Πρώτη Σειρά Ασκήσεων

Αυτή είναι η πρώτη σειρά ασκήσεων. Η προθεσμία για την παράδοση είναι στις 31 Μαρτίου 11:59 μ.μ. Κάνετε turn-in τον κώδικα σας, και παραδώστε τις υπόλοιπες ερωτήσεις είτε ηλεκτρονικά, είτε σε χαρτί. 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. Θα γίνει προφορική εξέταση των ασκήσεων μέσα στην εβδομάδα πριν το Πάσχα.

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.

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 τα βήματα 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 δίνονται 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 σχέση μεταξύ των στηλών B, C και της στήλης A. 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

Σε αυτή την άσκηση θα υλοποιήσετε και θα εφαρμόσετε τον Apriori αλγόριθμο για την εύρεση συχνών στοιχειοσυνόλων. Υλοποιήστε την βασική εκδοχή του αλγορίθμου, όπως περιγράψαμε στην τάξη.

Η εφαρμογή του αλγορίθμου θα γίνει πάνω σε ένα σύνολο από tweets που σας δίνονται στο αρχείο "tweets_dataset.txt" στη σελίδα του μαθήματος. Το αρχείο αυτό περιέχει tweets από followers του Donald Trump και της Hillary Clinton τα οποία συλλέχτηκαν στο διάστημα 25-30 Οκτωβρίου 2016, εν μέσω της προεκλογικής περιόδου. Ο στόχος μας είναι να βρούμε ενδιαφέρουσες συνεμφανίσεις από tags και handles στα tweets (αυτά θα είναι τα items για την περίπτωση μας).

Καταρχάς πρέπει να καθαρίσετε τα δεδομένα. Το αρχείο περιέχει tab-separated εγγραφές με 14 στήλες που αντιστοιχούν στα εξής πεδία:

Name, ScreenName, UserID, FollowersCount, FriendsCount, Location, Description, CreatedAt, StatusID, Language, Place, RetweetCount, FavoriteCount, Text

Η κάθε γραμμή του αρχείου είναι ένα tweet. Πετάξετε όλα τα tweets τα οποία είναι retweets (το κείμενο ξεκινάει με RT) και από το κείμενο κρατήσετε μόνο τα hashtags (λέξεις που ξεκινάνε με #) και τα handles (λέξεις που ξεκινάνε με @). Σε μερικά tweets υπάρχουν σημεία στίξεως στο τέλος του hashtag ή handle τα οποία πρέπει να καθαρίσετε ή είναι κολλημένα κάποια hashtags. Καθαρίστε επίσης άλλα προφανή λάθη (μπορείτε να πάρετε ένα δείγμα του αρχείου για να βρείτε προβληματικές περιπτώσεις). Δημιουργήστε ένα "καλάθι" για κάθε tweet που έχει τουλάχιστον ένα hashtag ή handle. Σώσετε το αποτέλεσμα σε ένα νέο αρχείο.

Χρησιμοποιώντας τα δεδομένα που δημιουργήσατε τυπώσετε κάποια στατιστικά ώστε να δώσετε μια ιδέα για το πώς μοιάζουν τα δεδομένα σας (π.χ., αριθμό από καλάθια, αριθμό από διακριτά items, μέσο αριθμό από items per basket, ιστόγραμμα του μεγέθους των καλαθιών, αριθμό hashtags vs handles, κλπ). Σχολιάστε τα αποτελέσματα.

Εφαρμόσετε τον Apriori αλγόριθμο για την εύρεση συχνών στοιχειοσυνόλων πάνω στα δεδομένα σας με support threshold μεταξύ 0.02% και 0.05% των καλαθιών. Βρείτε τα maximal συχνά στοιχειοσύνολα (συχνά στοιχειοσύνολα που δεν περιέχονται σε κάποιο μεγαλύτερο συχνό στοιχειοσύνολο). Δεδομένου ότι τα συχνά στοιχειοσύνολα δεν είναι πολλά χρησιμοποιήστε μια απλή υλοποίηση για να βρείτε τα maximal στοιχειοσύνολα. Δημιουργήστε ένα γράφημα που δείχνει πως αλλάζει ο αριθμός για διαφορετικά support thresholds.

Κοιτάξετε με το μάτι τα αποτελέσματα (ειδικά τα πιο συχνά στοιχειοσύνολα) και σχολιάστε αν υπάρχουν ενδιαφέρουσες συσχετίσεις και τι μπορεί να σημαίνουν. Σχολιάσετε στην αναφορά σας τις συσχετίσεις που έχουν ενδιαφέρον.

Παραδώσετε τον κώδικα σας (συνίσταται να κάνετε την άσκηση σε Python, αλλά μπορεί να είναι και σε άλλη γλώσσα), αρχεία με τα maximal συχνά στοιχειοσύνολα, και την αναφορά σας όπου θα έχετε τα γραφήματα και τις παρατηρήσεις σας (μπορείτε να βάλετε μαζί κώδικα και αναφορά σε ένα IPython notebook). Η αναφορά είναι πολύ σημαντική για τη βαθμολόγηση της άσκησης γιατί θα πρέπει να δείχνει τι ενδιαφέρον βρήκατε στα δεδομένα.

Bonus: Ένας διαφορετικός τρόπος να δημιουργήσουμε “καλάθια” είναι βάζοντας μαζί όλα τα hashtags και handles τα οποία έχει κάνει tweet ένας χρήστης (χωρίς επαναλήψεις). Βρείτε τα maximal συχνά στοιχειοσύνολα και σε αυτή την περίπτωση και συγκρίνετε με αυτά που βρήκατε όταν κάνατε ένα καλάθι ανά tweet.