Data Mining Project 1

Q56061040 湯立婷

IBM Data

1. Number of transactions in database
2. Average transaction length
3. Number of items

拿4種不同組合參數生成資料

1. ntrans=0.1, tlen=20, nitems=0.1
2. ntrans=0.1, tlen=20, nitems=0.5
3. ntrans=0.5, tlen=20, nitems=0.1
4. ntrans=0.5, tlen=40, nitems=0.1

IBM Data (Cont.)

左圖每行分別代表CustID, TransID, Item

右圖為DAT格式，將同一個Transaction的Item放在同一列

```
1 1 0
1 1 6
1 1 8
1 1 34
1 1 36
1 1 38
1 1 42
1 1 45
1 1 47
1 1 49
1 1 50
1 1 51
1 1 52
1 1 53
1 1 55
1 1 61
1 1 62
1 1 63
1 1 67
1 1 69
1 1 71
1 1 74
1 1 78
1 1 83
1 1 94
2 2 4
```

```
0,6,8,34,36,38,42,45,47,49,50,51,52,53,55,61,62,63,67,69,71,74,78,83,94,
4,8,9,11,14,17,18,35,36,38,39,40,41,42,43,59,63,69,73,80,81,85,87,93,97,
8,9,10,14,17,21,25,36,38,40,43,45,57,60,62,63,69,83,85,93,
0,4,7,11,12,13,14,20,23,26,29,38,42,46,61,62,63,72,73,85,86,91,96,
3,6,11,12,14,17,18,19,21,25,33,38,48,61,62,63,68,71,75,80,81,83,84,87,89,
0,11,13,38,40,60,63,74,80,
3,5,7,17,23,28,32,38,39,40,43,47,52,57,61,67,69,70,79,85,87,93,
3,29,43,45,69,72,74,78,89,93,95,97,
0,3,4,12,15,28,33,38,39,40,47,48,61,65,71,72,74,80,86,87,89,96,98,
17,21,33,36,43,45,47,48,61,69,73,78,81,83,87,
3,5,6,28,35,36,38,40,43,48,51,52,61,62,63,74,77,85,86,
3,5,8,11,14,17,20,21,28,33,36,40,51,52,57,59,63,66,81,83,85,89,90,
5,8,14,21,29,31,35,38,39,41,42,59,62,63,70,86,95,
1,3,5,7,11,14,28,31,33,35,36,43,48,61,63,69,80,82,86,91,
12,21,26,28,29,31,43,45,57,61,62,63,67,69,77,80,81,83,86,87,88,89,93,96,
3,8,11,12,17,20,23,28,29,35,38,40,43,47,48,50,62,63,66,71,72,73,81,83,92,97,
13,23,25,31,35,36,38,50,62,66,67,70,71,74,81,85,
21,36,43,51,59,62,63,69,70,87,89,97,
5,13,23,29,36,42,46,48,51,57,63,68,69,72,74,81,83,85,87,95,
3,8,9,17,21,27,28,36,38,40,43,45,48,49,57,66,67,69,71,74,78,81,84,85,86,87,89,91,93,95,
7,13,21,34,35,38,47,49,52,63,66,68,71,80,89,
3,8,10,13,25,26,28,36,38,39,48,51,62,63,69,72,73,81,83,85,93,
3,8,12,14,17,23,35,38,41,48,51,57,61,63,69,73,76,85,86,87,89,90,97,
```

Kaggle data

New Zealand Migration

Migration numbers to and from New Zealand from 1979 to 2016

每行feature分别是:

1. **Measure**: The signal type given in this row, one of: "Arrivals", "Departures", "Net"
2. **Country**: Country from where people arrived into to New Zealand (for Measure = "Arrivals") or to where they left (for Measure = "Departures"). Contains special values "Not Stated" and "All countries" (grand total)
3. **Citizenship**: Citizenship of the migrants, one of: "New Zealand Citizen", "Australian Citizen", "Total All Citizenships"
4. **Year**: Year of the measurement
5. **Value**: Number of migrants

Kaggle data (Cont.)

