Data Mining Project 1

學號: P76111262 姓名: 林晨鈞

- Prpgramming

1. Apriori:

請執行 main.py，Apriori 已完成

2. FP-Tree:

請執行 fp_main.py，因為本身設計的演算法有問題，所以執行較大資料集會執行過久，我依照老師上課講義產生出一模

TID	Items
1	Milk, Bread, Beer
2	Bread, Coffee
3	Bread, Egg
4	Milk, Bread, Coffee
5	Milk, Egg
6	Bread, Egg
7	Milk, Egg
8	Milk, Bread, Egg, Beer
9	Milk, Bread, Egg

一樣的測資(如右)

minsup = 2

我存成 fp_data.txt，欲執行可以複製以下指令:

python fp_main.py --min_sup 0.2 --min_sup 0.2 --dataset fp_data.txt

還請助教斟酌給分!

3. Kaggle(Bonus):

請執行 kaggle_main.py

因為此 kaggle 資料集還尚須去掉含 None 的 label 欄位...等資

料前處理，以符合我的程式的 input，所以新建一個

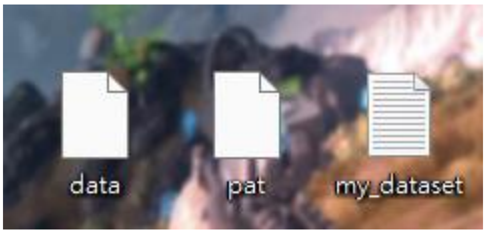
kaggle_main.py，欲執行請用此.py 檔，輸出檔案為

BreadBasket_DMS-kaggle.csv

- # 分析 Report

用 IBM Quest Data Generator 產生自己的 dataset，並取名為”my_dataset.txt”，以下 report 用此資料集進行實驗，此資料集 Transation ID 數為 4114

```
C:\Users\brian\Desktop>"IBM Quest Data Generator.exe" lit
```



1. What do you observe in the below 4 scenarios?

- ✓ 高的 support 值代表需要出現較多次的 Itemset 才能成為 Frequent Itemset，所以 Frequent Itemset 的數量會較少。反之亦然。
- ✓ 高的 confidence 值代表關聯性很高，形成 association rule 的條件變得嚴苛。反之亦然。

■ High support, High confidence:

因為 high support、high confidence 的原因，所以 frequent itemsets 較少且關聯性較高者才能形成 association rules，此種組合下，雖然可輕易看出關聯性高的項目，但同時也會失去很多具有不俗的關聯性的項目，因此參考價值較為不好。

- ➢ Min_sup:0.025 / Min_conf:0.9 Result: 11 個 association rules

```
C:\Users\brian\Desktop\python\hw1-example>python main.py --min_sup 0.025 --min_conf 0.9 --dataset my_dataset.txt
```

	A	B	C	D	E
1	antecedent	consequent	support	confidence	lift
2	{55776}	{40313}	0.028	0.9	23.735
3	{92490}	{51657}	0.025	0.972	37.36
4	{92490}	{12305}	0.025	0.981	26.21
5	{92490}	{95398}	0.025	0.972	36.675
6	{51657}	{12305}	0.025	0.972	25.965
7	{51657}	{92490}	0.025	0.963	37.36
8	{51657}	{95398}	0.026	0.981	37.038
9	{95398}	{51657}	0.026	0.963	37.038
10	{95398}	{92490}	0.025	0.945	36.675
11	{95398}	{12305}	0.026	0.972	25.979
12	{39749}	{64408}	0.026	0.921	17.224

■ High support, Low confidence:

因為 high support、low confidence 的原因，所以 frequent itemsets 較少，但關聯性不高者也能形成 association rules，此種組合下，大多數的 itemset 會在選擇 frequent itemset 時就被排除掉，此組合有一定的參考價值。

➢ Min_sup:0.025 / Min_conf:0.1 Result: 16 個 association rules

```
C:\Users\brian\Desktop\python\hw1-example>python main.py --min_sup 0.025 --min_conf 0.1 --dataset my_dataset.txt
```

	A	B	C	D	E
1	antecedent	consequent	support	confidence	lift
2	{95398}	{92490}	0.025	0.945	36.675
3	{95398}	{12305}	0.026	0.972	25.979
4	{95398}	{51657}	0.026	0.963	37.038
5	{40313}	{55776}	0.028	0.75	23.735
6	{64408}	{39749}	0.026	0.477	17.224
7	{12305}	{51657}	0.025	0.675	25.965
8	{12305}	{92490}	0.025	0.675	26.21
9	{12305}	{95398}	0.026	0.688	25.979
10	{51657}	{12305}	0.025	0.972	25.965
11	{51657}	{92490}	0.025	0.963	37.36
12	{51657}	{95398}	0.026	0.981	37.038
13	{92490}	{95398}	0.025	0.972	36.675
14	{92490}	{12305}	0.025	0.981	26.21
15	{92490}	{51657}	0.025	0.972	37.36
16	{55776}	{40313}	0.028	0.9	23.735
17	{39749}	{64408}	0.026	0.921	17.224

