Πανεπιστήμιο Ιωαννίνων, Εξόρυξη Δεδομένων – ΜΥΕ012/ΠΛΕ 059
Τμήμα Μηχανικών Πληροφορικής Χειμερινό Εξάμηνο, 2015

# Πρώτη Σειρά Ασκήσεων

Αυτό είναι το πρώτο μέρος της πρώτης σειράς ασκήσεων. Η προθεσμία για την παράδοση αυτού του κομματιού είναι στις 5 Νοεμβρίου, 1:00 μ.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 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 αντικειμένων με βάρη, να επιλέγει ένα αντικείμενο με πιθανότητα ανάλογη προς το βάρος του αντικειμένου. Δηλαδή αν τελικά το ρεύμα έχει $N$ αντικείμενα, και το συνολικό βάρος τους είναι $W = \sum_{i=1}^{N} w_i$ το αντικείμενο $i$ θα πρέπει να έχει πιθανότητα $w_i/W$ να επιλεγεί, για κάθε $1 \leq i \leq N$. Όπως και με τον κλασικό Reservoir Sampling αλγόριθμο, το $N$ δεν είναι γνωστό εκ των προτέρων και 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 $\boldsymbol{K}$ 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). (**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.
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 κείμενο. 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`”.