Data Mining Homework 1

Fatima AlSaadeh-fya7

fatima.alsaadeh@rutgers.edu

1 Map-Reduce

src/main/java/question$_1$*.java*

$MapReduceAlgorithmRecommendsthetop10friendsbasedonthenumberofmutualfriends$
$TheAlgorithm$:

Map Procedure:
- Loop over all the users list:
- Pick the user A:
- Loop over the user A friends:
- Pair the user A to the already friend users and 0: Key A, Pair (friend1, 0)
- Loop over each friend of user A friends and Pair them together with 1
which means they have one mutual friend A: Key: friend1, Pair (friend2,1)

The resulted pairs will be shuffled and sorted, combined by the similar keys and sent to the produce

Produce Procedure:
- Get a key with all its pairs
- Loop over these pairs:
- If the pair with count 0, ignore it because it means the user and suggested are already friends
- If the pair with count=1, if it was counted before increase the counter,
else add it with count 1
- Will end up for key A : (friend1,1), (friend2,4), (friend3,9), where friend1 is suggested and count is how many mutual friends between them
- Sort the friends by their mutual friends count and pick the top 10

Output Examples:

924 439,2409,6995,11860,15416,43748,45881
8941 8938,8942,8946,8939,8943,8944,8945,8940
8942 8938,8939,8941,8945,8946,8940,8943,8944
9019 320,9018,9016,9017,9020,9021,9022,317,9023
9020 9021,320,9016,9017,9018,9019,9022,317,9023
9021 9020,320,9016,9017,9018,9019,9022,317,9023
9022 9019,9020,9021,317,320,9016,9017,9018,9023
9990 9987,9988,9989,9993,9994,35667,9991,9992,13134,13478
9992 9987,9989,9988,9990,9993,9994,35667,9991
9993 9990,9994,9987,9988,9989,9991,35667,9992,13134,13478

2 Association Rules

src/main/java/question$_2$.*java*
(*a*)**A drawback of using confidence is that it ignores Pr(B). Why is this a drawback? Explain w**
$conf(A \to B) = P(B|A) = P(BandA)/P(A)$Confidence ignoring P(B) will give a misleading rule because the

$$conf(A \to B)$$

might be high because it happens that product A and B to occur together very often (high P(A and B)) or B occurs very often. For example if

$$P(milkandegg) = .15$$

$$P(milk) = .2$$

$$P(egg) = .9$$

$$conf(milk \to egg) = .75$$

but if P(egg)=.9 this means it appears in the basket very often and this rule is misleading where

$$conf(notmilk \to egg)$$

can be high as well Support(B) which is used in both lift and conviction it is using S(B) which is the probability of B divided by the number of baskets giving the real numbers.
(b)**A measure is symmetrical if**

$$conf(A \to B) = conf(B \to A)$$

Which of the measures presented here are symmetrical? For each measure, please provide either a proof that the measure is symmetrical, or a counterexample that shows the measure is not symmetrical.
Confidence : this measure isn't symmetric Having :

$$conf(A \to B) = P(B|A) = P(B \cap A)/P(A)$$

$$conf(B \rightarrow A) = P(A|B) = P(A \cap B)/P(B)$$

Unless P(A) = P(B) -which is very rarely to happen- these two values won't be equal. For the example above the probability of occurrence of eggs in the basket if the basket already contains milk isn't equal to the probability of occurrence of eggs in the basket if the basket already contains milk

Lift : this measure is symmetric, the proof mathematically Having :

$$lift(A \rightarrow B) = \frac{conf(A \rightarrow B)}{P(B)} = \frac{P(B \cap A)}{P(A)} * \frac{1}{P(B)}$$

$$lift(B \rightarrow A) = \frac{conf(B \rightarrow A)}{P(A)} = \frac{P(A \cap B)}{P(B)} * \frac{1}{P(A)}$$

And these values are equal.

