HW1 Solutions (Data Mining)

Problem 1

a Since everyone gets to a hotel in 100 days,
Pr(a person visits a hotel on any day)=0.01
Pr(2 particular persons visit a hotel on any day)= 0.01 * 0.01 = 10^{-4}
Since, there are 10^5 hotels,
Pr(2 particular persons visit same hotel on any day)= $10^{-4}/10^5 = 10^{-9}$
Pr(2 particular persons visit same hotel on two different days) =$(10^{-9})^2 = 10^{-18}$
No of possible pairs of persons = ${}^{10^9}C_2 = 5 * 10^{17}$ (approx.) **
No of possible pairs of days = ${}^{2000}C_2 = 2 * 10^6$(approx.) **

Expected no. of evil doer pairs =Pr(2 particular persons visit same hotel on two different days) * No. of Possible pairs of persons * No of possible pairs of days =$10^{-18} * 5 * 10^{17} * 2 * 10^6 = 10^6$

** ${}^nC_2$ is approximately equal to $\frac{n^2}{2}$ when n is sufficiently large

b Since everyone gets to a hotel in 100 days,
Pr(a person visits a hotel on any day)=0.01
Pr(2 particular persons visit a hotel on any day)= 0.01 * 0.01 = 10^{-4}
Since, there are $2 * 10^5$ hotels,
Pr(2 particular persons visit same hotel on any day)= $10^{-4}/(2*10^5) = 5*10^{-10}$
Pr(2 particular persons visit same hotel on two different days) =$(5 * 10^{-10})^2 = 2.5 * 10^{-19}$
No of possible pairs of persons = ${}^{2*10^9}C_2 = 2 * 10^{18}$ (approx.) **
No of possible pairs of days = ${}^{1000}C_2 = 5 * 10^5$(approx.) **

Expected no. of evil doer pairs =Pr(2 particular persons visit same hotel on two different days) * No. of Possible pairs of persons * No of possible pairs of days =$2.5 * 10^{-19} * 2 * 10^{18} * 5 * 10^5 = 2.5 * 10^5$

** ${}^nC_2$is approximately equal to $\frac{n^2}{2}$when n is sufficiently large

Problem 2 For any numbers a and b, a mod b $=a-(b*quotient(a,b))$
where quotient(a,b) $=\frac{a}{b}$ (integral division as in java)
Let gcd(a,b) denote the greatest common divisor of a and b.
The above equation can be written as,
a mod b $=gcd(a,b)*(\frac{a}{gcd(a,b)}-\frac{b}{gcd(a,b)}*quotient(a,b))$
Now back to our problem,
If $gcd(c,15)>1$, then $gcd(x,15)>1$
h(x) = x mod 15 = $gcd(x,15)*(\frac{x}{gcd(x,15)}-\frac{15}{gcd(x,15)}*quotient(x,15))$
Therefore h(x) will be a multiple of gcd(x,15)
eg. for c=3 , gcd(c,15) =3
Possible values of h(x) – 0, 3, 6, 9, 12
eg. for c=20, gcd(c,15) =5
Possible values of h(x) – 0,5,10

So, if$gcd(c,15)>1$, then h(x) won't take the values that are not multiples of gcd(c,15) Therefore, gcd(c,15) should be 1 if we want uniform hashing

Appropriate values of c:-1,2,4,7,8,11,13,14,16,17,19,22,.....

