

연관 분석 개념

예) 기저귀-맥주(미국 월마트 분석)

1. 고객들은 어떤 상품들을 동시에 구매하는가?
2. 라면을 구매한 고객은 주로 다른 어떤 상품을 구매하는가?

위와 같은 질문에 대한 분석을 토대로 고객들에게 SMS를 보낸다든가, 판촉용 전화를 한다든가 묶음 판매를 기획함.

이와 같은 질문에 대한 답은 연관규칙을 이용하여 구할 수 있습니다. 연관규칙은 상업 데이터베이스에서 가장 흔히 쓰이는 도구로, 어떤 사건이 얼마나 자주 동시에 발생하는가를 표현하는 규칙 또는 조건을 의미.

support 지지도는 품목 A와 B를 동시에 구매할 확률인 $P(A \cap B)$ 를 나타냅니다

confidence: 신뢰도는 품목 A가 구매하고나서, 품목 B가 구매될 확률

lift 향상도는 A를 구매한 사람이 B를 구매할 확률과 A의 구매와 상관없이 B를 구매할 확률의 비율

lift > 1 이면 관련도가 높고 **lift < 1** 이면 A구매자가 B를 구매하지 않을 확률이 높음

참고사이트: <https://ratsgo.github.io/machine%20learning/2017/04/08/apriori/>

*연관분석, 장바구니 분석

***지지도(Support)**: 전체 집합군에서 [조건] 자료가 포함된 집합수, 비율,
[조건1]자료수 / 전체자료수

***신뢰도(Confidence)**: [조건1]가 있을때 [조건2]도 같이 있는 확률
[조건1]->[조건2] 라고 하면
[조건1],[조건2] 가 같이나온 자료수/[조건1] 자료수

즉: [조건1],[조건2] 지지도 / [조건1] 지지도

***향상도(Lift:Improvement)**:
[조건1][조건2]가 같이 나온 자료수/[조건1]자료수/전체자료수

https://m.blog.naver.com/PostView.nhn?blogId=leedk1110&logNo=220785911828&proxyReferer=http%3A%2F%2Fwww.google.co.kr%2Furl%3Fsa%3Dt%26rct%3Dj%26q%3D%26esrc%3Ds%26source%3Dweb%26cd%3D3%26ved%3D2ahUKEwji-bmbtMTcAhVD_GEKHSv5DyIQFjACegQIARAB%26url%3Dhttp%253A%252F%252Fm.blog.naver.com%252Fleedk1110%252F220785911828%26usg%3DAOvVaw2dQ91WI-N0WuNhdsE3wRbj

=> 지지도, 신뢰도, 향상도 개념잡기

<http://rfriend.tistory.com/190>

<http://rfriend.tistory.com/191>

<http://rfriend.tistory.com/194?category=706118>

<http://rfriend.tistory.com/193>

=> 데이터 프레임형태로 되어 있는 자료일때.

<http://blog.daum.net/sys4ppl/6>

=> 데이터 전처리 필요할때(NA등)

<http://leebaro.tistory.com/entry/Association-Analysis-Association-Rule-Apriori-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98-2-of-3>

=> 구글검색

filetype:pdf R연관분석

=>전체정리

<http://blog.naver.com/PostView.nhn?blogId=leedk1110&logNo=220788082381&parentCategoryNo=&categoryNo=8&viewDate=&isShowPopularPosts=false&from=postView>

연관분석 - 화장품전문점 패키지 구성방법?

분류	내용
예제 데이터	<ul style="list-style-type: none"> ■ B화장품전문점에서 판매된 트랜잭션 데이터
변수명	<ul style="list-style-type: none"> ■ 단일변수 <ul style="list-style-type: none"> - Nail Polish(매니큐어), Brushes(브러시), - Concealer(컨실러: 피부 결점을 감추어 주는 화장품) - Bronzer(피부를 햇볕에 그을린 것처럼 보이게 하는 화장품) - Lip liner(입술 라이너), Mascara(마스카라: 속눈썹용 화장품) - Eye shadow(아이섀도: 눈꺼풀에 바르는 화장품) - Foundation(파운데이션: 가루분), Lip Gloss(립글로스: 입술 화장품) - Lipstick(립스틱), Eyeliner(아이 라이너: 눈의 윤곽 그림)
분석문제	<ul style="list-style-type: none"> ■ 전체 트랜잭션 개수와 상품아이템 유형은 몇 개인가? ■ 가장 발생빈도가 높은 상품아이템은 무엇인가? ■ 지지도를 10%로 설정했을 때의 생성되는 규칙의 가지수는? ■ 상품아이템 중에서 가장 발생확률이 높은 아이템과 낮은 아이템은 무엇인가? ■ 가장 발생가능성이 높은 <2개 상품간>의 연관규칙은 무엇인가? ■ 가장 발생가능성이 높은 <2개 상품이상에서> <제3의 상품으로>의 연관규칙은?

