ITE4005 Data Science course

Assignment 1

2016026080 한다진

* **Experiment environments**

OS Windows 10

Language python3

Tools Jupyter notebook

Pure python script

* **Apriori algorithm**

“*sets*” is the name of list has the data about item\_set\_object. Item\_set\_object format is like this {‘*items’*:{1,2,3}, ‘*nums’*:5, ‘*idxs’*:[4,5,6,7,8]}. For this case, it means, the item\_set {1,2,3} is appeared 5 times and is appeard in 4th, 5th, 6th, 7th, 8th transaction.

Every step, item\_set\_objects are newly added to *sets.* And start\_point is the index of first item\_set\_objects among newly added item\_set\_objects to *sets*.

Get initial *sets* with function init\_sets().

Set initial start\_point and initial end\_point to 0 and len(sets)

while start point is less than end\_point(=len(sets)).

1. Get the item\_set\_object\_A(obj\_A) = sets[i] (start\_point <= i <end\_point)
2. Get the item\_set\_object\_B(obj\_B) = sets[j] (0 <= j < i)
3. 1. Check if obj\_A[‘items’] and obj\_B[‘items’] are mutually exclusive. if yes, go 3-2  
   2. Check times the item\_set(=obj\_A[‘items’]|obj\_B[‘items’]) appeard in transactions. If times are bigger than minimum\_support, go 3-3  
   3. Add the new item\_set\_object with item\_set (obj\_A[‘items’]|obj\_B[‘items’]) and make new rules.
4. Execute with every (i,j) cases
5. Update start\_point to end\_point, and end\_point to len(sets). and run one more while loop

* **Code explanation**
  + Import Libraries



Import sys libraries to get the arguments

* + **Apriori Class**



Make apriori class and use the internal function in apriori class. Detail of each function is below.

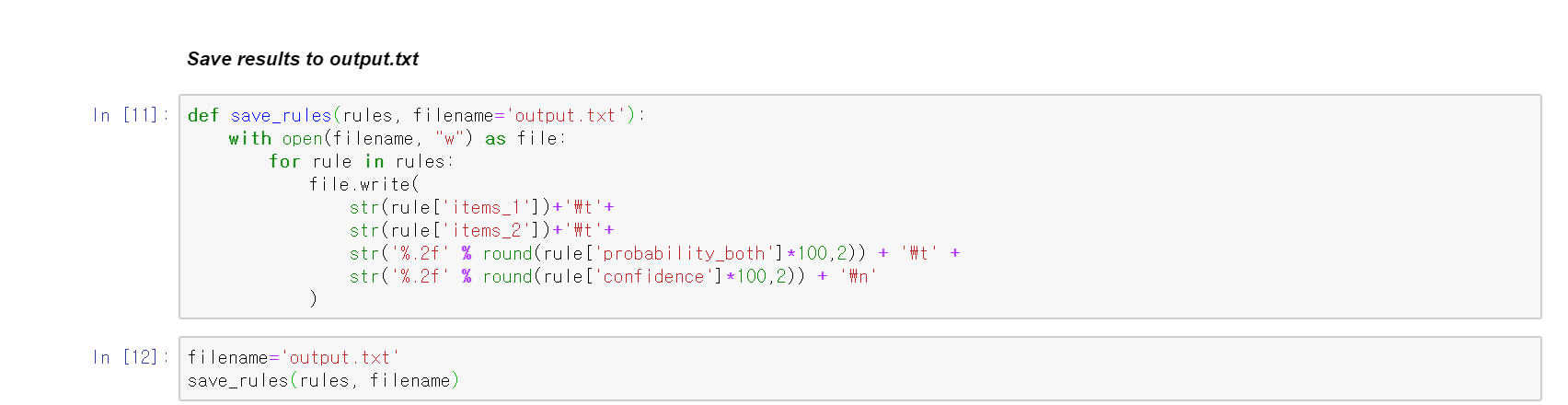
* + Load input file



Load input file and store it to *transactions*.

Every items in each transactions are managed with int type.