將公民身份別為紐西蘭，以及移民人數大於1000的資料抓出來，以同年份代表同一Transaction ID，希望算出紐西蘭至別的國家移民的關聯

123	1989	Hong-Kong
124	1989	Japan
125	1989	Malaysia
126	1989	Singapore
127	1989	Taiwan
128	1989	Europe
129	1989	UK
130	1989	Americas
131	1989	USA
132	1989	Not-stated
133	1990	Oceania
134	1990	Australia
135	1990	Fiji
136	1990	New-Zealand
137	1990	Samoa
138	1990	Asia

```
Oceania,Australia,Fiji,Asia,Hong-Kong,Japan,Malaysia,Taiwan,Europe,UK,USSR,
Oceania,Australia,Fiji,Asia,Hong-Kong,Japan,South-Korea,Malaysia,Taiwan,Eu
ica-and-the-Middle-East,South-Africa,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,M
K,Americas,USA,Africa-and-the-Middle-East,South-Africa,Not-stated,
Oceania,Australia,Fiji,Samoa,Asia,China,Hong-Kong,India,Japan,South-Korea,M
K,Americas,Canada,USA,Africa-and-the-Middle-East,South-Africa,Not-stated,
```

Implement Apriori Algorithm

min_support=0.01, min_conf=0.3

FP Growth Algorithm - IBM Data

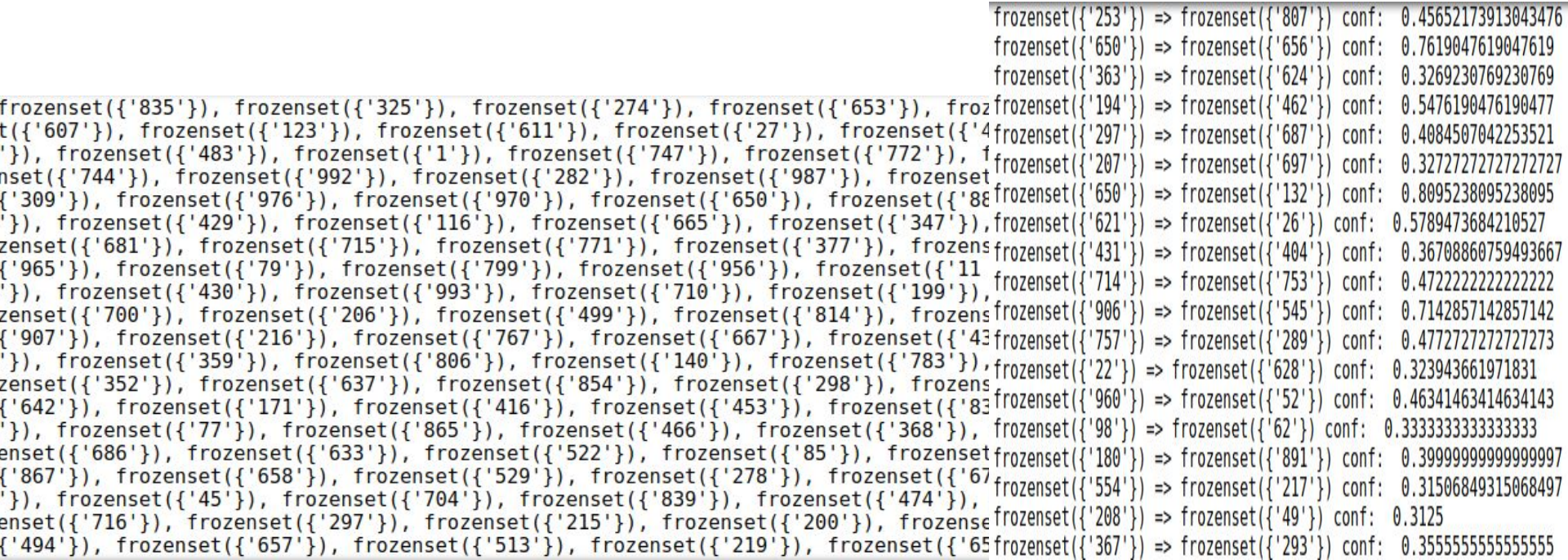
min_support=0.01

```
[['42', '397', '471', '510', '553', '629', '644', '656', '716', '772', '790', '838'],
['56', '293', '521', '729'], ['321', '459', '524', '543', '578', '592', '599', '622'],
['29', '39', '46', '152', '306', '336', '412', '429', '432', '543', '547', '553', '766',
'794', '819', '870', '894', '902', '981'], ['80', '108', '304', '388', '625', '962', '96
4', '994'], ['62', '107', '129', '163', '437', '629', '705', '815', '904'], ['137', '18
8', '214', '405', '418', '420', '443', '447', '566', '813', '837', '858'], ['51', '63',
'86', '146', '325', '374', '395', '444', '571', '628', '684', '970'], ['38', '60', '85',
