ASSIGNMENT 2

STUDENT: ΑΝΤΩΝΙΟΣ ΠΡΟΜΠΟΝΑΣ

ID: 03400232

1. ΑΣΚΗΣΗ 1

Τα 4 τελευταία ψηφία του ακαδημαϊκού μητρώου μου είναι 0232.

Άρα, A=02 --> A=2 και B=32. Ωστόσο, το 32 υπάρχει ήδη στη λίστα των δοσμένων αριθμών και έπειτα από υπόδειξη της εκφώνησης της άσκησης, στη τιμή B αναθέτω τη τιμή 26(B=26).

Το σύνολο των αριθμών που πρέπει να διαχειριστούμε είναι 28.Η άσκηση μας ζητάει να δημιουργήσουμε κάδους με βάθος 4. Αυτό σημαίνει ότι πρέπει κάθε κάδος περιέχει 4 τιμές. Άρα, πρέπει να δημιουργήσουμε 28/4=7 κάδους.

Πρώτα απ΄ όλα όμως, πρέπει να ταξινομήσουμε τις τιμές αυτές. Έπειτα, από την ταξινόμηση οι τιμές εμφανίζονται με αυτή τη σειρά:

2,13,14,18,23,26,27,29,32,36,41,44,46,55,60,61,62,64,65,73,75,78, 80,84,87,93,96,99

Το επόμενο βήμα είναι να διαμερίσω ομοιόμορφα τις τιμές στους 7 κάδους.

Bin 1: 2,13,14,18

Bin 2:23,26,27, 29

Bin 3: 32,36,41,44

Bin 4:46,55, 60,61

Bin 5: 62,64,65,73

Bin 6: 75,78,80,84

Bin 7: 87,93,96,99

Τώρα, είμαστε έτοιμοι να εφαρμόσουμε ‘smoothing by bin means’ και ‘smoothing by bin boundaries’.

i. Smoothing by bin means

Για να εφαρμόσω τη μέθοδο αυτή πρέπει σε κάθε κάδο να υπάρχουν 4 τιμές , οι οποίες είναι ο μέσος όρος των τιμών του κάδου. Συγκεκριμένα:

Μέσος όρος του κάδου 1: (2+13+14+18)/4= 47/4 = 11.75 ~= 12

Μέσος όρος του κάδου 2: (23+26+27+29)/4= 105/4 = 26.25 ~= 26

Μέσος όρος του κάδου 3: (32+36+41+44)/4= 153/4 = 38.25 ~= 38

Μέσος όρος του κάδου 4: (46+55+60+61)/4= 222/4 = 55.5 ~= 56

Μέσος όρος του κάδου 5: (62+64+65+73)/4= 264/4 = 66

Μέσος όρος του κάδου 6: (75+78+80+84)/4= 317/4 = 79.25 ~= 79

Μέσος όρος του κάδου 7: (87+93+96+99)/4= 375/4 = 93.75 ~= 94

Συνεπώς:

Smoothing by bin means:

Bin 1: 12,12,12,12

Bin 2: 26,26,26,26

Bin 3: 38,38,38,38

Bin 4: 56,56,56,56

Bin 5: 66,66,66,66

Bin 6: 79,79,79,79

Bin 7: 94,94,94,94

ii. Smoothing by bin boundaries

Στη μέθοδο αυτή , παίρνω τα 2 άκρα τιμών του κάδου και υπολογίζω τη μέση τιμή τους. Οι τιμές του κάδου από το ένα άκρο του κάδου μέχρι το ακέραιο μέρος της μέσης τιμής, προσωποποιούνται με το 1ο άκρο του κάδου και οι υπόλοιπες τιμές προσωποποιούνται με το 2ο άκρο του κάδου. Συγκεκριμένα:

Bin 1: Τα άκρα είναι: 2,18. Η μέση τιμή είναι (2+18)/2=20/2=10. Άρα για τιμές από [2,10], οι τιμές γίνονται 2 και για τιμές από [11,18] οι τιμές γίνονται 18.

Bin 2: Τα άκρα είναι: 23,29. Η μέση τιμή είναι (23+29)/2=52/2=26. Άρα για τιμές από [23,26], οι τιμές γίνονται 23 και για τιμές από [27,29] οι τιμές γίνονται 29.

Bin 3: Τα άκρα είναι: 32,44. Η μέση τιμή είναι (32+44)/2=76/2=38. Άρα για τιμές από [32,38], οι τιμές γίνονται 32 και για τιμές από [39,44] οι τιμές γίνονται 44.