■ Low support, Low confidence:

因為 low support、low confidence 的原因，所以 frequent itemsets 很多且關聯性低者也形成 association rules，此種組合下，多數的 itemset 之間皆高機率形成 association rules，非常不好看出區別，因此參考價值十分差。

➢ Min_sup:0.01 / Min_conf:0.5 Result: 17557 個 association rules

```
C:\Users\brian\Desktop\python\hw1-example>python main.py --min_sup 0.01 --min_conf 0.5 --dataset my_dataset.txt
```

	A	B	C	D	E
17545	{10413 157	{31717}	0.012	0.98	60.199
17546	{10413 157	{31717 625	0.011	0.922	61.151
17547	{10413 317	{86958}	0.011	1	60.5
17548	{10413 317	{15755}	0.011	1	49.566
17549	{10413 317	{15755 869	0.011	1	61.403
17550	{10413 157	{83249}	0.011	0.94	60.424
17551	{10413 317	{15755}	0.011	1	49.566
17552	{10413 157	{31717}	0.011	0.979	60.124
17553	{10413 157	{62583}	0.011	0.94	61.383
17554	{10413 157	{78656}	0.011	0.904	59.023
17555	{15755 317	{10413}	0.011	0.94	60.424
17556	{10413 157	{60042}	0.011	0.979	61.035
17557	{10413 157	{86958}	0.011	1	60.5

■ Low support, High confidence:

因為 low support、high confidence 的原因，所以 frequent itemsets 較多但要形成 association rules 的條件變得嚴苛，此種組合下，形成的 association rules 較完整且數量也夠多，因此有一定的參考價值。

- ➢ Min_sup:0.01 / Min_conf:0.99　Result: 440 個 association rules

```
C:\Users\brian\Desktop\python\hw1-example>python main.py --min_sup 0.01 --min_conf 0.96 --dataset my_dataset.txt
```

	A	B	C	D	E
428	{10413 31'	{15755}	0.011	1	49.566
429	{10413 31'	{15755 86!	0.011	1	61.403
430	{6080 249.	{14148}	0.011	1	44.237
431	{10413 31'	{15755}	0.013	1	49.566
432	{10413 15'	{86958}	0.012	1	60.5
433	{10413 31'	{15755}	0.012	1	49.566
434	{10413 15'	{86958}	0.012	1	60.5
435	{10413 31'	{15755}	0.012	1	49.566
436	{10413 60(	{15755}	0.012	1	49.566
437	{6080 141.	{37321}	0.011	1	70.931
438	{31717 60(	{15755}	0.012	1	49.566
439	{10413 31'	{15755}	0.012	1	49.566
440	{10413 31'	{15755}	0.011	1	49.566
441	{10413 15'	{86958}	0.011	1	60.5

● Bonus:

kaggle 資料集簡介:

The dataset consists of 21293 observations from a bakery.

The data file contains four variables, Date, Time, Transaction ID and Item.

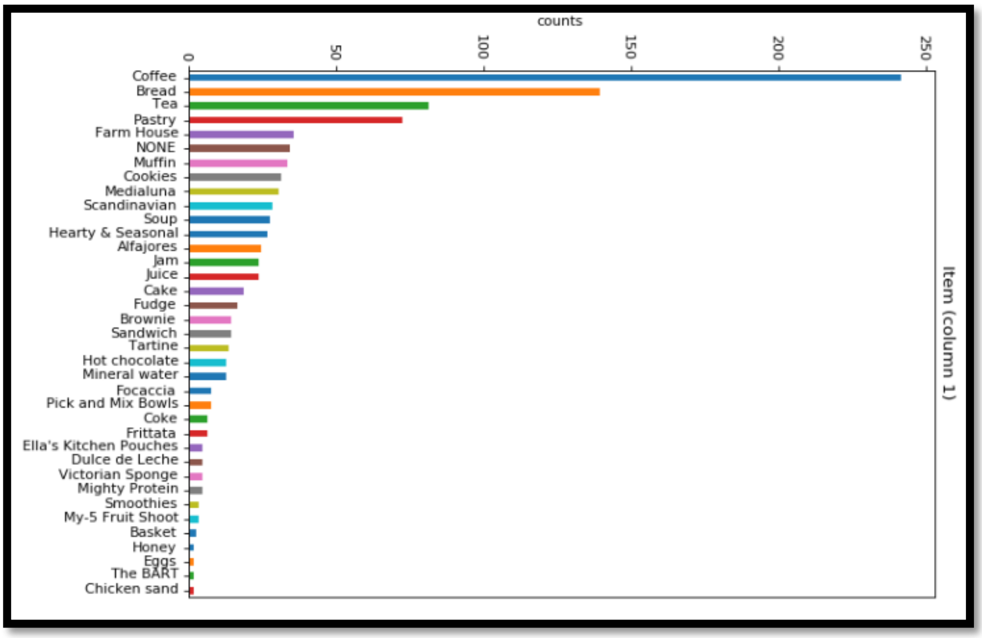
Transaction ID ranges from 1 through 9684.