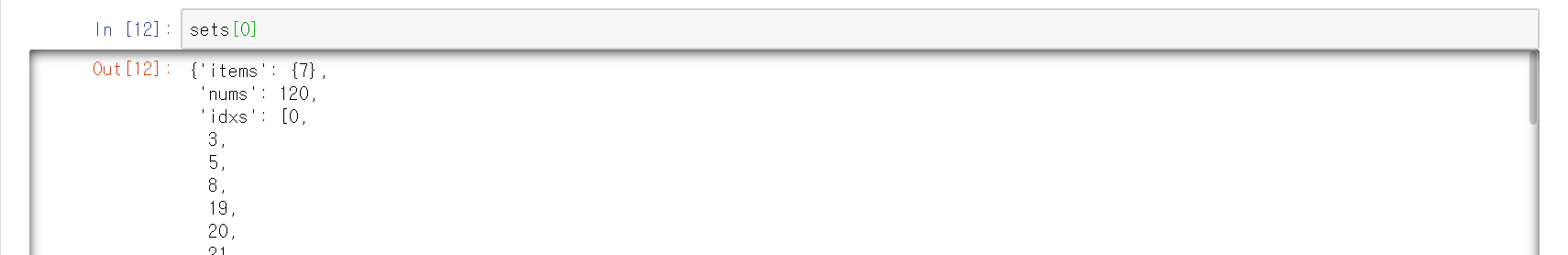
* + Save rules to output.txt



Save *rules* to output.txt.

* + Make initial sets with transaction file





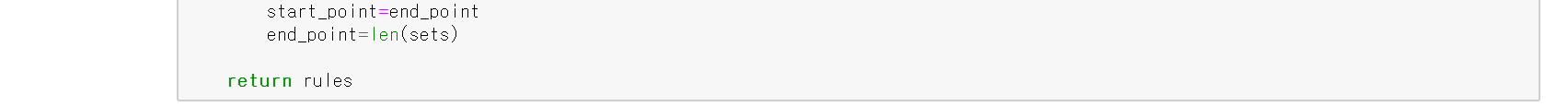
First, find every item in *transactions* and store it to *items*. *Items* is a dictionary. Its key is an item and its value is consist of ‘*nums’* and ‘*idxs’*. ‘*nums’* means the number of times an item has appeared. ‘*idxs’* is the list of indexs which current item is appeard in *transactions*.

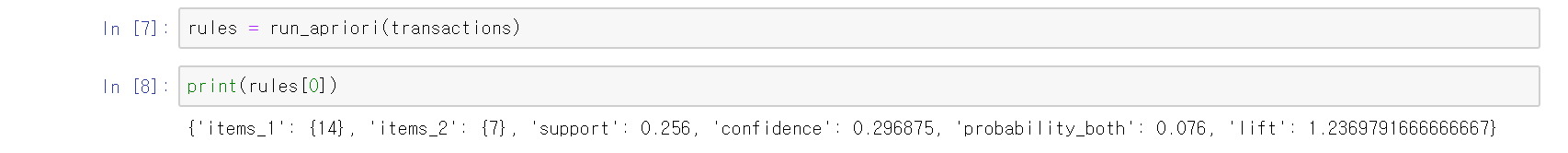
Then make list of dictionary with data in *items*. Each dictionary is consist of ‘items’, ‘nums’ and ‘idxs’. ‘nums’ and ‘idxs’ are same with those of *items*. ‘items’ means the set of item.

* + Run apriori algorithm









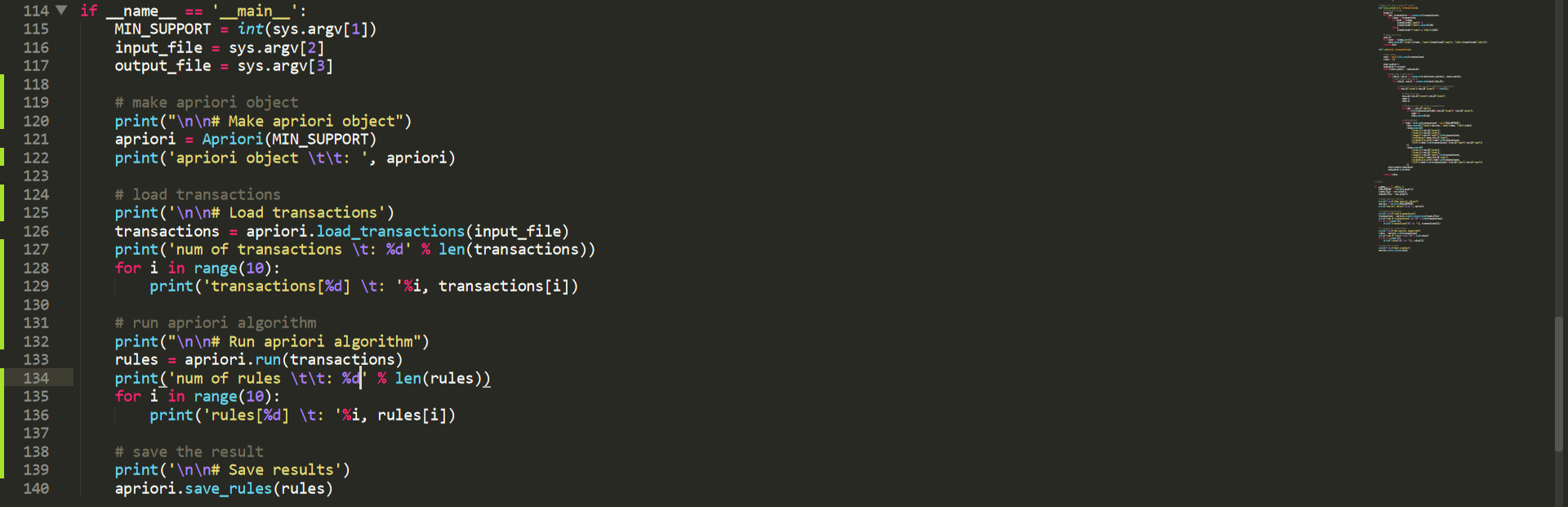
Details are explained in ***Apriori Algorithm***