사과, 치즈, 생수
 생수, 호두, 치즈, 고등어
 수박, 사과, 생수
 생수, 호두, 치즈, 옥수수

사과를 구매한 고객이 치즈도 함께구매할 연관성에 대해 분석

지지도 = $P(A \cap B)$ 신뢰도 = $P(A \cap B) / P(A)$ 향상도 = $\text{신뢰도}(A, B) / \text{지지도}(B)$

▶ 지지도=[사과][치즈]가 같이 나온 자료/전체자료 => 1/4 => 0.25

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
3	고등어
	수박
	사과
4	생수
	호두
	치즈
	옥수수

▶ 신뢰도=[사과][치즈]가 같이 나온 자료/[사과]자료 => 1/2 => 0.5

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
3	고등어
	수박
	사과
4	생수
	호두
	치즈
	옥수수

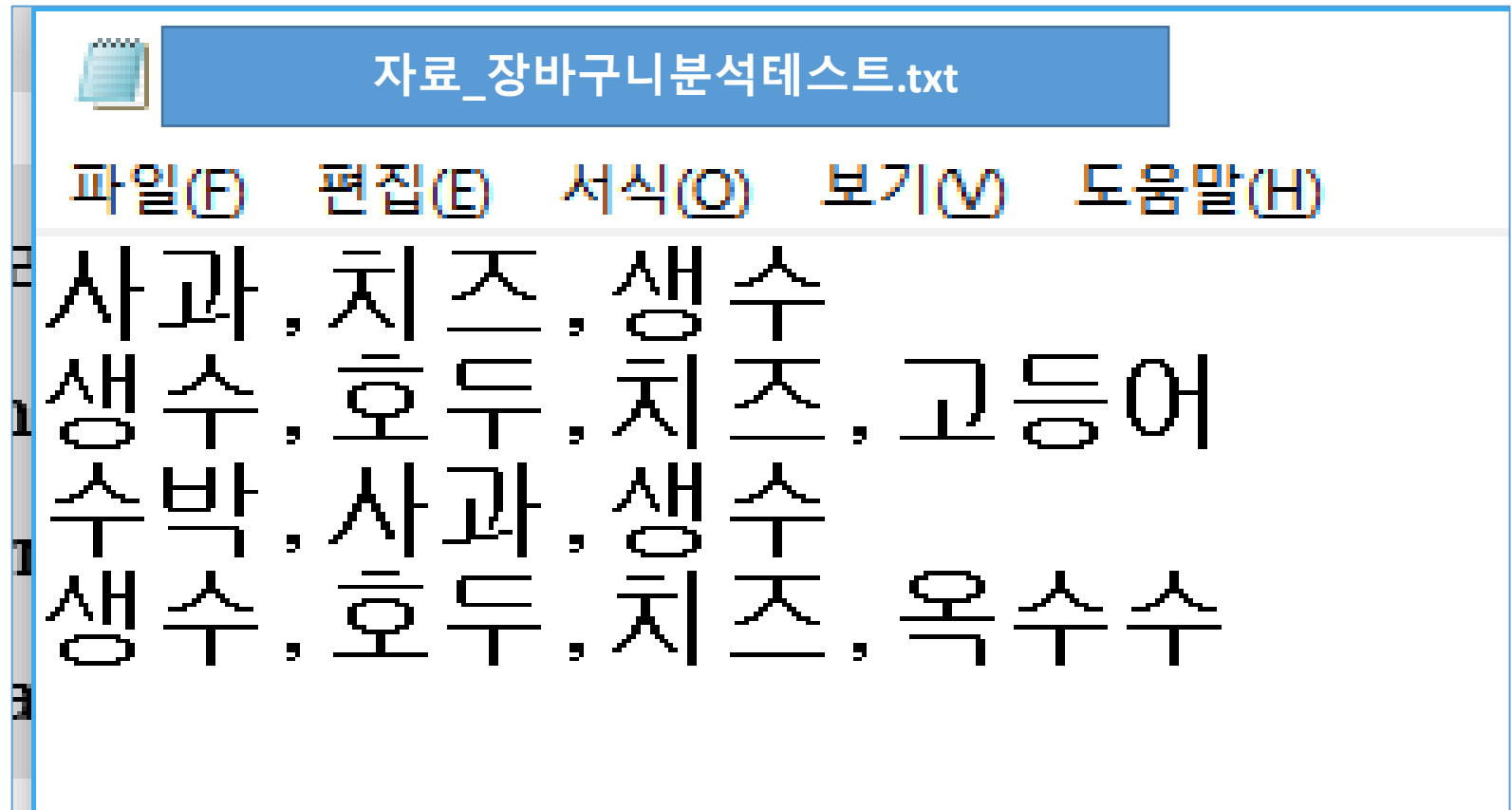
▶ 향상도 = $0.5 / 0.75 = 0.6666667$

구매자번호	제품명
1	사과
	치즈
	생수
2	생수
	호두
	치즈
3	고등어
	수박
	사과
4	생수
	호두
	치즈
	옥수수

항목별 지지도[Support]			
번호	제품명	지지도(자료수/4)	
1	고등어	1	0.25
2	사과	2	0.5
3	생수	4	1
4	수박	1	0.25
5	옥수수	1	0.25
6	치즈	3	0.75
7	호두	2	0.5

/연관분석 폴더안에 '자료_장바구니분석테스트.txt' 파일 있는지 확인

*쉼표로 각 항목이 분리된 자료이며 enter가 들어간(개행키) 위치까지가 한 행자료가 됨. 거래장부 데이터가 아래와 같이 쉼표로 분리되어 나오는 경우가 있음



```
install.packages("arules")
library(arules)

setwd("c:/data_r") # 디렉토리는 상황에 맞추어서 세팅
tr<-read.transactions (" 자료_장바구니분석테스트.txt",format="basket",sep=",")
tr
#지지도, 향상도 0.1 이상 자료 (0.1은 10%를 의미함 숫자값은 사용자가 임의로 넣음)
rules=apriori(tr,parameter=list(supp=0.1,conf=0.1))
inspect(rules)
```

15개의 규칙이 발견되었음을 의미함
여기 0이 나오면
지지도향상도 최소값
을 더 작게 변경해야
함

```
> rules=apriori(tr,parameter=list(supp=0.3,conf=0.3)) #