Conviction : this measure is not symmetric, the proof mathematically again Having :

$$Conviction(A \rightarrow B) = \frac{1 - P(B)}{1 - conf(A \rightarrow B)}$$

$$Conviction(B \rightarrow A) = \frac{1 - P(A)}{1 - conf(B \rightarrow A)}$$

And these values are not equal. A and B are independent if the conviction value is high and the conf is small.

(c)A measure is desirable if its value is maximal for rules that hold 100 of the time (such rules are called perfect implications). This makes it easy to identify the best rules. Which of the above measures have this property? Explain why. Confidence is desired because it will be equal 1 if A and B always occurs together. Lift is not equal for every time A and B appears together because it depends on the probability of B as well, it will be maximal if the confidence is equal to the P(B). Conviction is desired, it's the inverse of the lift value and when it's equal to 1 it always mean that A doesn't relate to B.

(d)Identify pairs of items (X, Y) such that the support of X, Y is at least 100. For all such pairs, compute the confidence scores of the corresponding association rules: X Y, Y X. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Break ties, if any, by lexicographically increasing order on the left hand side of the rule.

$$[DAI93865, FRO40251] = 1.0$$

$$[FRO40251, DAI93865] = 0.053594434424117494$$

$$[GRO85051, FRO40251] = 0.999176276771005$$

$$[FRO40251, GRO85051] = 0.0$$

$$[GRO38636, FRO40251] = 0.9906542056074766$$

$$[FRO40251, GRO38636] = 0.027312548312290647$$

$$[ELE12951, FRO40251] = 0.9905660377358491$$

$$[FRO40251, ELE12951] = 0.027054882762174697$$

$$[DAI88079, FRO40251] = 0.9867256637168141$$

$$[FRO40251, DAI88079] = 0.11491883535171347$$

(e) Identify item triples (X, Y, Z) such that the support of X, Y, Z is at least 100. For all such triples, compute the confidence scores of the corresponding association rules: (X, Y) Z, (X, Z) Y, and (Y, Z) X. Sort the rules in decreasing order of confidence scores and list the top 5 rules in the writeup. Order the left-hand-side pair lexicographically and break ties, if any, by lexicographical order of the first then the second item in the pair.

