Data Analytics Week 6 Assignment

202011431 산업공학과 차승현

Association Rule Mining

건국대학교

Analysis Procedure

1) 필요한 패키지를 로드하고 파일을 읽어 data_lst라는 변수에 데이터셋을 생성한다.

In [18]:
```
import pandas as pd
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori
import csv
```

In [23]:
```
data = open('data_week6.txt', encoding='UTF-8')
csv_reader = csv.reader(data)

data_lst = []

[data_lst.append(num) for num in csv_reader]

data_lst
```

Out[23]:
```
[['식스센스', '반지의제왕1', '해리포터1', '해리포터2', '쇼생크탈출'],
 ['어벤져스', '스타워즈', '아바타'],
 ['반지의제왕1', '반지의제왕2'],
 ['어벤져스', '스타워즈', '식스센스'],
 ['어벤져스', '스타워즈', '식스센스', '해리포터1'],
 ['어벤져스', '아바타'],
 ['해리포터1', '해리포터2', '반지의제왕1', '반지의제왕2'],
 ['어벤져스', '스타워즈', '쇼생크탈출'],
 ['어벤져스', '스타워즈', '식스센스'],
 ['반지의제왕1', '식스센스', '어벤져스', '쇼생크탈출'],
 ['어벤져스', '스타워즈', '반지의제왕1', '반지의제왕2'],
 ['해리포터1', '해리포터2', '반지의제왕1', '반지의제왕2'],
 ['식스센스', '쇼생크탈출', '타이타닉'],
 ['쇼생크탈출', '아바타'],
 ['반지의제왕1', '반지의제왕2', '타이타닉'],
 ['해리포터1', '해리포터2'],
 ['해리포터1', '해리포터2', '쇼생크탈출'],
 ['어벤져스', '스타워즈', '타이타닉'],
 ['어벤져스', '아바타'],
 ['어벤져스', '아바타', '해리포터1'],
 ['해리포터1', '반지의제왕1'],
 ['타이타닉', '쇼생크탈출'],
 ['타이타닉', '식스센스'],
 ['어벤져스', '아바타', '타이타닉'],
 ['해리포터1', '타이타닉']]
```

2) 가나다 순으로 Column값을 생성하여, (반지의제왕1, 반지의제왕2, 쇼생크

탈출, …, 해리포터1, 해리포터2)에서 n번째(n= number of datasets)의 데이터에 해당 column이 표시되어 있다면, True값으로 표기하고, 없다면 False값으로 표기한다.

```
te = TransactionEncoder()
te_ary = te.fit(data_lst).transform(data_lst)
df = pd.DataFrame(te_ary, columns=te.columns_) #위에서 나온걸 보기 좋게 데이터프레임으로 변경
```

```
df
```

	반지의제왕1	반지의제왕2	쇼생크탈출	스타워즈	식스센스	아바타	어벤져스	타이타닉	해리포터1	해리포터2
0	True	False	True	False	True	False	False	False	True	True
1	False	False	False	True	False	True	True	False	False	False
2	True	True	False	False	False	False	False	False	False	False
3	False	False	False	True	True	False	True	False	False	False
4	False	False	False	True	True	False	True	False	True	False
5	False	False	False	False	False	True	True	False	False	False
6	True	True	False	False	False	False	False	False	True	True
7	False	False	True	True	False	False	True	False	False	False
8	False	False	False	True	True	False	True	False	False	False
9	True	False	True	False	True	False	True	False	False	False
10	True	True	False	True	False	False	True	False	False	False
11	True	True	False	False	False	False	False	False	True	True
12	False	False	True	False	True	False	False	True	False	False
13	False	False	True	False	False	True	False	False	False	False
14	True	True	False	False	False	False	False	True	False	False
15	False	False	False	False	False	False	False	False	True	True
16	False	False	True	False	False	False	False	False	True	True
17	False	False	False	True	False	False	True	True	False	False
18	False	False	False	False	False	True	True	False	False	False
19	False	False	False	False	False	True	True	False	True	False
20	True	False	False	False	False	False	False	False	True	False
21	False	False	True	False	False	False	False	True	False	False
22	False	False	False	False	True	False	False	True	False	False
23	False	False	False	False	False	True	True	True	False	False
24	False	False	False	False	False	False	False	True	True	False