Apriori

Parameter specification:
confidence minval smax arem aval originalSupport maxtime support minlen maxlen target ext
0.3 0.1 1 none FALSE TRUE 5 0.3 1 10 rules FALSE

Algorithmic control:
filter tree heap memopt load sort verbose
0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 1

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[7 item(s), 4 transaction(s)] done [0.00s].
sorting and recoding items ... [4 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [15 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
> inspect(rules)
```

	lhs	rhs	support	confidence	lift	count
[1]	{}	{사과}	0.50	0.5000000	1.0000000	2
[2]	{}	{호두}	0.50	0.5000000	1.0000000	2
[3]	{}	{치즈}	0.75	0.7500000	1.0000000	3
[4]	{}	{생수}	1.00	1.0000000	1.0000000	4
[5]	{사과}	{생수}	0.50	1.0000000	1.0000000	2
[6]	{사과}	{치즈}	0.50	0.5000000	1.0000000	2
[7]	{호두}	{치즈}	0.50	1.0000000	1.3333333	2
[8]	{호두}	{치즈}	0.50	0.6666667	1.3333333	2
[9]	{호두}	{생수}	0.50	1.0000000	1.0000000	2
[10]	{생수}	{치즈}	0.50	0.5000000	1.0000000	2
[11]	{생수}	{호두}	0.50	0.5000000	1.0000000	2
[12]	{생수}	{치즈}	0.75	0.7500000	1.0000000	3
[13]	{치즈}	{호두}	0.50	1.0000000	1.0000000	2
[14]	{치즈}	{호두}	0.50	1.0000000	1.3333333	2
[15]	{치즈}	{호두}	0.50	0.6666667	1.3333333	2

