## **Q1** Information and Honor Code

0 Points

In this assignment, you complete the Colab2 worksheet and obtain results from it. Your answer would be an integer, a float number, or a string. The float value should be a decimal number rounded to the **nearest 0.001**. For example, 0.2435 would become `0.244`. The string should be all **lowercase**, no quote or bracket, but may contain spaces or comma. For example, 'Organic Garlic, Limes' would become `organic garlic, limes`.

You can submit as many times as you want, and the last submission will be graded. Only the fully corrected answer will receive 1 point. No late day is allowed for any Colab assignment.

Please verify that you have read the above instructions and the Stanford Honor Code and that you have not given or received unpermitted aid while completing this assignment.

If you have any questions about how the Honor Code applies to Colab assignments or other parts of the course, please contact the teaching staff for clarification.

◉ I have read and understood the above information

## **Q2** Frequent pattern in spark

8 Points

Visualize the frequent itemsets in the dataset with a threshold of 0.01 for support, and a threshold of 0.5 for confidence.

First, you want to know the number of frequent itemsets and the number of association rules.

Q2.1, how many frequent itemsets do you obtain? (Integer)

Q2.2, how many association rules do you obtain? (Integer)

Now decrease the threshold of support to be 0.001, while the threshold of confidence stays unchanged as 0.5. With the new threshold values for all the following questions:

Q2.3, how many itemsets are frequent? (Integer)

Q2.4, how many association rules do you obtain? (Integer)

Furthermore, visualize the top 20 frequent itemsets.

Q2.5, what is the most frequent itemsets? (String)

Q2.6, what is the frequency for 'Strawberries' in the top 20 list? (Integer)

Q2.7-2.8, what is the `confidence` and `lift` for the association rule with `antecedent` as [Organic Broccoli, Organic Hass Avocado] and `consequent` as [Bag of Organic Bananas]? The answers should be float numbers, and you may find the method `df.show(False)` to be useful for avoiding truncation.

## **Q2.1** Number of frequent itemsets

1 Point

## **Q2.2** Number of association rules

1 Point

## **Q2.3** Number of frequent itemsets with the new threshold

1 Point

**Q2.4** Number of association rules with the new threshold
1 Point

**Q2.5** The most frequent item is
1 Point

**Q2.6** Strawberries' frequency
1 Point

**Q2.7** Confidence for the association [Organic Broccoli, Organic Hass Avocado] -> [Bag of Organic Bananas]
1 Point

**Q2.8** Lift for the association [Organic Broccoli, Organic Hass Avocado] -> [Bag of Organic Bananas]
1 Point