3) 최소 표시 지지도(support)를 0.5로 설정하여 apriori algorithm을 실행

하면 다음과 같은 결과가 출력된다.

```
frequent_itemsets = apriori(df, min_support=0.5, use_colnames=True)
frequent_itemsets
```

	support	itemsets

4) 3)의 결과는 최소 지지도를 너무 높게 설정하여 발생한 issue이므로, 최소 지지도를 낮게 설정하여 다시 algorithm을 실행하였다.

```
frequent_itemsets = apriori(df, min_support=0.01, use_colnames=True)
frequent_itemsets
```

	support	itemsets
0	0.32	(반지의제왕1)
1	0.20	(반지의제왕2)
2	0.28	(쇼생크탈출)
3	0.28	(스타워즈)
4	0.28	(식스센스)
...	...	...
82	0.04	(해리포터1, 쇼생크탈출, 반지의제왕1, 해리포터2)
83	0.04	(해리포터1, 반지의제왕1, 식스센스, 해리포터2)
84	0.04	(해리포터1, 쇼생크탈출, 식스센스, 해리포터2)
85	0.04	(해리포터1, 어벤져스, 스타워즈, 식스센스)
86	0.04	(식스센스, 해리포터1, 쇼생크탈출, 해리포터2, 반지의제왕1)

5) 적당한 최소 지지도(0.12)를 설정하여 반복적으로 algorithm을 실행한 결과이다.

```
frequent_itemsets = apriori(df, min_support=0.12, use_colnames=True)
frequent_itemsets
```

	support	itemsets
0	0.32	(반지의제왕1)
1	0.20	(반지의제왕2)
2	0.28	(쇼생크탈출)
3	0.28	(스타워즈)
4	0.28	(식스센스)
5	0.24	(아바타)
6	0.48	(어벤져스)
7	0.28	(타이타닉)
8	0.36	(해리포터1)
9	0.20	(해리포터2)
10	0.20	(반지의제왕1, 반지의제왕2)
11	0.16	(해리포터1, 반지의제왕1)
12	0.12	(반지의제왕1, 해리포터2)
13	0.12	(식스센스, 쇼생크탈출)
14	0.12	(식스센스, 스타워즈)
15	0.28	(어벤져스, 스타워즈)
16	0.16	(식스센스, 어벤져스)
17	0.20	(어벤져스, 아바타)
18	0.20	(해리포터1, 해리포터2)
19	0.12	(해리포터1, 반지의제왕1, 해리포터2)
20	0.12	(식스센스, 어벤져스, 스타워즈)

위의 결과로 보아, (in 0 row) 반지의제왕1을 택할 확률은 0.32이고, (in 10 row) (반지의제왕1, 반지의제왕2)를 같이 택할 확률이 0.20이다.

6) 신뢰도의 한계점이 0.5인 것들로 추리면 다음과 같다.

```
from mlxtend.frequent_patterns import association_rules
association_rules(frequent_itemsets, metric="confidence", min_threshold=0.5)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(반지의제왕1)	(반지의제왕2)	0.32	0.20	0.20	0.625000	3.125000	0.1360	2.133333
1	(반지의제왕2)	(반지의제왕1)	0.20	0.32	0.20	1.000000	3.125000	0.1360	inf
2	(반지의제왕1)	(해리포터1)	0.32	0.36	0.16	0.500000	1.388889	0.0448	1.280000
3	(해리포터2)	(반지의제왕1)	0.20	0.32	0.12	0.600000	1.875000	0.0560	1.700000
4	(어벤져스)	(스타워즈)	0.48	0.28	0.28	0.583333	2.083333	0.1456	1.728000
5	(스타워즈)	(어벤져스)	0.28	0.48	0.28	1.000000	2.083333	0.1456	inf
6	(식스센스)	(어벤져스)	0.28	0.48	0.16	0.571429	1.190476	0.0256	1.213333
7	(아바타)	(어벤져스)	0.24	0.48	0.20	0.833333	1.736111	0.0848	3.120000
8	(해리포터1)	(해리포터2)	0.36	0.20	0.20	0.555556	2.777778	0.1280	1.800000
9	(해리포터2)	(해리포터1)	0.20	0.36	0.20	1.000000	2.777778	0.1280	inf
10	(해리포터1, 반지의제왕1)	(해리포터2)	0.16	0.20	0.12	0.750000	3.750000	0.0880	3.200000
11	(해리포터1, 해리포터2)	(반지의제왕1)	0.20	0.32	0.12	0.600000	1.875000	0.0560	1.700000
12	(반지의제왕1, 해리포터2)	(해리포터1)	0.12	0.36	0.12	1.000000	2.777778	0.0768	inf
13	(해리포터2)	(해리포터1, 반지의제왕1)	0.20	0.16	0.12	0.600000	3.750000	0.0880	2.100000
14	(식스센스, 어벤져스)	(스타워즈)	0.16	0.28	0.12	0.750000	2.678571	0.0752	2.880000
15	(식스센스, 스타워즈)	(어벤져스)	0.12	0.48	0.12	1.000000	2.083333	0.0624	inf

7) 결과 해석

X->Y(antecedents -> consequents)의 관계에서, support(지지도)의 측면으로 보았을 때(antecedent support와 consequent support가 아닌 일반적 support), 가장 높은 영화는 (어벤져스, 스타워즈)와 (스타워즈, 어벤져스)가 0.28의 값으로 가장 우수했다.

Confidence(신뢰도)의 측면에서 보았을 때, (반지의제왕2, 반지의제왕1), (스타워즈, 어벤져스), (해리포터2, 해리포터1), ((반지의제왕1, 해리포터2), 해리포터1), ((식스센스, 스타워즈), (어벤져스))가 1.0의 값으로 가장 우수했다. 이는 antecedents의 영화를 택한 사람들이 100%의 확률로 consequents의 영화를 택했다고 해석할 수 있다.

Lift(향상도)의 측면에서 보았을 때, ((해리포터1, 반지의제왕1), 해리포터2), (해리포터2, (해리포터1, 반지의제왕1))이 3.75의 값으로 가장 우수했다. 그러나, 모든 row의 값이 lift가 1을 초과하므로 모든 antecedents와 consequents가 강한 양의 상관관계를 가지고 있다고 해석할 수 있다.

