίες). Η μεθοδολογία σας θα πρέπει να ορίζει μια μετρική που μας λέει πόσο σημαντική είναι μια κατηγορία για την γειτονιά. Μπορείτε να επεξεργαστείτε με διάφορους τρόπους τα δεδομένα και να πετάξετε κάποια από τα δεδομένα που θεωρείτε ότι δεν προσφέρουν χρήσιμη πληροφορία. Κάνετε τις επιλογές σας σαφείς μέσα στην περιγραφή της μεθοδολογίας.

Αναφέρετε για κάθε γειτονιά τις κατηγορίες που βγάζει η μέθοδος σας (απλή αναφορά είναι ΟΚ, αλλά αν θέλετε να κάνετε κάποιας μορφής οπτικοποίηση θα προσμετρηθεί θετικά). Στη συνέχεια για κάθε γειτονιά ψάξετε να βρείτε τα πραγματικά της χαρακτηριστικά και αναφέρετε κατά πόσο οι κατηγορίες που βρήκατε περιγράφουν σωστά τη γειτονιά. Χρησιμοποιείστε το διαδίκτυο γι αυτό το σκοπό: υπάρχουν αρκετοί τουριστικοί οδηγοί online που δίνουν πληροφορίες ανά γειτονιά, και το Airbnb έχει περιλήψεις για τις διάφορες γειτονιές.

Στην αναφορά σας θα πρέπει να έχετε μια περιγραφή της μεθοδολογίας σας, τα αποτελέσματα σας, και σχολιασμό των αποτελεσμάτων.

Bonus: Για πολλές επιχειρήσεις υπάρχει και ένα πεδίο Ambiance όπου για διάφορες λέξεις-κλειδιά που περιγράφουν της ατμόσφαιρας της επιχείρησης έχουμε μια true/false τιμή. Εφαρμόστε την μεθοδολογία σας χρησιμοποιώντας αυτές τις λέξεις κλειδιά αντί για τις κατηγορίες και σχολιάστε την περιγραφή της γειτονιάς που παίρνουμε με αυτές τις λέξεις.