Problem 3

a If MinHash signature has size r,
Probablity that $I_1 and I_2$ have same signature in a hash table= x^r
Probablity that $I_1 and I_2$ do not have same signature in a hash table = $1-x^r$
Probablity that $I_1 and I_2$ do not have same signature in any of b hash tables = $(1-x^r)^b$
Probablity that $I_1 and I_2$ have same signature in at least one hash table (out of b) = $1-(1-x^r)^b$

b Using following approximation formula from lecture 1, if $a << 1$
$(1-a)^b = ((1-a)^{\frac{1}{a}})^{a*b} = e^{-a*b}(approx)$
$Since, x^r << 1,$
$(1-x^r)^b = e^{-x^r*b}(approx)$

Now back to the problem,
$1–(1–x^r)^b = \frac{1}{2}$
$(1–x^r)^b = \frac{1}{2}$
$e^{-x^r*b} = \frac{1}{2}$
$-x^r*b = ln(\frac{1}{2})$
$-x^r*b = -ln(2)$
$b = \frac{ln(2)}{x^r}$

Problem 4

a Items 1-20
An item i can be in basket numbers $i, 2*i, 3*i, 4*i, 5*i, 6*i, , , , ,$
Let's consider item number 21,
Since $5*21 > 100$, item 21 can only appear in 4 baskets
So, items from 1 to 20 appear in atleast 5 baskets and are therefore frequent

b Let the two items in the pair be x and y.
Let LCM(x,y) be the least common multiple of x and y
Items x and y both exist in these baskets - (LCM(x,y) , 2*LCM(x,y), 3*LCM(x,y), 4*LCM(x,y) , 5*LCM(x,y))
If $LCM(x, y) <= 20$, then $5 * LCM(x, y) <= 100$
Therefore, if $LCM(x, y) <= 20$, pair (x,y) is present in at least 5 baskets and is frequent
Frequent pairs:
(1, 2) (1, 3) (1, 4) (1, 5) (1, 6) (1, 7) (1, 8) (1, 9) (1, 10) (1, 11) (1, 12) (1, 13) (1, 14) (1, 15) (1, 16) (1, 17) (1, 18) (1, 19) (1, 20) (2, 3) (2, 4) (2, 5) (2, 6) (2, 7) (2, 8) (2, 9) (2, 10) (2, 12) (2, 14) (2, 16) (2, 18) (2, 20) (3, 4) (3, 5) (3, 6) (3, 9) (3, 12) (3, 15) (3, 18) (4, 5) (4, 6) (4, 8) (4, 10) (4, 12) (4, 16) (4, 20) (5, 10) (5, 15) (5, 20) (6, 9) (6, 12) (6, 18) (7, 14) (8, 16) (9, 18) (10, 20)

c Let's consider a basket numbered n.
Number of items in n will be equal to the number of factors of n
So, the largest basket will be the basket number having maximum number of factors

It can be seen that the following numbers have 12 factors each
$60 - (1, 2, 3, 4, 5, 6, 10, 12, 15, 20, 30, 60)$
'$72 - (1, 2, 3, 4, 6, 8, 9, 12, 18, 24, 36, 72)$
$84 - (1, 2, 3, 4, 6, 7, 12, 14, 21, 28, 42, 84)$
$90 - (1, 2, 3, 5, 6, 9, 10, 15, 18, 30, 60, 90)$
$96 - (1, 2, 3, 4, 6, 8, 12, 16, 24, 31, 48, 96)$

12 is the maximum number of factors for numbers from 1 - 100
So, the baskets with maximum items are 60,72,84,90,96

d $Confidence(5,7->2) = \frac{Support(5,7,2)}{Support(5,7)}$
LCM(5,7,2)=70
LCM(5,7) = 35
i=1 is maximum value such that $i * LCM(5,7,2) <= 100$
i=2 is maximum value such that $i * LCM(5,7) <= 100$
Therefore,
Support(5,7,2)= 1
Support(5,7)=2
$Confidence(5,7->2) = \frac{1}{2}$

$Confidence(2,3,4->5) = \frac{Support(2,3,4,5)}{Support(2,3,4)}$
LCM(2,3,4,5)= 60
LCM(2,3,4) = 12
i=1 is maximum value such that $i * LCM(2,3,4,5) <= 100$
i=8 is maximum value such that $i * LCM(2,3,4) <= 100$
Therefore,
Support(2,3,4,5)= 1
Support(2,3,4)=8
$Confidence(2,3,4->5) = \frac{1}{8}$