```
as_rules.sort_values('support', ascending=False)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
4	(어벤져스)	(스타워즈)	0.48	0.28	0.28	0.583333	2.083333	0.1456	1.728000
5	(스타워즈)	(어벤져스)	0.28	0.48	0.28	1.000000	2.083333	0.1456	inf
0	(반지의제왕1)	(반지의제왕2)	0.32	0.20	0.20	0.625000	3.125000	0.1360	2.133333
1	(반지의제왕2)	(반지의제왕1)	0.20	0.32	0.20	1.000000	3.125000	0.1360	inf
7	(아바타)	(어벤져스)	0.24	0.48	0.20	0.833333	1.736111	0.0848	3.120000
8	(해리포터1)	(해리포터2)	0.36	0.20	0.20	0.555556	2.777778	0.1280	1.800000
9	(해리포터2)	(해리포터1)	0.20	0.36	0.20	1.000000	2.777778	0.1280	inf

```
as_rules.sort_values('confidence', ascending=False)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	(반지의제왕2)	(반지의제왕1)	0.20	0.32	0.20	1.000000	3.125000	0.1360	inf
5	(스타워즈)	(어벤져스)	0.28	0.48	0.28	1.000000	2.083333	0.1456	inf
9	(해리포터2)	(해리포터1)	0.20	0.36	0.20	1.000000	2.777778	0.1280	inf
12	(반지의제왕1, 해리포터2)	(해리포터1)	0.12	0.36	0.12	1.000000	2.777778	0.0768	inf
15	(식스센스, 스타워즈)	(어벤져스)	0.12	0.48	0.12	1.000000	2.083333	0.0624	inf
7	(아바타)	(어벤져스)	0.24	0.48	0.20	0.833333	1.736111	0.0848	3.120000

```
as_rules.sort_values('lift', ascending=False)
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
10	(해리포터1, 반지의제왕1)	(해리포터2)	0.16	0.20	0.12	0.750000	3.750000	0.0880	3.200000
13	(해리포터2)	(해리포터1, 반지의제왕1)	0.20	0.16	0.12	0.600000	3.750000	0.0880	2.100000
0	(반지의제왕1)	(반지의제왕2)	0.32	0.20	0.20	0.625000	3.125000	0.1360	2.133333
1	(반지의제왕2)	(반지의제왕1)	0.20	0.32	0.20	1.000000	3.125000	0.1360	inf
8	(해리포터1)	(해리포터2)	0.36	0.20	0.20	0.555556	2.777778	0.1280	1.800000
9	(해리포터2)	(해리포터1)	0.20	0.36	0.20	1.000000	2.777778	0.1280	inf
12	(반지의제왕1, 해리포터2)	(해리포터1)	0.12	0.36	0.12	1.000000	2.777778	0.0768	inf
14	(식스센스, 어벤져스)	(스타워즈)	0.16	0.28	0.12	0.750000	2.678571	0.0752	2.880000
4	(어벤져스)	(스타워즈)	0.48	0.28	0.28	0.583333	2.083333	0.1456	1.728000
5	(스타워즈)	(어벤져스)	0.28	0.48	0.28	1.000000	2.083333	0.1456	inf
15	(식스센스, 스타워즈)	(어벤져스)	0.12	0.48	0.12	1.000000	2.083333	0.0624	inf
3	(해리포터2)	(반지의제왕1)	0.20	0.32	0.12	0.600000	1.875000	0.0560	1.700000
11	(해리포터1, 해리포터2)	(반지의제왕1)	0.20	0.32	0.12	0.600000	1.875000	0.0560	1.700000
7	(아바타)	(어벤져스)	0.24	0.48	0.20	0.833333	1.736111	0.0848	3.120000
2	(반지의제왕1)	(해리포터1)	0.32	0.36	0.16	0.500000	1.388889	0.0448	1.280000
6	(식스센스)	(어벤져스)	0.28	0.48	0.16	0.571429	1.190476	0.0256	1.213333