CS550: Massive Data Mining and Learning Spring 2018

Problem Set 1

Due 11:59pm Sunday, February 25, 2018

Only one late period is allowed for this homework (11:59pm Monday 2/26)

**Submission Instructions**

**Assignment Submission**: Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Sakai. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

**Late Day Policy**: Each student will have a total of **two** free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

**Honor Code**: Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

I acknowledge and accept the Honor Code.

(Signed) \_\_\_\_FT\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

If you are not printing this document out, please type your initials above.

**Answer to Questions 1**

ii)During the mapping stage:

I checked every user and pair their existed friends. The structure of the key-value pair is (ID, (possible suggested friend,1)). The existed friends are grouped in the structure (ID,(existed friend, -1)).

While during reducing, all the pairs with same keys will be put together and then I looped the all values and maintain a hash map to count the same ids of suggested friends, omit the existed friends. Then I sorted the map and output the top ten recommendations.

iii)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| user | Re1 | Re2 | Re3 | Re4 | Re5 | Re6 | Re7 | Re8 | Re9 | Re10 |
| 924 | 439 | 2409 | 6995 | 11860 | 15416 | 43478 | 45881 |  |  |  |
| 8491 | 8493 | 8488 | 8489 | 8492 | 8494 | 8499 | 8501 | 8503 | 8504 |  |
| 8492 | 8939 | 8940 | 8943 | 8944 |  |  |  |  |  |  |
| 9019 | 9022 | 317 | 9023 |  |  |  |  |  |  |  |
| 9020 | 9021 | 9016 | 9017 | 9022 | 317 | 9023 |  |  |  |  |
| 9021 | 9020 | 9016 | 9017 | 9022 | 317 | 9023 |  |  |  |  |
| 9022 | 9019 | 9020 | 9021 | 317 | 9016 | 9017 | 9023 |  |  |  |
| 9990 | 13134 | 13478 | 13877 | 34299 | 34485 | 34642 | 37941 |  |  |  |
| 9992 | 9987 | 9989 | 35667 | 9991 |  |  |  |  |  |  |
| 9993 | 9991 | 13134 | 13478 | 13877 | 34299 | 32285 | 34642 | 37941 |  |  |

**Answer to Questions 2(a)**

When A and B are independent, , which means that is only determined by . Therefore, A could be seen as frequent when is high though they are independent. This may lead to a wrong conclusion.

Because , we know and take into account.

**Answer to Questions 2(b)**

Therefore, only lift is symmetric.

For example, if we have four baskets*, A, B, AB* and *BC.*

Then ,

,

,

,

**Answer to Questions 2(c)**

If A and B occur simultaneously every time, then

.

, it is determined by .

.

Therefore, confidence and conviction are desirable.

For example, we have three baskets, AB, CD, CDE where A occurs every time B occurs and C occurs every time D occurs.

,

Though, these two rules are both 100% rules, they have different lift scores.

**Answer to Questions 2(d)**

DAI93865 -> FRO40251 1.0

GRO85051 -> FRO40251 0.999176276771005

GRO38636 -> FRO40251 0.9906542056074766

ELE12951 -> FRO40251 0.9905660377358491

DAI88079 -> FRO40251 0.9867256637168141

**Answer to Questions 2(e)**

(DAI23334,ELE92920)->DAI62779 1.0

(DAI31081,GRO85051)->FRO40251 1.0

(DAI55911,GRO85051)->FRO40251 1.0

(DAI62779,DAI88079)->FRO40251 1.0

(DAI75645,GRO85051)->FRO40251 1.0

**Answer to Questions 3(a)**

If a column has m 1’s, the number of columns with m 1’s of n is.

The number of these columns without 1 in one of k chosen rows is equal to the number that all m columns are not chosen

The probability of no 1 in the k chosen rows is p=

**Answer to Questions 3(b)**

Therefore,

**Answer to Questions 3(c)**

For example, the matrix with two columns is . The Jaccard Similarity is

When the second row and the third row are chosen as the random row r to be put at the first row, their minhash values are same. The probability is 0.5.