지지도

신뢰도

향상도

지지도, 신뢰도30%이상인 15개의 자료나옴
사과,치즈는 지지도가 0.25 이므로 나타나지 않음

치즈->생수
지지도: 0.75
신뢰도: 0.75
향상도: 1

#10개 항목만 보기 앞쪽의 Rules에서 10개 미만일때
##아래와 같이 1:10을 하면 에러나옴. 본인의 상황에
맞추어서 개수를 작업해야함.

```
inspect(rules[1:10])
```

```
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로
```

```
set item appearances ... [10 item(s)] done [0.00s].
set transactions ... [7 item(s), 4 transactions] done [0.00s].
sorting and recoding items ... [4 items] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
[15 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
> inspect(rules)
```

아래내용은 선택사항임. 필수작업 아님

```
연관결과<-inspect(sort(rules,by="lift"))
```

```
head(연관결과)
```

```
subset(연관결과,subset=(lift>=1))
```

#lift(향상도) 값이 1이상인값만 추출

#lift(향상도) 값이 100이상이면서 support(지지도)가 50이상

```
subset(연관결과,subset=(lift>=1 & support>=0.5))
```

```
#####
```

```
##### grep(검색단어, 검색위치),
```

```
##### -grep(검색단어,검색위치) : -는 제외하고 뜻임
```

```
#####
```

```
연관결과[grep("사과",연관결과$lhs),] # lhs 변수에 '사과' 가 포함된 자료만 추출
```

```
연관결과[-grep("사과",연관결과$lhs),] # lhs 변수에 '사과' 글자 없는 자료만 추출
```

```
사과연관분석<-연관결과[grep("사과",연관결과$lhs),]
```

```
사과_lift_1이상<-subset(사과연관분석,subset=(lift>=1))
```

```
사과_lift_1이상
```

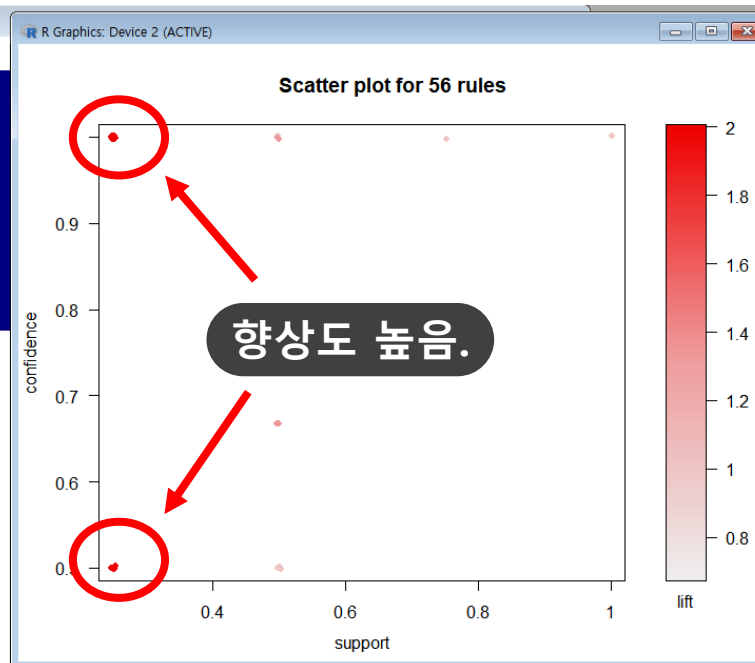
```
#####
```

```
#####필요한 항목만 csv로 저장가능 #####
```

```
#####
```

```
write.csv(연관결과, "c:/data_r/연관분석결과.csv")
```

