1η Ανάθεση για το Σπίτι (σελ. 1 από 2)

Algorithms for Big Data

(1) Έστω ότι μελετάμε εκτιμήσεις της Jaccard Similarity μεταξύ 0/1-διανυσμάτων **X**,**Y** $\in \{0,1\}^{nx1}$ χρησιμοποιώντας ένα σύνολο από διαφορετικές μεταξύ τους μεταθέσεις των γραμμών τους (δηλαδή, του { 0, 1, 2, …, n-1 }). Συγκεκριμένα, η εκτίμηση που επιστρέφουμε είναι το **ποσοστό** των μεταθέσεων όπου οι MinHash τιμές των δύο στηλών συμπίπτουν. Έστω το 4x2 μητρώο **M** = [0 , 0 ; 0 , 1 ; 1 , 0 ; 1 , 1], όπου τα κόμματα διαχωρίζουν στοιχεία εντός της ίδιας γραμμής, ενώ τα ερωτηματικά διαχωρίζουν γραμμές. Απαριθμούμε τις γραμμές του **M** ως: **0** = [0,0], **1** = [0,1], **2** = [1,0], **3** = [1,1].

(1α) Να υπολογιστεί η MinHash τιμή κάθε στήλης, για ΟΛΕΣ (πόσες είναι?) τις μεταθέσεις των γραμμών του **M**.

(1β) Να υπολογιστεί η εκτίμηση της Jaccard Similarity των δυο στηλών του **M**, που προκύπτει αν λάβουμε υπόψη μας ΟΛΕΣ τις μεταθέσεις των γραμμών του **M**.

(1γ) Να υπολογιστεί η εκτίμηση της Jaccard Similarity που προκύπτει αν λάβουμε υπόψη μας ΜΟΝΟ (όλες) τις κυκλικές μεταθέσεις των γραμμών, δηλαδή εκείνες τις μεταθέσεις που επιλέγουν (με όλους τους δυνατούς τρόπους) το πρώτο στοιχείο (γραμμή) r$\in${**0**,**1**,**2**,**3**} της μετάθεσης, και τα επόμενα στοιχεία είναι το (r+1) **mod** 4, (r+2) **mod** 4 και (r+3) **mod** 4. Πχ, μια τέτοια μετάθεση είναι η **3**-**0**-**1**-**2**, αλλά όχι η **3**-**1**-**0**-**2**.

(1δ) Να αποδείξετε ότι παίρνοντας ΟΛΕΣ τις μεταθέσεις ενός nx2 μητρώου **M** $\in \{0,1\}^{nx2}$ και εκτιμώντας τη Jaccard Similarity, sim($\mathbf{M_1}$,$\mathbf{M_2}$), σύμφωνα με το **1α**, καταλήγουμε στην πραγματική τιμή της Jaccard Similarity (δλδ, το σφάλμα εκτίμησης είναι μηδενικό).

1η Ανάθεση για το Σπίτι (σελ. 2 από 2)

(2) Να κατασκευαστεί πρόγραμμα (σε python, ή σε c/c++) που να εκτελεί τις εξής εργασίες:

(2α) Ως είσοδο δέχεται:

(ι) Όνομα αρχείου κειμένου (TXT) της μορφής:

```
0 0 1 ... 0
1 1 0 ... 0
   ...
0 1 0 ... 0
```

που περιγράφει ένα KxN 0/1-μητρώο **A** (κάθε γραμμή του αρχείου περιγράφει και μια διαφορετική γραμμή του **A**). Θα πρέπει, κατά την ανάγνωση του αρχείου, να υπολογιστούν οι διαστάσεις K,N του **A** (δεν είναι γνωστές εκ των προτέρων).

(ιι) Φυσικό αριθμό n (το πλήθος των μεταθέσεων γραμμών που θέλουμε να χρησιμοποιήσουμε για τις υπογραφές μας). Θεωρήστε ως τυπική τιμή το n=10.

(2β) Θα εκτελεί τη **MinHashing** τεχνική για δημιουργία (και εκτύπωση στην οθόνη ή/και σε αρχείο TXT) του nxN μητρώου υπογραφών **S**, όπου κάθε υπογραφή είναι μια στήλη με τιμές από το σύνολο { 0, 1, ... , K-1 }. Θεωρούμε ότι $n \leq 22$. Η υλοποίηση θα πρέπει να γίνει στη λογική που περιγράφει η ενότητα 3.3.5 του βιβλίου: Αντί για n τυχαίες μεταθέσεις, θα χρησιμοποιηθούν n συναρτήσεις κατακερματισμού της μορφής $\phi(X) = \alpha*X+\beta$ **mod** 23, για τυχαία επιλεγμένα (ένα ανά συνάρτηση κατακερματισμού) ζεύγη φυσικών αριθμών $\alpha,\beta \in \{1,2,3,...,22\}$.

(2γ) Θα υπολογίζει (και θα τυπώνει στην οθόνη ή/και σε αρχείο) τις **ποσοστά ομοιότητας** για (όλα, ή επιλεγμένα που θα ορίζει ο χρήστης) ζεύγη στήλών από το **S**, αλλά και τιμές **Jaccard-ομοιότητας** για τα αντίστοιχα ζεύγη στηλών του **A**.