Market Basket Analysis

At Shopee, sellers list thousands of products for sale on our platform. A better understanding of users' tastes and preferences for products can help Shopee design better promotions and recommendations for our users. To do that, we conduct market basket analysis which allows us to identify the relationship between different combinations of products that users buy.

We are interested in finding ***association rules*** between combinations of different products. These *association rules* can help to uncover regularities in purchasing behaviors of our users.

For example, an ***association rule*** between 3 products, {Product A & Product B} → {Product C}, would indicate that a user buying both Product A & Product B would likely buy Product C as well.

Confidence is a measure that is used to indicate such tendencies and can be used to determine the association for varying numbers of products. For the purpose of this question, we will be using confidence to calculate the association for 2 products and 3 products.

***Confidence* for two products:**

$$Confidence\ (A > B) = \frac{(No.\ of\ orders\ containing\ both\ product\ A\ \&\ B)}{(No.\ of\ orders\ containing\ product\ A)}$$

***Confidence* for three products:**

$$Confidence\ (A > B\&C) = \frac{(No.\ of\ orders\ containing\ both\ product\ A, B\ \&C)}{(No.\ of\ orders\ containing\ product\ A)}$$

Or

$$Confidence\ (A\&B > C) = \frac{(No.\ of\ orders\ containing\ both\ product\ A, B\ \&C)}{(No.\ of\ orders\ containing\ product\ A\&B)}$$

Basic Concepts

Confidence is defined as the tendency that given product A is purchased, that product B will also be purchased.

Each orderid represents a distinct transaction that has occurred.

Each itemid represents a unique product that is sold on Shopee.

A transaction can contain 1 or more itemid(s). If 2 or more itemid(s) share the same orderid, they are purchased together in a single transaction.

An itemid can appear many times in different orderid(s), which means that the product was purchased many times in different transactions.

Task

Please calculate the ***confidence*** values for all the ***association rules*** provided in the ***rules.csv*** file.

Tips:

1. A > B and B > A have different ***confidence*** and should be calculated separately
2. A & B > C and B & A > C are identical ***association rules*** and will yield the same ***confidence***

Examples

Case 1: A > B
8 orderid have itemid 7917849
(31338643584868, 31364354557783, 31368958440199, 31369772179043, 31371954695064, 31375314731607, 31377601474289, 31379328498817)

6 orderid out of above have both itemid 7917849 and itemid 18642183
(31338643584868, 31368958440199, 31369772179043, 31371954695064, 31375314731607, 31377601474289)

***Confidence* (7917849 > 18642183)**

= 6 / 8

= 0.750

orderid	itemid
31338643584868	7917849
31338643584868	18642183
31364354557783	7917849
31368958440199	7917849
31368958440199	18642183
31369772179043	7917849
31369772179043	18642183
31371954695064	7917849
31371954695064	18642183
31375314731607	7917849
31375314731607	18642183
31377601474289	7917849
31377601474289	18642183
31379328498817	7917849

Case 2: A&B > C

7 orderid have itemid 2363580843 and itemid 2002243261
(31342449702678, 31365563352719, 31366764361012, 31371701813987, 31372163437582, 31373610230585, 31381568386099)

6 orderid out of above have all itemid 2363580843, itemid 2002243261 and itemid 1993068031
(31342449702678, 31365563352719, 31366764361012, 31372163437582, 31373610230585, 31381568386099)

***Confidence* (2363580843 & 2002243261 > 1993068031)**

= 6 / 7

= 0.857 (rounded to 3 decimal places)

orderid	itemid
31342449702678	2363580843
31342449702678	1993068031
31342449702678	2002243261
31365563352719	2363580843
31365563352719	2002243261
31365563352719	1993068031
31366764361012	2363580843
31366764361012	1993068031
31366764361012	2002243261
31371701813987	2363580843
31371701813987	2002243261
31372163437582	2363580843
31372163437582	2002243261
31372163437582	1993068031
31373610230585	2002243261
31373610230585	2363580843
31373610230585	1993068031
31381568386099	2002243261
31381568386099	1993068031
31381568386099	2363580843

Case 3: A > B&C
9 orderid have itemid 1089203645
(31351735245918, 31367488312991, 31372554805324, 31373458010259, 31373724807962, 31374927925523, 31375318612401, 31375354382289, 31384570619582)

7 orderid out of above have all itemid 1089203645, 431391770 and 1216842899
(31351735245918, 31372554805324, 31373458010259, 31373724807962, 31374927925523, 31375318612401, 31375354382289)

***Confidence* (1089203645 > 431391770 & 1216842899)**

= 7 / 9

= 0.778 (rounded to 3 decimal places)

orderid	itemid
31351735245918	431391770
31351735245918	1089203645
31351735245918	1216842899
31367488312991	1089203645
31372554805324	1216842899
31372554805324	431391770
31372554805324	1089203645
31373458010259	1216842899
31373458010259	1089203645
31373458010259	431391770
31373724807962	1089203645
31373724807962	431391770
31373724807962	1216842899
31374927925523	1089203645
31374927925523	431391770
31374927925523	1216842899
31375318612401	1216842899
31375318612401	431391770
31375318612401	1089203645
31375354382289	1216842899
31375354382289	431391770
31375354382289	1089203645
31384570619582	1089203645

Data Description

Edit

association_order.csv: It contains transaction order information. Columns: [orderid, itemid]
- Each orderid represents a distinct transaction that has occurred.

- Each itemid represents a unique product that is sold on Shopee.

rules.csv: It contains a list of association rules for which you are required to find the confidence values for.

- These association rules are selected because the combinations of products have appeared together in quite a number of orders and we would like to find out their confidence values

Evaluation

Evaluation Metric

Submissions are scored based on Categorisation Accuracy:

Evaluation Description

Your submission will be evaluated based on categorisation correctness. The score is the number of correct results that your submitted file contains.

$$score = \frac{1}{N}\sum_{i=0}^{N} p(x_i, y_i)$$

Where

- N is number of test samples.

- x_i is the predicted category for i th test sample.

- y_i is the ground truth for i th test sample.

- $p(x_i, y_i)$ is calculated as 1 if $x_i = y_i$ and 0 otherwise.

Submission Format

Submission Format

Please calculate the confidence values for all the association rules provided in the rules.csv file.

Two columns required:

1. rule (provided in rules.csv file)

2. confidence: Please times the confidence by 1000 and round down to integer. e.g.

(1) If a case has confidence = 5/16 = 0.3125, then 0.3125*1000 = 312.5 and round down to integer 312, please submit 312

(2) When you find confidence = 1, please submit 1000

(3) If a case has confidence = 1/16 = 0.0625, then 0.0625*1000 = 62.5 and round down to integer 62, please submit 62. (No need to add 0 prior to 6 to make it 062)

rule	confidence
7917849>18642183	750
2002243261&2363580843>1993068031	857
1089203645>431391770&1216842899	777

Your submission should have **14,238 rows** (excluding the headers) , each with **2 columns.** Participants may make up to 20 submissions as a team for this challenge.