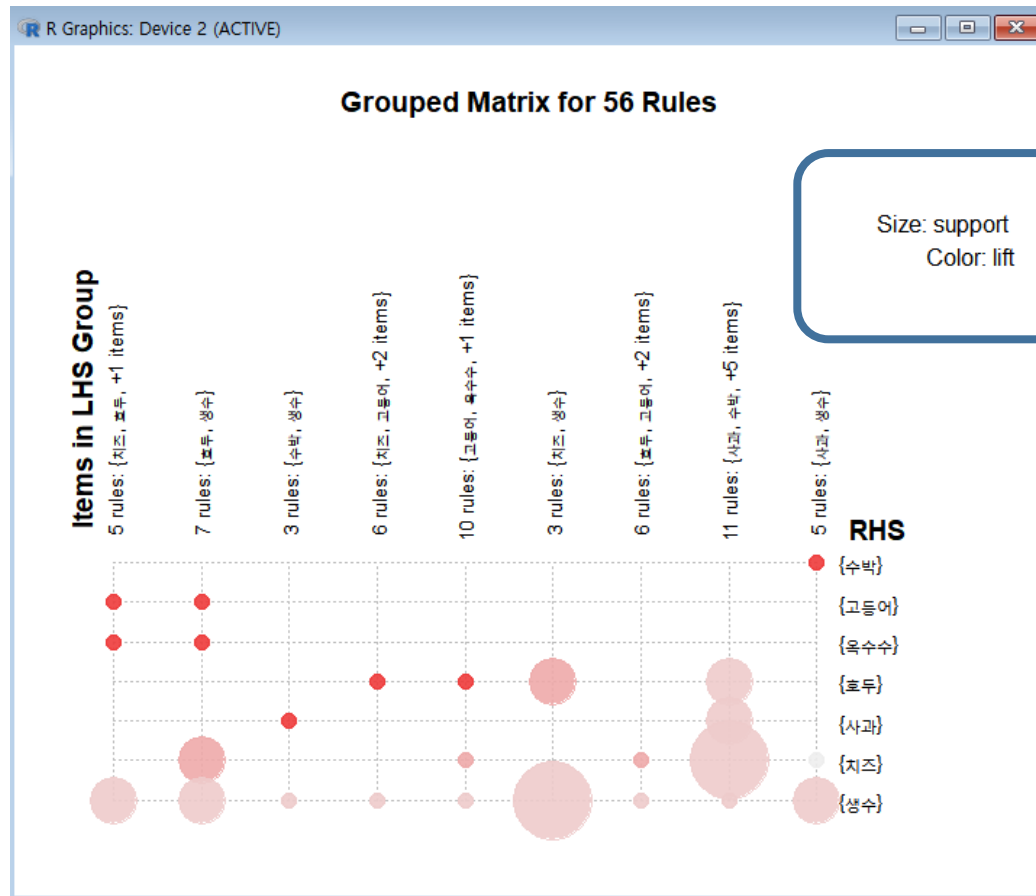
가로(지지도), 세로(신뢰도), 색상(향상도)
#아래 자료는 지지도 0.25, 신뢰도 0.5와 1일때 향상도가 높음, 진한빨강색이 표시됨.

[illegible]