However, there are some skipped numbers in Transaction IDs.

Source Website : https://www.kaggle.com/sulmansarwar/transactions-from-a-bakery?select=BreadBasket_DMS.csv

(以下為 BreadBasket_DMS.csv 格式，此資料共 9684 筆交易)

1	Date	Time	Transactior	Item
2	2016/10/30	09:58:11	1	Bread
3	2016/10/30	10:05:34	2	Scandinavian
4	2016/10/30	10:05:34	2	Scandinavian
5	2016/10/30	10:07:57	3	Hot chocolate
6	2016/10/30	10:07:57	3	Jam
7	2016/10/30	10:07:57	3	Cookies
8	2016/10/30	10:08:41	4	Muffin
9	2016/10/30	10:13:03	5	Coffee
10	2016/10/30	10:13:03	5	Pastry
11	2016/10/30	10:13:03	5	Bread



因為此 kaggle 資料集還尚須去掉含 None 的 label 欄位...等資料前處理，以符合我的程式的 input，所以新建一個 kaggle_main.py，欲執行請用此.py 檔，輸出檔案為 BreadBasket_DMS-kaggle.csv

High support, High confidence:

- Min_sup:0.009 / Min_conf:0.6　Result: 1 個 association rules
- 此 min_sup 和 min_conf 的組合會產生一條 association rules，可以發現 Toast 和 Coffee 兩個商品間有非常高的關聯性

```
C:\Users\brian\Desktop\python\hw1-example>python kaggle_main.py --min_sup 0.009 --min_conf 0.6
```

	antecedent	consequent	support	confidence	lift
2	{Toast}	{Coffee}	0.023	0.704	1.247
3					

■ High support, Low confidence:

- ➢ Min_sup:0.009 / Min_conf:0.4　Result: 16 個 association rules

- ➢ 此 min_sup 和 min_conf 的組合，可以發現 Coffee 這個商品，最常出現在 consequent，大多數人在買完其他商品都會買 Coffee

```
C:\Users\brian\Desktop\python\hw1-example>python kaggle_main.py --min_sup 0.009 --min_conf 0.4
```

1	antecedent	consequent	support	confidence	lift
2	{Toast}	{Coffee}	0.023	0.704	1.247
3	{Scone}	{Coffee}	0.018	0.523	0.926
4	{Sandwich}	{Coffee}	0.037	0.47	0.831
5	{Muffin}	{Coffee}	0.018	0.481	0.852
6	{Cake}	{Coffee}	0.053	0.505	0.895
7	{Spanish Brunch}	{Coffee}	0.011	0.599	1.06
8	{Medialuna}	{Coffee}	0.034	0.541	0.957
9	{Juice}	{Coffee}	0.02	0.528	0.935
10	{Alfajores}	{Coffee}	0.019	0.504	0.892
11	{Soup}	{Coffee}	0.015	0.439	0.776
12	{Brownie}	{Coffee}	0.019	0.491	0.869
13	{Pastry}	{Coffee}	0.046	0.526	0.931
14	{Hot chocolate}	{Coffee}	0.029	0.475	0.84
15	{Cookies}	{Coffee}	0.028	0.494	0.875
16	{Bread Cake}	{Coffee}	0.01	0.43	0.761
17	{Tea Cake}	{Coffee}	0.01	0.422	0.747

■ Low support, Low confidence:

- ➢ Min_sup:0.003 / Min_conf:0.4　Result: 44 個 association rules

- ➢ 此 min_sup 和 min_conf 的組合，可以發現 Coffee 這個商品，最常出現在 consequent，大多數人在買完其他商品都會買 Coffee，但是 antecedent 相對較雜亂，比較不好看出更細的細節。

```
C:\Users\brian\Desktop\python\hw1-example>python kaggle_main.py --min_sup 0.003 --min_conf 0.4
```

