Πρώτη σειρά ασκήσεων Εξόρυξη Δεδομένων

Βασιλειάδης Μιλτιάδης 2944

Ερώτηση 1.

Α)

1.)Ο αλγόριθμος που περιεγράφηκε στην διάλεξη επιλέγει ένα αντικείμενο με πιθανότητα 1/k. Ζητήθηκε να μετατραπεί ο αλγόριθμος έτσι ώστε να επιλέγονται Κ αντικείμενα με πιθανότητα Κ/Ν. Για να το γίνει αυτό χρησιμοποιώ μία απλή λίστα μεγέθους Κ όπου θα αποθηκεύονται τα πρώτα Κ αντικείμενα όπως έρχονται από το stream κρατώντας έναν μετρητή των αντικειμένων που έχουν διαβαστεί από το stream. Για κάθε νέο αντικείμενο που έρχεται αυξάνω τον μετρητή κατόπιν επιλέγεται τυχαία ένας αριθμός ανάλογος των διαβασμένων μέχρι εκείνη την στιγμή αντικειμένων αν ο αριθμός αυτός είναι μεταξύ του 0 και του k-1 το αντικείμενο που μόλις διαβάστηκε αντικαθιστά το στοιχείο στην λίστα με index τον αριθμό που επιλέχτηκε τυχαία.

2.)Η απόδειξη του παραπάνω ισχυρισμού γίνεται με χρήση επαγωγής.

Αρχικά αν το stream περιέχει Κ αντικείμενα τότε τα αντικείμενα επιλέγονται με πιθανότητα $\frac{k}{k} = 1$. Το επόμενο αντικείμενο που θα έρθει $i = k + 1$ τότε αυτό το αντικείμενο θα έχει πιθανότητα να εκλεγεί ίση με $\frac{k}{k+1}$ Τα υπόλοιπα k αντικείμενα μέσα στο reservoir έχουν το καθένα πιθανότητα $\frac{1}{k+1}$ πιθανότητα να μην αντικατασταθούν και πιθανότητα $\frac{k-1}{k}$ ότι ένα από τα υπόλοιπα αντικείμενα στην λίστα θα αντικατασταθεί.

Προκύπτει αθροιστικά $\frac{1}{k+1} + \frac{k-1}{k} * \frac{k}{k+1} = \frac{k}{k+1}$

Υποθέτουμε ότι αυτό ισχύει για τα k+1,k+2….N αντικείμενα στο stream. Κάθε προηγούμενο αντικείμενο είχε πιθανότητα $\frac{k}{N}$ να διατηρηθεί. Για N=N+1 κάθε αντικείμενο είχε πιθανότητα $\frac{k}{N+1}$ να διατηρηθεί στην λίστα. Η πιθανότητα να μην επιλεγεί το τρέχον αντικείμενο είναι $\frac{N-k+1}{N+1}$ και η πιθανότητα να επιλεγεί είναι $\frac{k}{N+1}$ αλλά η πιθανότητα να επιλεγεί κάποιο άλλο αντικείμενο από την λίστα προς αντικατάσταση είναι $\frac{k-1}{k}$. Άρα η πιθανότητα υπολογίζεται παίρνοντας το άθροισμα $\frac{k}{N}\left(\frac{N-k+1}{N+1} + \frac{k}{N+1} * \frac{k-1}{k}\right) = \frac{k}{N}\left(\frac{N-k+1}{N+1} + \frac{k-1}{N+1}\right) = \frac{k}{N+1}$

Άρα για το αντικείμενο $i = n + 1$ που έρχεται η πιθανότητα να επιλεγεί $\frac{k}{N+1}$

Β)

Για να μετατρέψουμε τον αρχικό αλγόριθμο να κάνει επιλογή με βάση βάρη κρατάμε έναν μετρητή για το συνολικό βάρος. Αρχικά επιλέγεται το πρώτο αντικείμενο του stream και υπολογίζεται το βάρος του χρησιμοποιώντας γεννήτρια τυχαίων αριθμών αυτό το βάρος προστίθεται στο συνολικό. Για κάθε επόμενο στοιχείο που έρχεται υπολογίζουμε το βάρος