#매트릭스차트

lhs(가로축)-조건(x아이템)과 rhs(세로축)-결과(y아이템) 으로구성한매트릭스그래프

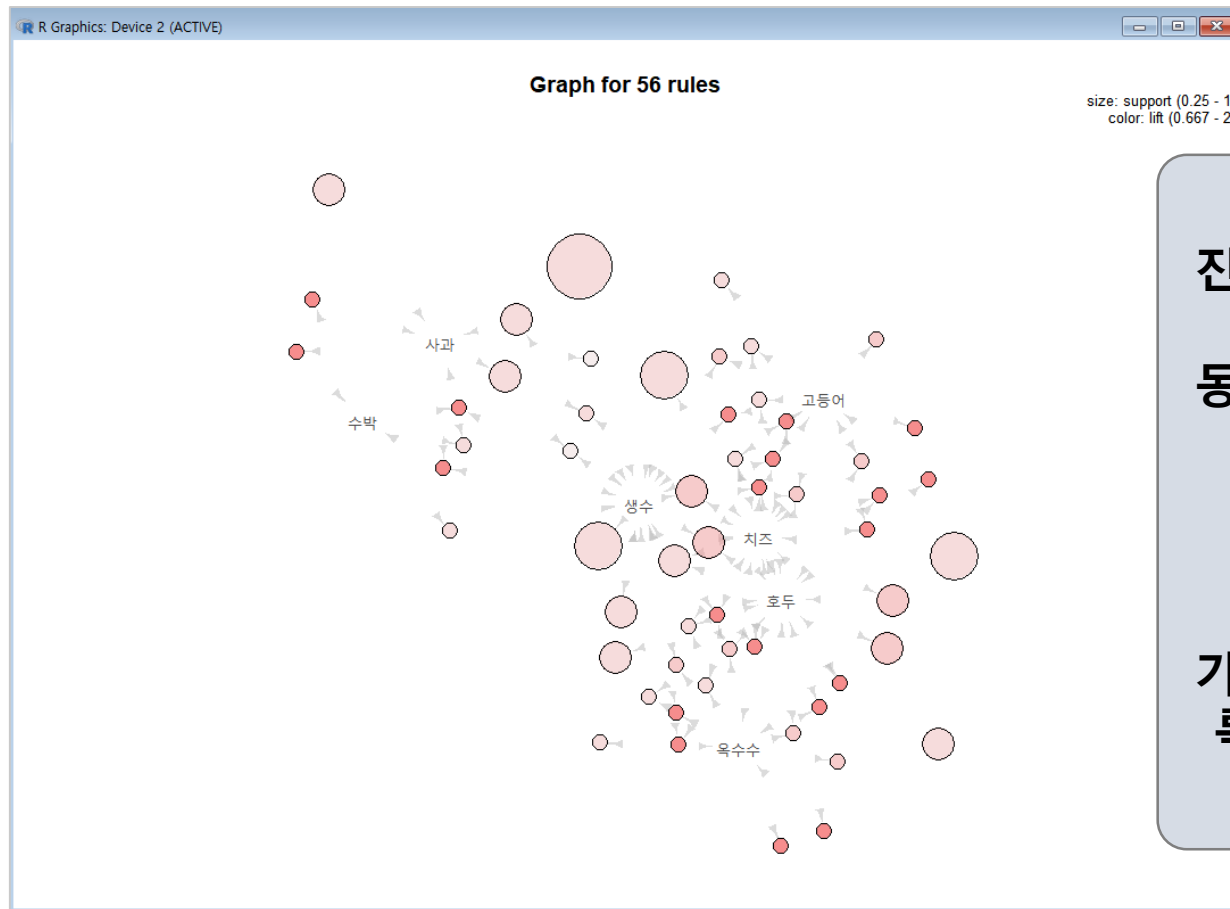
```
plot(rules,method="grouped")
```



진한 빨간색일수록 향상도가
높은 자료임.
동그라미가 클수록 지지도

(많이 나온 빈도수가
높은 자료임)

```
# 각규칙별로어떤아이템들이 연관되어뭉여있는지 보여주는네트워크그래프
#네트워크차트
plot(rules,method="graph")
```



**진한 빨간색일수록 향상도가 높은 자료임.
동그라미가 클수록 지지도**

(많이 나온 빈도수가 높은 자료임)

가까운곳에 있는 자료일수록 연관도가 높은 자료임

연관분석 (데이터 프레임일때)

- 데이터 프레임구조는 ID(구매고객)를 기준으로 item을 나누어서 작업해야함.
- 아래 사이트 자료임. 파일명: dvdtrans.csv

<http://blog.daum.net/sys4ppl/6>

dvdtrans.csv

자료_빅카인즈_1만.txt

자료_구매내역_1만.txt

A1

A

B

C

1	ID	Item
2	1	Sixth Sense
3	1	LOTR1
4	1	Harry Potter1
5	1	Green Mile
6	1	LOTR2
7	2	Gladiator
8	2	Patriot
9	2	Braveheart
10	3	LOTR1
11	3	LOTR2
12	4	Gladiator
13	4	Patriot

좌측의 데이터 프레임을
우측의 구조로 변경하는 작업이 필요함.

	A	B		A	B	C	D
1	장바구니번호	구매내역	→	1	넥타이	셔츠	양말
2		1 넥타이		2	양말	벨트	장갑
3		1 셔츠		3	지갑	넥타이	셔츠
4		1 양말		4	양말	벨트	장갑
5		2 양말		5			바지
6		2 벨트					
7		2 장갑					
8		2 셔츠					
9		3 지갑					
10		3 넥타이					
11		3 셔츠					
12		4 양말					
13		4 벨트					
14		4 장갑					
15		4 바지					

<http://blog.daum.net/sys4ppl/6>

메모리 지우기 연관분석 패키지설치

```
rm(list=ls())
```

```
install.packages("arules")  
library(arules)  
install.packages("arulesViz")  
library(arulesViz)
```

```
setwd("c:/data_r") # 본인 상황에 맞추어서 작업  
dvd<-read.csv ("dvdtrans.csv", as.is=TRUE)  
head(dvd)
```

```
dvd.list <- split(dvd$item, dvd$ID) # id를 기준으로 item을 나눔.  
dvd.list
```

```
dvd.trans <- as(dvd.list, "transactions")  
dvd.trans
```

```
#지지도, 향상도 0.1 이상 자료 (0.1은 10%를 의미함 숫자값은 사용자가 임의로 넣음)  
rules=apriori(dvd.trans,parameter=list(supp=0.05,conf=0.05))  
inspect(rules)  
inspect(rules[1:10])  
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로 10개
```

```
# 가로(지지도), 세로(신뢰도), 색상(향상도)  
#아래 자료는 지지도 0.25, 신뢰도 0.5와 1일때 향상도가 높음, 진한빨강색이 표시됨.
```

```
plot(rules)
```