'371', '387', '456', '764', '840'], ['8', '50', '60', '117', '360', '557', '578', '592',
'624', '753', '868', '870', '877', '954'], ['7', '35', '99', '102', '127', '172', '194',
'221', '266', '462', '607'], ['106', '351', '432', '599'], ['124', '740', '988', '989',
'990'], ['39', '155', '167', '169', '214', '287', '374', '497', '593', '733', '778', '80
6', '848', '850'], ['25', '106', '107', '123', '135', '167', '179', '182', '374', '414',
'571', '589', '601', '612', '673', '729', '820', '847', '911', '934'], ['3', '105', '11
6', '123', '773', '806'], ['36', '201', '404', '416', '559', '668', '682', '719', '803',
'916'], ['12', '47', '148', '238', '278', '368', '446', '471', '544', '584'], ['47', '32
8', '737', '772', '855', '970'], ['21', '255', '447', '456', '692', '803', '934', '981',
'994'], ['150', '371', '425', '477', '773', '966', '994'], ['7', '245', '412', '644', '7
32', '733', '744', '813', '874', '903', '946', '991'], ['15', '29', '119', '274', '283',
'472', '490', '571', '682', '776', '835', '884', '909'], ['182', '441', '442', '456', '4
73', '487', '506', '682', '756', '801', '820', '832', '845'], ['17', '144', '280', '33
```

```
946 frozenset: {'222'}
947 frozenset: {'238'}
948 frozenset: {'127'}
949 frozenset: {'444'}
950 frozenset: {'827'}
951 frozenset: {'132'}
952 frozenset: {'874'}
953 frozenset: {'38'}
954 frozenset: {'707'}
955 frozenset: {'63'}
956 frozenset: {'800'}
957 frozenset: {'221'}
958 frozenset: {'432'}
959 frozenset: {'416'}
960 frozenset: {'571'}
961 frozenset: {'592', '571'}
962 frozenset: {'553'}
963 frozenset: {'709'}
964 frozenset: {'592'}
```