$$[DAI23334, ELE92920, DAI62779] = 1.0$$

$$[DAI23334, DAI62779, ELE92920] = 0.5238095238095238$$

$$[ELE92920, DAI62779, DAI23334] = 0.1630558722919042$$

$$[DAI31081, GRO85051, FRO40251] = 1.0$$

$$[DAI31081, FRO40251, GRO85051] = 0.36428571428571427$$

$$[GRO85051, FRO40251, DAI31081] = 0.08408903544929926$$

$$[DAI55911, GRO85051, FRO40251] = 1.0$$

$$[DAI55911, FRO40251, GRO85051] = 0.5732758620689655$$

$$[GRO85051, FRO40251, DAI55911] = 0.10964550700741962$$

$$[DAI62779, DAI88079, FRO40251] = 1.0$$

$$[DAI62779, FRO40251, DAI88079] = 0.10934579439252337$$

$$[DAI88079, FRO40251, DAI62779] = 0.2623318385650224$$

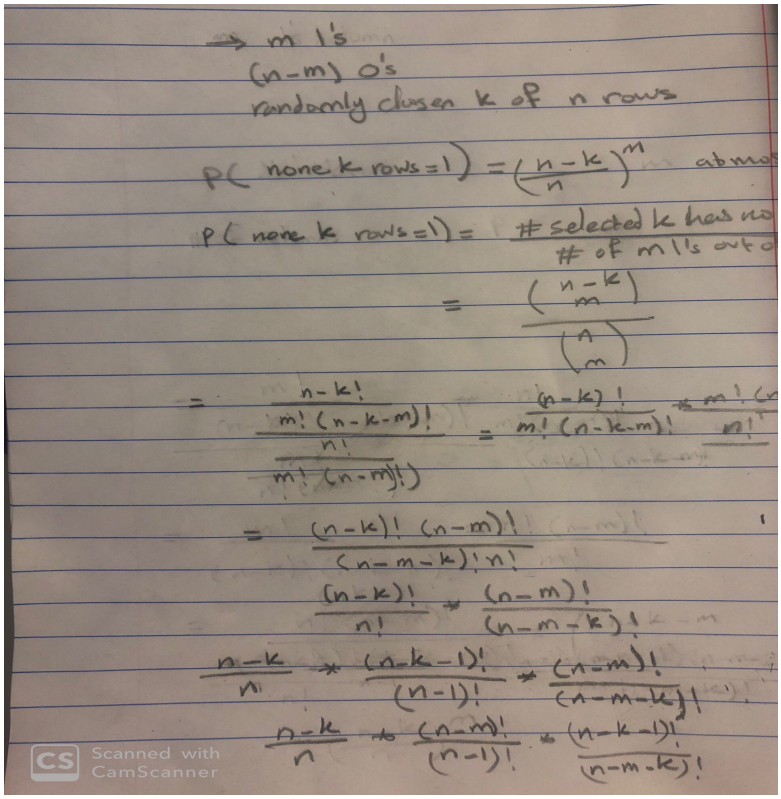
$$[DAI75645, GRO85051, FRO40251] = 1.0$$

$$[DAI75645, FRO40251, GRO85051] = 0.3149920255183413$$

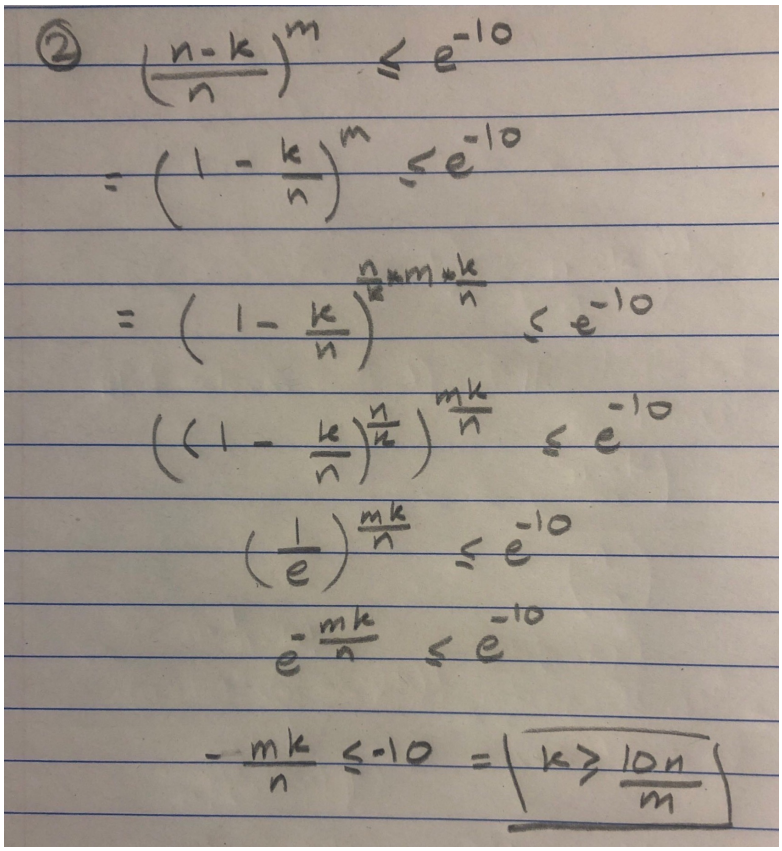
$$[GRO85051, FRO40251, DAI75645] = 0.325638911788953$$