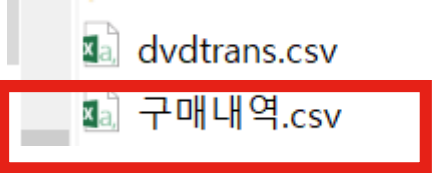
```
#매트릭스차트  
# lhs(가로축)-조건(x아이템)과 rhs(세로축)-결과(y아이템) 으로구성한매트릭스그래프
```

```
plot(rules,method="grouped")
```

```
# 각규칙별로어떤아이템들이 연관되어있어있는지 보여주는네트워크그래프  
plot(rules,method="graph") #네트워크차트
```

연관분석 에러

<http://blog.naver.com/PostView.nhn?blogId=gywlsangel&logNo=221325303378&parentCategoryNo=&categoryNo=88&viewDate=&isShowPopularPosts=true&from=search>



번호	고객명	구매항목
1	홍길동	새우깡
2	홍길동	맛동산
3	홍길동	맥주
4	일지매	짬구
5	일지매	감자깡
6	강감찬	감자깡
7	강감찬	새우깡
8	전우치	자갈치
9	전우치	맛동산
10	홍길동	짬구
11	어우동	빠다코코넛
12	어우동	맛동산
13	강감찬	포카칩
14	강감찬	맥주
15	김유신	자갈치
16	김유신	짬구
17	김유신	맛동산
18	홍길동	맛동산
19	전우치	초코칩쿠키
20	강감찬	크라운산도

구매항목을 고객명으로 나누면

홍길동 고객
새우깡, 맛동산, 맥주, 맛동산

- ➔ 으로 맛동산 항목이 중복됨.
- ➔ 연관분석은 중복데이터가 있으면 에러

위의 사이트를 참조하여서 반드시 실습

자료확인: '자료_빅카인즈_햇반.txt'

빅카인즈(신문기사 분석 플랫폼)에서 단어가 다 분리된 자료를 이용하여서도 연관분석 가능함.
이자료는 '햇반' 키워드로 자료 다운로드한 자료의 p열 특성추출은 Q열의 본문자료를 단어단위로
썹표분리한 자료임.

직접 SNS 상의 자료를 가져와서, 텍스트를 단어로 분리하여서 아래와 같이 단어, 단어, 단어 로 분리하는 작업을 하여서 텍스트간의 연관성을 찾는 ‘텍스트 마이닝’의 한 작업임

표시 형식	스타일					
이마트,이벤탈,선착순,e장날,구매고객,네이버,검색어,신라면,게티,햇반,할인쿠폰,반값쿠폰,달걀,실시간,급상승,기존회 갈무리,배송,차돌박이,수박,상품						
J	K	L	M	N	O	P
사건/사고	사건/사고	인물	위치	기관	키워드	특성추출
				이마트,네	이마트,e장날,이벤트,50%,제공,할	이마트,이벤탈,선착순
		김경연		CJ제일제당	CJ제일제당,CJ더마켓,가정간편식	온라인,식
		김경연		CJ제일제당	CJ제일제당,CJ더마켓,가정간편식	온라인,쿡킷,소비자,hr
		김경연		CJ더마켓	CJ10시,HMR,메뉴,CJ더마켓,오픈,CJ	온라인,cj더마켓,hmr,
		김경연		CJ제일제당	CJ제일제당,온라인,식품,사업,강화	쿡킷,cj,소비자,cj더마
		김경연		CJ제일제당	CJ제일제당,자체,온라인인물,강화,CJ	쿡킷,cj더마켓,소비자,
		김경연		CJ더마켓	CJ제일제당,CJ더마켓,강화,온라인	온라인,쿡킷,cj더마켓,
		김경연		CJ제일제당	CJ제일제당,온라인,식품,사업,강화	소비자,쿡킷,온라인,hr
		돈스파이크		킹스,서울	돈스파이크,다이어트,가능,작곡가	돈스파이크,2개,온종일
		돈스파이크		킹스,돈스파	1식,감량,돈스파이크,하루,음식	돈스파이크,2개,온라인
		돈스파이크		방송인	감량,돈스파이크,16kg,다이어트,온	돈스파이크,매일경제,
				MBN스타	돈스파이크,16kg,감량,공개,근황,현	돈스파이크,이목구비,
의일반>살인		고유정	제주	제주지검	고유정,3장,사진,전리품,기록,습관	고유정,제주,3장,휴대
의일반>살인		고유정	제주	제주지검	고유정,남편,살해,범행,장면,사진,계	고유정,고씨,줄피땀,전
의일반>살인			조천읍	제주동부검	고유정,보관,범행,장소,사진,혐의,남	고유정,줄피땀,완도행
의 사고>교통사고>해상		고유정	제주	충북,검찰	고유정,자신,행동,기록,카레,그릇,전	고씨,줄피땀,강씨,제주
의일반>살인			제주	검찰	고유정,남편,살해,사진,범행,장면,범	고유정,3장,피해자,햇
				덴마크,비	급부상,브랜드,비비고,짜파게티,글	칸탈월드패널,비비고, [
의일반>살인		고유정	청주 제	청주지검	고유정,사진,범행,장면,검찰,조사,관	고씨,줄피땀,강씨,고유