Apriori Algorithm - IBM Data

min_support=0.001, min_conf=0.3



FP Growth Algorithm - IBM Data

min_support=0.001

```
[['42', '397', '471', '510', '553', '629', '644', '656', '716', '772', '790', '838'
['56', '293', '521', '729'], ['321', '459', '524', '543', '578', '592', '599', '622
['29', '39', '46', '152', '306', '336', '412', '429', '432', '543', '547', '553', '
'794', '819', '870', '894', '902', '981'], ['80', '108', '304', '388', '625', '962'
4', '994'], ['62', '107', '129', '163', '437', '629', '705', '815', '904'], ['137',
8', '214', '405', '418', '420', '443', '447', '566', '813', '837', '858'], ['51', '
'86', '146', '325', '374', '395', '444', '571', '628', '684', '970'], ['38', '60',
'371', '387', '456', '764', '840'], ['8', '50', '60', '117', '360', '557', '578', '
'624', '753', '868', '870', '877', '954'], ['7', '35', '99', '102', '127', '172', '
'221', '266', '462', '607'], ['106', '351', '432', '599'], ['124', '740', '988', '9
'990'], ['39', '155', '167', '169', '214', '287', '374', '497', '593', '733', '778'
6', '848', '850'], ['25', '106', '107', '123', '135', '167', '179', '182', '374', '
'571', '589', '601', '612', '673', '729', '820', '847', '911', '934'], ['3', '105',
6', '123', '773', '806'], ['36', '201', '404', '416', '559', '668', '682', '719', '
'916'], ['12', '47', '148', '238', '278', '368', '446', '471', '544', '584'], ['47'
8', '737', '772', '855', '970'], ['21', '255', '447', '456', '692', '803', '934', '
'994'], ['150', '371', '425', '477', '773', '966', '994'], ['7', '245', '412', '644
32', '733', '744', '813', '874', '903', '946', '991'], ['15', '29', '119', '274', '
'472', '490', '571', '682', '776', '835', '884', '909'], ['182', '441', '442', '456
73', '487', '506', '682', '756', '801', '820', '832', '845'], ['17', '144', '280',
```

```
946 frozenset: {'222'}
947 frozenset: {'238'}
948 frozenset: {'127'}
949 frozenset: {'444'}
950 frozenset: {'827'}
951 frozenset: {'132'}
952 frozenset: {'874'}
953 frozenset: {'38'}
954 frozenset: {'707'}
955 frozenset: {'63'}
956 frozenset: {'800'}
957 frozenset: {'221'}
958 frozenset: {'432'}
959 frozenset: {'416'}
960 frozenset: {'571'}
961 frozenset: {'592', '571'}
962 frozenset: {'553'}
963 frozenset: {'709'}
964 frozenset: {'592'}
```

Performance Comparation - **IBM Data**

	Apriori	**FP Growth**
minSup=0.01, min_conf=0.3	3117.908341	33.07
minSup=0.001, min_conf=0.3	2810.978982	38.62

Discussion - IBM Data

Apriori的時間複雜度為FP-Growth的100倍左右，比較特別的是將minSupport 值調低，Apriori的時間花費減少了，而FP-Growth時間增加

FP Growth Algorithm - Kaggle Data

```
user@user-System-Product-Name: /media/user/67a2dc9c-17dd-44b5-969d-d7ac6dcfcc
檔案(F)  編輯(E)  檢視(V)  搜尋(S)  終端機(T)  求助(H)
(Australia Europe Oceania USA) ==> (Not-stated ) confidence=
(Not-stated) ==> (Australia Europe UK ) confidence= 1.00
(Australia) ==> (Not-stated Europe UK ) confidence= 0.97
(Europe) ==> (Not-stated Australia UK ) confidence= 0.97
(UK) ==> (Not-stated Australia Europe ) confidence= 0.97
(Not-stated Australia) ==> (Europe UK ) confidence= 1.00
(Not-stated Europe) ==> (Australia UK ) confidence= 1.00
(Not-stated UK) ==> (Australia Europe ) confidence= 1.00
(Australia Europe) ==> (Not-stated UK ) confidence= 0.97
(Australia UK) ==> (Not-stated Europe ) confidence= 0.97
(Europe UK) ==> (Not-stated Australia ) confidence= 0.97
(Not-stated Australia Europe) ==> (UK ) confidence= 1.00
(Not-stated Australia UK) ==> (Europe ) confidence= 1.00
(Not-stated Europe UK) ==> (Australia ) confidence= 1.00
(Australia Europe UK) ==> (Not-stated ) confidence= 0.97
(Not-stated) ==> (Australia Europe UK USA ) confidence= 1.00
(Australia) ==> (Not-stated Europe UK USA ) confidence= 0.97
(Europe) ==> (Not-stated Australia UK USA ) confidence= 0.97
(UK) ==> (Not-stated Australia Europe USA ) confidence= 0.97
(USA) ==> (Not-stated Australia Europe UK ) confidence= 0.97
```