‘**support’** means the probability of an transaction ‘items\_1’ occurring.

‘**confidence’** means the probability of an transaction ‘items\_2’ occurring where transaction ‘items\_1’ occuered

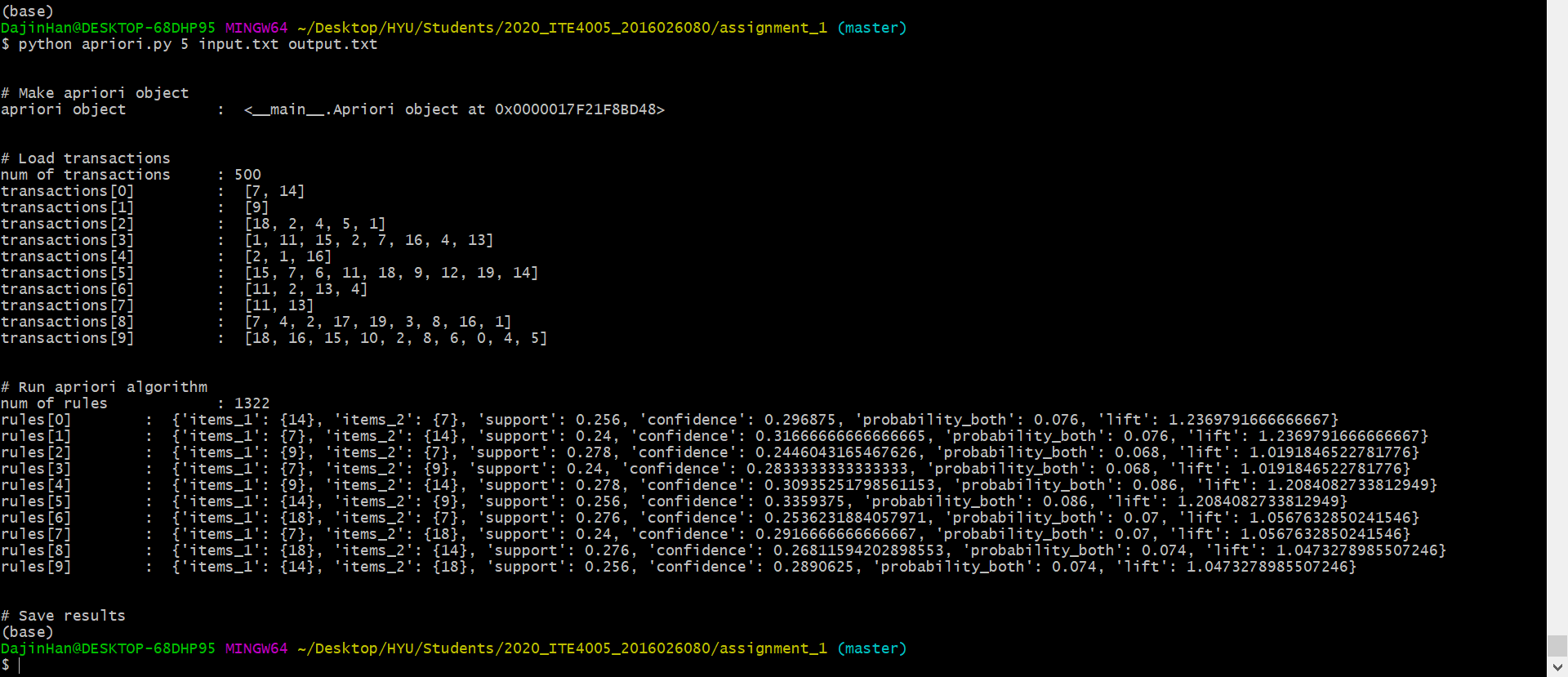
**‘probability\_both’** means the probability of an transaction ’items\_1’&items\_2’ occuring

* **Experiment**



This is the python code in apriori.py to check the result in console screen.

Show up to 10th transaction and 10rules.



This is the result screen when excuting the command “python apriori.py 5 input.txt output.txt”