Bin 4: Τα άκρα είναι: 46,61. Η μέση τιμή είναι (46+61)/2=107/2=53.5. Άρα για τιμές από [46,53], οι τιμές γίνονται 46 και για τιμές από [54,61] οι τιμές γίνονται 61.

Bin 5: Τα άκρα είναι: 62,73. Η μέση τιμή είναι (62+73)/2=135/2=67.5. Άρα για τιμές από [62,67], οι τιμές γίνονται 62 και για τιμές από [68,73] οι τιμές γίνονται 73.

Bin 6: Τα άκρα είναι: 75,84. Η μέση τιμή είναι (75+84)/2=159/2=79.5. Άρα για τιμές από [75,79], οι τιμές γίνονται 75 και για τιμές από [80,84] οι τιμές γίνονται 84.

Bin 7: Τα άκρα είναι: 87,99. Η μέση τιμή είναι (87+99)/2=186/2=93. Άρα για τιμές από [87,93], οι τιμές γίνονται 87 και για τιμές από [94,99] οι τιμές γίνονται 99.

Συνεπώς:

Smoothing by bin boundaries:

Bin 1: 2,18,18,18

Bin 2: 23,23,29,29

Bin 3: 32,32,44,44

Bin 4: 46,61,61,61

Bin 5: 62,62,62,73

Bin 6: 75,75,84,84

Bin 7: 87,87,99,99

2. ΑΣΚΗΣΗ 2

Το συγκεκριμένο bloom filter έχει 20 bits και τη χρονική στιγμή t, έχει τη συγκεκριμένη μορφή:

[1,0,1,0,0,1,1,0,1,0,0,1,0,0,1,0,1,1,0,0]

Έχουμε 2 hash functions.

Ο Α.Μ. ειναι : 03400232. Τα τελευταία 2 ψηφία είναι 32. Άρα, Α=3 και Β=2.

1.h1(x)=(3x+11)mod20

2.h2(x)=(2x+2)mod20

i. Για y=8:

h1(8)=(3*8+11)mod20=(24+11)mod20=35mod20=15

h1(8)=20

h2(8)=(2*8+2)mod20=(16+2)mod20=18

h2(8)=18

Άρα, ψάχνουμε στη θέση 15 και 18 του bloom filter. Και στις 2 θέσεις υπάρχει μονάδα. Άρα, υπάρχει πολύ μεγάλη πιθανότητα το στοιχείο y=8, να έχει περάσει από το stream.

ii. Τώρα, θα υπολογίσουμε τη false positive πιθανότητα για n1=15 και για n2=18. Ο τύπος υπολογισμού της false positive πιθανότητας είναι :

F=(1-e^(-nw/m))^w, όπου w είναι ο αριθμός των hash functions και w , ο αριθμός των bits στο bloom filter. Άρα, m=20 και w=2

Η false positive πιθανότητα για n1=15 είναι:

F=(1-e^(-2*15/20))^2=(0.77)^2=0.6

Η false positive πιθανότητα για n1=18 είναι:

F=(1-e^(-2*18/20))^2=(0.83)^2=0.69

Συνεπώς, υπάρχει 60% πιθανότητα να υπάρχει false positive στην 15η θέση του stream και 69% πιθανότητα να υπάρχει false positive στην 18η θέση του stream.

iii. Τη χρονική στιγμή t+1, εισάγουμε στο bloom filter την τιμή x=13. Συγκεκριμένα, υπολογίζουμε μέσω των hash functions τις αντίστοιχες θέσεις-τιμές του bloom filter .

h1(13)=(3*13+11)mod20=(39+11)mod20=50mod20=10

h1(13)=10

h2(13)=(2*13+2)mod20=(26+2)mod20=28mod20=8

h2(13)=8

Το bloom filter πριν τη χρονική στιγμή t+1 έχει τη μορφή:
[1,0,1,0,0,1,1,0,1,0,0,1,0,0,1,0,1,1,0,0]

Στη θέση 8 , υπάρχει 0 , το οποίο μετατρέπεται σε 1. Το ίδιο ισχύει και στη θέση 10. Άρα, το bloom filter τη χρονική στιγμή t+1 έχει τη μορφή:

[1,0,1,0,0,1,1,1,1,1,0,1,0,0,1,0,1,1,0,0]

iv. Ένα bloom filter μπορεί να υπολογίσει κατά προσέγγιση το N αριθμό διακριτών στοιχείων που έχουν περάσει από το stream. Επειδή, ο αριθμός των bits με τιμή 0 είναι μικρότερος του 20, N=m*ln(m/m0)/w, όπου m0 το πλήθος των 0 bits στο bloom filter , w το πλήθος των hash functions και m το πλήθος των bits στο bloom filter. Άρα, m0=9, w=2 και m=20.