3 Locality-Sensitive Hashing

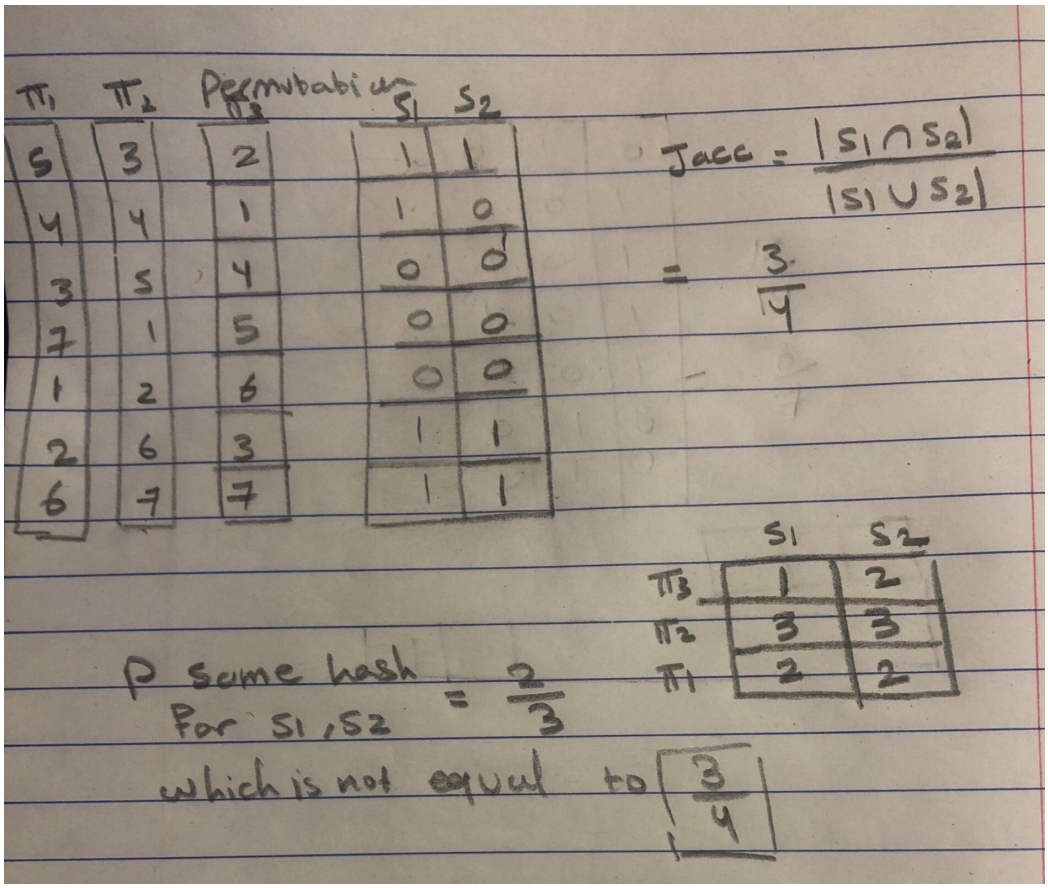
(a)Suppose a column has m 1's and therefore (n-m) 0's. Prove that the probability we get "don't know" as the min-hash value for this column is at most $(n - k/n) * m$.



(b) Suppose we want the probability of "don't know" to be at most e^-10 Assuming n and m are both very large (but n is much larger than m or k), give a simple approximation to the smallest value of k that will assure this probability is at most e^-10 Hints: (1) You can use $(n-k/n)*m$ as the exact value of the probability of "don't know." (2) Remember that for large x, $(1 - 1/X)^x 1/e$



(c) Give an example of two columns such that the probability (over cyclic permutations only) that their min-hash values agree is not the same as their Jaccard similarity. In your answer, please provide (a) an example of a matrix with two columns (let the two columns correspond to sets denoted by S1 and S2) (b) the Jaccard similarity of S1 and S2, and (c) the probability that a random cyclic permutation yields the same min-hash value for both S1 and S2.


π1 π2 Permutation π3 S1 S2
5 3 2 1 1
4 4 1 1 0
3 5 4 0 0
7 1 5 0 0
1 2 6 0 0
2 6 3 1 1
6 7 7 1 1
Jacc = |S1 ∩ S2| / |S1 ∪ S2|
= 3/4
S1 S2
π3 1 2
π2 3 3
π1 2 2
P same hash For S1, S2 = 2/3
which is not equal to 3/4

References