

Πρώτη Σειρά Ασκήσεων

Αυτό είναι το πρώτο μέρος της πρώτης σειράς ασκήσεων. Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 8 Νοεμβρίου, 11:00 π.μ., στο ξεκίνημα του μαθήματος. 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 (Weighted Reservoir Sampling)

Στην τάξη περιγράψαμε τον αλγόριθμο Reservoir Sampling για τη δειγματοληψία ενός αντικειμένου από ένα ρεύμα N αντικειμένων. Σε αυτή την άσκηση θα πρέπει να τροποποιήσετε τον αλγόριθμο ώστε να κάνει **σταθμισμένη δειγματοληψία**. Υποθέτουμε ότι το κάθε αντικείμενο i έχει βάρος w_i . Θα τροποποιήσετε τον αλγόριθμο δειγματοληψίας ώστε από ένα ρεύμα N αντικειμένων με βάρη, να επιλέγει ένα αντικείμενο με πιθανότητα ανάλογη προς το βάρος του αντικειμένου, δηλαδή το αντικείμενο i θα επιλεγεί με πιθανότητα ανάλογη του w_i . Όπως και με τον κλασικό Reservoir Sampling αλγόριθμο, ο αλγόριθμος θα πρέπει να δουλεύει με σταθερό χώρο μνήμης, ανεξάρτητο του N . Αποδείξτε την ορθότητα του αλγορίθμου σας.

Ερώτηση 2 (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. 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).
(Υπόδειξη: Σε ένα ομοιόμορφα τυχαίο δείγμα, κάθε στοιχείο έχει πιθανότητα K/N να εμφανιστεί στο δείγμα.)
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.
(Υπόδειξη: Ποια είναι η πιθανότητα αν επιλεγεί το i -οστό στοιχείο, να αντικατασταθεί αργότερα από το j -οστό, για $j > i$?)

3. Γράψτε ένα πρόγραμμα που υλοποιεί τον αλγόριθμο σας (σε όποια γλώσσα θέλετε). Το πρόγραμμα 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 πρόγραμμα σας θα πρέπει να τρέχει ως εξής:
`"sample <K> <inputfile> <outputfile>".`

Ερώτηση 3

Στη σελίδα με τις ασκήσεις του μαθήματος σας δίνεται το αρχείο "stocks.txt" με δύο στήλες χωρισμένες με tab. Η κάθε γραμμή αντιστοιχεί στις τιμές δύο μετοχών μια μέρα, και το αρχείο έχει τις τιμές για 100 μέρες. Φανταστείτε ότι είσαστε data analyst σε μία τράπεζα και σας ενδιαφέρει να βρείτε αν υπάρχει κάποια σχέση μεταξύ των δύο μετοχών. Αναλύστε τα δεδομένα και γράψετε μια μικρή αναφορά για την ανάλυση που κάνατε και τα αποτελέσματα της. Η τελική σας αναφορά θα πρέπει να έχετε πειστικές ενδείξεις (με νούμερα ή/και γραφικές παραστάσεις) που να τεκμηριώνουν τη σχέση των δύο μετοχών.