data	202
자료_빅카인즈_햇반.txt	202
자료_장바구니분석테스트.txt	202

엑셀자료중 P열의 특성추출
부분만 복사해서 붙임
txt로 저장

소스, 박카인즈, 햇반 txt - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

햇북, 햇반, 미니북, 소비자, 서울, cj 제일제당, 네이버, 1인, 문
 햇북, 햇반, 미니북, cj 제일제당, 관계자, 1인, 신개념, 텍스트
 햇북, 햇반, 소비자, 문지현, 시들, 맛들, 네이버, cj 제일제당,
 햇북, 햇반, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 네
 햇북, 햇반, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 네
 가정 간편식, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 네
 천, 손편지, 인천, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 네
 가격 인상, 식음료, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 영
 아산시, 박람회, 아산, 은행, 미니북, cj 제일제당, 1인, 서울, 혼밥, 소비자, 영
 편의점, gs25, 햇반, 밥솥, 세는날레는, 5인, 유어스, cu, 1만, 보
 곳 직원, 이재민, 이재민, 강원도, 영니트, 스타벅스, 김천수
 cj 제일제당, cj 제일제당, 봉사주간, 봉사활동, 구성원, 햇반, 폐기
 cj 제일제당, 임직원, 봉사활동, 봉사주간, 구성원, 햇반, 임직
 구성원, 봉사활동, 봉사주간, 구성원, 햇반, 임직원, 햇반, 임직원

중앙 기록 보관이 되게다. 남평 북 산타 http
 주 전남편 살해 사건'의 피의자 고유정 http
 남편 살해 혐의로 기소된 고유정이 범
 주 전 남편 살해 사건' 피의자 고유정(http
 남편을 살해한 혐의로 재판에 넘겨져 http
 지혜 [기자] 글로벌 마케팅 리서치업체 http
 북한 조사과정서 촬영 이윤 목자 '목우 http


```
install.packages("arules")
library(arules)
install.packages("arulesViz")
library(arulesViz)

setwd("c:/data_r")
tr<-read.transactions ("자료_빅카인즈_햇반.txt",format="basket",sep=",")
tr
#지지도, 향상도 0.1 이상 자료 (0.1은 10%를 의미함 숫자값은 사용자가 임의로 넣음)
rules=apriori(tr,parameter=list(supp=0.05,conf=0.05))
inspect(rules)
inspect(rules[1:10])
inspect(sort(rules,by="lift")[1:10]) # lift(향상도) 높은순으로 10개

# 가로(지지도), 세로(신뢰도), 색상(향상도)
#아래 자료는 지지도 0.25, 신뢰도 0.5와 1일때 향상도가 높음, 진한빨강색이 표시됨.

plot(rules)

#매트릭스차트
# lhs(가로축)-조건(x아이템)과 rhs(세로축)-결과(y아이템) 으로구성한매트릭스그래프

plot(rules,method="grouped")

# 각규칙별로어떤아이템들이 연관되어있어있는지 보여주는네트워크그래프
plot(rules,method="graph") #네트워크차트
```