του με την χρήση της γεννήτριας τυχαίων αριθμών και προστίθεται στο συνολικό βάρος, επιλέγουμε έναν τυχαίο αριθμό αν ο αριθμός αυτός είναι μικρότερος από το πηλίκο $\frac{w_i}{\sum w}$ τότε αυτό το στοιχείο αντικαθιστά το προηγουμένως επιλεγμένο. Η απόδειξη είναι παρόμοια με την απόδειξη του Reservoir Sampling θα χρησιμοποιηθεί η επαγωγή.

Για το πρώτο στοιχείο η πιθανότητα επιλογής είναι ίση με $\frac{w_1}{\sum_1^1 w_i}$ = 1

Έστω για το k στοιχείο η πιθανότητα να επιλεγεί είναι $\frac{w_k}{\sum_{i=1}^k wi}$

Για το k στοιχείο η πιθανότητα να επιβιώσει για N-K επόμενα στοιχεία
$\frac{w_k}{\sum_{i=1}^k wi}\left(1-\frac{w_{k+1}}{\sum_{i=1}^k wi+ w_{k+1}}\right)\left(1-\frac{w_{k+2}}{\sum_{i=1}^{k+1} wi+ w_{k+2}}\right)\dots\left(1-\frac{w_n}{\sum_{i=1}^n wi}\right)=\frac{w_k}{W}$

Εστω ότι εχει εκλεγεί το N

Για N+1 η πιθανότητα να παραμείνει το N επιλεγμένο είναι

$$\frac{w_n}{W}\left(1-\frac{w_{n+1}}{W+w_{n+1}}\right)=\frac{w_n}{W}-\frac{w_{n+1}}{W+w_{n+1}}$$

Ερώτηση 2.

1) Υποθέτουμε ότι ο ισχυρισμός $C\mathrm{w}(f) \leq 3Cs(f)$ δεν ισχύει.
Στην ρόδα οι κόμβοι που ανήκουν στον κύκλο είναι συνδεδεμένοι με 2 άλλους γειτονικούς κόμβους που επίσης ανήκουν στον κύκλο και με τον ένα κόμβο που είναι στο κέντρο της ρόδας. Ο κεντρικός κόμβος κ του κύκλου προφανώς συνδέεται άμεσα με όλους τους υπολοίπους. Επειδή το d(u,v) είναι πάντα λιγότερο ακριβό από το u,v < - u,z + z,v λόγω τριγωνικής ανισότητας. Επίσης η ρόδα έχει διάμετρο 2 . Για να μεταβούμε από το u,v το πολύ να περάσουμε μία φορά από το κεντρικό κόμβο κ καθώς αυτός ενώνει με όλους τους υπόλοιπους. Άρα η παραπάνω υπόθεση καταλήγει σε άτοπο.

2) Επειδή οι αναθέσεις t και ο είναι οι βέλτιστες για το S και για το W αντίστοιχα τότε η απόσταση του u από τον v στον W με την ανάθεση t για το S θα είναι βέλτιστη για τις αποστάσεις που χρησιμοποιούν τον κεντρικό κόμβο, οποιαδήποτε μετάβαση μπορεί να γίνει βέλτιστα με το πολύ 2 hops το ένα περνά από το κέντρο του κύκλου κ εκτός αν ο u και v είναι γειτονικοί κόμβοι του κύκλου που μεταξύ τους ίσως να μην εχουν την βέλτιστη απόσταση. Η ανάθεση εγγυάται ότι οι αποστάσεις από κόμβο του C για τους άλλους κόμβους είναι βέλτιστες για τους 3 γειτονικούς και για τους υπολοίπους ισχύει η τριγωνική ανισότητα.
$C\ W\ (u) \leq 3C\ W\ (o)$

Οι αναφορές των 3 και 4 περιέχονται στα Jupyter Notebooks