N=20ln(20/9)/2=10*0.79=7.9~=8

Άρα, τη χρονική στιγμή t+2, θα έχουν περάσει από το stream, 8 στοιχεία.

Μία άλλη σκέψη για το πόσοι αριθμοί έχουν περάσει από το bloom filter είναι να μετρήσουμε το πλήθος του αριθμού “1” που υπάρχει στο bloom filter. Η αρχική μορφή του συγκεκριμένου bloom filter είναι και στις 20 θέσεις 0. Ωστόσο, τώρα υπάρχουν 11 “1”, που σημαίνει ότι είναι πολύ πιθανό να έχουν περάσει 11/2(hash functions)~=6 αριθμοί από αυτό , συν κάποια πιθανά collisions. Συνεπώς, ο αριθμός 8 που υπολόγισα παραπάνω φαίνεται αρκετά ρεαλιστικός.

3. ΑΣΚΗΣΗ 3

Ο Α.Μ. είναι 02300232. Το τελευταίο ψηφίο είναι το 2. Άρα, ο πίνακας κατάστασης μετάβασης είναι:

	s1	s2
s1	0.2	0.8
s2	0.2	0.8

Η πιθανότητα αρχικής κατάστασης είναι:

s1	s2
0	1

Η πιθανότητα να δημιουργηθεί το εκάστοτε σύμβολο δίνεται από τον παρακάτω πίνακα:

	a	b	c
s1	0.5	0.5	0
s2	0.5	0	0.5

Το πρώτο πράγμα που καταλαβαίνουμε εύκολα, με βάση τα παραπάνω δεδομένα είναι ότι υπάρχει 0% πιθανότητα να ξεκινήσουμε από την κατάσταση s1 και 100% να ξεκινήσουμε από την κατάσταση s2. Από την s2, υπάρχει 80% πιθανότητα να μεταβούμε στην s2 και 20% πιθανότητα να μεταβούμε στην s1. Το πρώτο στοιχείο που θέλουμε να μεταβούμε είναι το c, το οποίο έχει 50% και 0 % πιθανότητα από s2 και s1 αντίστοιχα. Άρα, υπάρχει μοναδικός τρόπος το c να είναι το πρώτο στοιχείο της συμβολοσειράς. Ο τρόπος αυτός είναι ξεκινώντας από s2, παραμένοντας σε s2 και από

εκεί μεταβαίνοντας στο c. Αυτό έχει πιθανότητα: 0.8*0.5=0.4 (40%).

Για τη συμβολοσειρά που θέλουμε να υπολογίσουμε τη πιθανότητα ύπαρξης, το 2ο στοιχείο είναι το b.

Από το σημείο που βρισκόμαστε τώρα, ένας τρόπος είναι να ξαναγυρίσουμε στο s2 και από εκεί στο b. Ωστόσο, η πιθανότητα από s2 να μεταβούμε στο b είναι 0. Συνεπώς, μόνο από s1 μπορούμε να μεταβούμε στο b. Άρα, η διαδρομή είναι από s2 σε s1 και από s1 σε b. Η συγκεκριμένη διαδρομή έχει πιθανότητα 0.2*0.5=0.1. Τη συγκεκριμένη πιθανότητα την πολλαπλασιάζουμε με την ήδη υπάρχον πιθανότητα. Άρα, η πιθανότητα τα πρώτα 2 στοιχεία της συμβολοσειράς να είναι {cb}, είναι: 0.4*0.1=0.04.

Αυτή τη στιγμή λοιπόν, βρισκόμαστε στο στοιχείο b της s1, με πιθανότητα 0.04. Το τελευταίο στοιχείο της συμβολοσειράς που θέλουμε να υπολογίσουμε τη πιθανότητα δημιουργίας είναι το a. Στο a μπορούμε να μεταβούμε είτε από s1 είτε από s2 ως εξής.

s1: 0.04*0.2*0.5=0.004

s2:0.04*0.8*0.5=0.016

P(cba)=0.004+0.016=0.02.

Συνεπώς, υπάρχει 2% πιθανότητα να δημιουργηθεί η συμβολοσειρά {cba}.