	A	B	C	D	E
3	{Alfajores}	{Coffee}	0.019	0.504	0.892
4	{Toast}	{Coffee}	0.023	0.704	1.247
5	{Cookies}	{Coffee}	0.028	0.494	0.875
6	{Scone}	{Coffee}	0.018	0.523	0.926
7	{Cake}	{Coffee}	0.053	0.505	0.895
8	{Smoothies	{Coffee}	0.004	0.494	0.874
9	{Sandwich	{Coffee}	0.037	0.47	0.831
10	{Extra Sala	{Coffee}	0.003	0.816	1.444
11	{Spanish B	{Coffee}	0.011	0.599	1.06
12	{The Nome	{Coffee}	0.003	0.534	0.946
13	{Soup}	{Coffee}	0.015	0.439	0.776
14	{Salad}	{Coffee}	0.006	0.626	1.109
15	{Tiffin}	{Coffee}	0.008	0.548	0.97
16	{Frittata}	{Coffee}	0.004	0.531	0.94
17	{Muffin}	{Coffee}	0.018	0.481	0.852
18	{Juice}	{Coffee}	0.02	0.528	0.935
19	{Hearty &	{Coffee}	0.006	0.505	0.893
20	{Keeping I	{Coffee}	0.005	0.81	1.433
21	{Medialuna	{Coffee}	0.034	0.541	0.957
22	{Hot choco	{Coffee}	0.029	0.475	0.84
23	{Vegan mi	{Coffee}	0.003	0.556	0.983
24	{Jammie D	{Coffee}	0.007	0.504	0.892
25	{Brownie}	{Coffee}	0.019	0.491	0.869
26	{Cookies J	{Coffee}	0.004	0.603	1.068
27	{Bread Cal	{Coffee}	0.01	0.43	0.761
28	{Bread Sar	{Coffee}	0.007	0.422	0.748
29	{Soup Sanc	{Coffee}	0.004	0.654	1.157
30	{Hot choco	{Coffee}	0.004	0.667	1.18
31	{Pastry Tea	{Coffee}	0.005	0.484	0.856
32	{Bread Alf	{Coffee}	0.004	0.418	0.741
33	{Medialuna	{Coffee}	0.004	0.455	0.805
34	{Juice Cak	{Coffee}	0.004	0.552	0.977
35	{Cake San	{Coffee}	0.005	0.677	1.198
36	{Tea Scone	{Coffee}	0.003	0.405	0.717
37	{Cookies C	{Coffee}	0.004	0.58	1.026
38	{Hot choco	{Coffee}	0.004	0.614	1.087
39	{Bread Hot	{Coffee}	0.006	0.457	0.808
40	{Pastry Me	{Coffee}	0.005	0.529	0.936
41	{Bread Toa	{Coffee}	0.004	0.473	0.837
42	{Hot choco	{Coffee}	0.007	0.602	1.065
43	{Tea Toast	{Coffee}	0.003	0.508	0.9
44	{Tea Cake	{Coffee}	0.01	0.422	0.747
45	{Bread Me	{Coffee}	0.007	0.4	0.708

- # Low support, High confidence:
  - Min_sup:0.003 / Min_conf:0.6　Result:10 個 association rules
  - 此 min_sup 和 min_conf 的組合，可以發現 Coffee 這個商品，最常出現在 consequent，大多數人在買完其他商品都會買 Coffee，antecedent 相對較 Low sup/Low conf 組合，篩選出關聯度較高的 association rules，以利於分析。

```
C:\Users\brian\Desktop\python\hw1-example>python kaggle_main.py --min_sup 0.003 --min_conf 0.6
```

	A	B	C	D	E
1	antecedent	consequent	support	confidence	lift
2	{Extra Salami or Feta}	{Coffee}	0.003	0.816	1.444
3	{Toast}	{Coffee}	0.023	0.704	1.247
4	{Salad}	{Coffee}	0.006	0.626	1.109
5	{Keeping It Local}	{Coffee}	0.005	0.81	1.433
6	{Cookies Juice}	{Coffee}	0.004	0.603	1.068
7	{Cake Sandwich}	{Coffee}	0.005	0.677	1.198
8	{Hot chocolate Cookies}	{Coffee}	0.004	0.614	1.087
9	{Hot chocolate Pastry}	{Coffee}	0.004	0.667	1.18
10	{Hot chocolate Cake}	{Coffee}	0.007	0.602	1.065
11	{Soup Sandwich}	{Coffee}	0.004	0.654	1.157

■ 結論:

從上面實驗的例子，我最大的發現就是，在這間麵包店裡面，最夯的產品就是咖啡，當一個客人到這間麵包店裡面買任何商品時，高機率都會再配上咖啡。