minsupport =1000

FP Growth Algorithm - Kaggle Data

```
user@user-System-Product-Name: /media/user/67a2dc9c-17dd-44b5-969d-d7ac6dcfcc
檔案(F)  編輯(E)  檢視(V)  搜尋(S)  終端機(T)  求助(H)
(Australia Europe Oceania USA) ==> (Not-stated ) confidence=
(Not-stated) ==> (Australia Europe UK ) confidence= 1.00
(Australia) ==> (Not-stated Europe UK ) confidence= 0.97
(Europe) ==> (Not-stated Australia UK ) confidence= 0.97
(UK) ==> (Not-stated Australia Europe ) confidence= 0.97
(Not-stated Australia) ==> (Europe UK ) confidence= 1.00
(Not-stated Europe) ==> (Australia UK ) confidence= 1.00
(Not-stated UK) ==> (Australia Europe ) confidence= 1.00
(Australia Europe) ==> (Not-stated UK ) confidence= 0.97
(Australia UK) ==> (Not-stated Europe ) confidence= 0.97
(Europe UK) ==> (Not-stated Australia ) confidence= 0.97
(Not-stated Australia Europe) ==> (UK ) confidence= 1.00
(Not-stated Australia UK) ==> (Europe ) confidence= 1.00
(Not-stated Europe UK) ==> (Australia ) confidence= 1.00
(Australia Europe UK) ==> (Not-stated ) confidence= 0.97
(Not-stated) ==> (Australia Europe UK USA ) confidence= 1.00
(Australia) ==> (Not-stated Europe UK USA ) confidence= 0.97
(Europe) ==> (Not-stated Australia UK USA ) confidence= 0.97
(UK) ==> (Not-stated Australia Europe USA ) confidence= 0.97
(USA) ==> (Not-stated Australia Europe UK ) confidence= 0.97
```

minsupport =0.001

Time = 0.05745s

Apriori Algorithm - Kaggle Data

min_support=0.001, min_conf=0.3, Time=0.121113s

```
(UK, Australia, Asia, Americas, USA, Oceania, Not-stated) ==> (Europe)  confidenc
e = 1.0
(UK, Australia, Europe, Americas, USA, Oceania, Not-stated) ==> (Asia)  confidenc
e = 1.0
(UK, Australia, Europe, Asia, USA, Oceania, Not-stated) ==> (Americas)  confidenc
e = 1.0
(UK, Australia, Europe, Asia, Americas, Oceania, Not-stated) ==> (USA)  confidenc
e = 1.0
(UK, Australia, Europe, Asia, Americas, USA, Not-stated) ==> (Oceania)  confidenc
e = 1.0
(UK, Australia, Europe, Asia, Americas, USA, Oceania) ==> (Not-stated)  confidenc
e = 0.974
(Australia, Asia, Americas, USA, Oceania, Europe) ==> (UK)  confidence = 0.974
(UK, Asia, Americas, USA, Oceania, Europe) ==> (Australia)  confidence = 0.974
(UK, Australia, Asia, Americas, USA, Oceania) ==> (Europe)  confidence = 0.974
(UK, Australia, Americas, USA, Oceania, Europe) ==> (Asia)  confidence = 0.974
(UK, Australia, Asia, USA, Oceania, Europe) ==> (Americas)  confidence = 0.974
(UK, Australia, Asia, Americas, Oceania, Europe) ==> (USA)  confidence = 0.974
(UK, Australia, Asia, Americas, USA, Europe) ==> (Oceania)  confidence = 0.974
```

Discussion - Kaggle Data

每年都會有固定的移民從紐西蘭去別的城市，沒有因為不同年代而有太大的改變，應該是資料前處理時，移民數設大於1000人就取出來的關係，會導致取出來的資料沒什麼代表性，數字要再取大一點，Frequent itemset才不會這麼多。