

Πρώτη Σειρά Ασκήσεων

Αυτό είναι το πρώτο μέρος της πρώτης σειράς ασκήσεων. Το κομμάτι αυτό θα πρέπει να παραδοθεί στο ξεκίνημα του μαθήματος που θα γίνει την εβδομάδα της 27^{ης} Οκτωβρίου (Πέμπτη ή Παρασκευή). 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.

Ερώτηση 1 (Reservoir Sampling)

Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον απλό (χωρίς βάρη) αλγόριθμο Reservoir Sampling ώστε να κάνει δειγματοληψία K αντικειμένων από ένα ρεύμα N αντικειμένων.

1. Αποδείξτε ότι αν επιλέξουμε K στοιχεία ομοιόμορφα τυχαία από ένα σύνολο N στοιχείων η πιθανότητα ένα συγκεκριμένο στοιχείο να βρίσκεται στο δείγμα είναι $\frac{K}{N}$.
2. Περιγράψτε 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).
3. 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.
4. Γράψτε ένα πρόγραμμα που υλοποιεί τον αλγόριθμο σας (σε όποια γλώσσα θέλετε). Το πρόγραμμά σας θα πρέπει να παράγει ένα δείγμα με K τυχαίες γραμμές ενός αρχείου εισόδου. Θα πρέπει να μπορούμε να τρέξουμε το πρόγραμμα από την γραμμή εντολών, και θα παίρνει σαν ορίσματα εντολής την τιμή του K , το όνομα του αρχείου εισόδου, και το όνομα του αρχείου εξόδου. Δηλαδή, το πρόγραμμά σας θα πρέπει να τρέχει ως εξής:
“sample <K> <inputfile> <outputfile>”.

Ερώτηση 2

Στη σελίδα με τις ασκήσεις του μαθήματος σας δίνονται δύο αρχεία “dataset1.txt” και “dataset2.txt”. Το κάθε αρχείο έχει δύο στήλες χωρισμένες με tab με 100 τιμές. Η κάθε στήλη αντιστοιχεί σε μία διαφορετική χρονο-σειρά, και οι χρονο-σειρές μέσα στο ίδιο αρχείο συνδέονται με κάποιο τρόπο. Για κάθε σύνολο δεδομένων, αναλύστε τα δεδομένα και γράψτε μια μικρή αναφορά για την ανάλυση που κάνατε και τα αποτελέσματά της. Η τελική σας αναφορά θα πρέπει να έχετε πειστικές ενδείξεις (με νούμερα ή/και γραφικές παραστάσεις) που να τεκμηριώνουν τη σχέση των δύο χρονο-σειρών.