연관 분석 개념

연관분석개념

예) 기저귀-맥주(미국월마트분석)

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∩B)를 나타냅니다

 # confidence: 신뢰도는 품목 A가 구매하고나서, 품목 B 가 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-N0WuNhsdE3wRbj

=> 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	▪ B화장품전문점에서 판매된 트랜잭션 데이터
변수명	▪ 단일변수 - 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	▪ 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?

연관성 분석(지지도, 신뢰도, 향상도) 계산

연관분석개념

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∩B) **신뢰도=P(A∩B)/P(A)** **향상도=신뢰도(A,B)/지지도(B)**

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

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

▶ 향상도=0.5/0.75=0.6666667

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

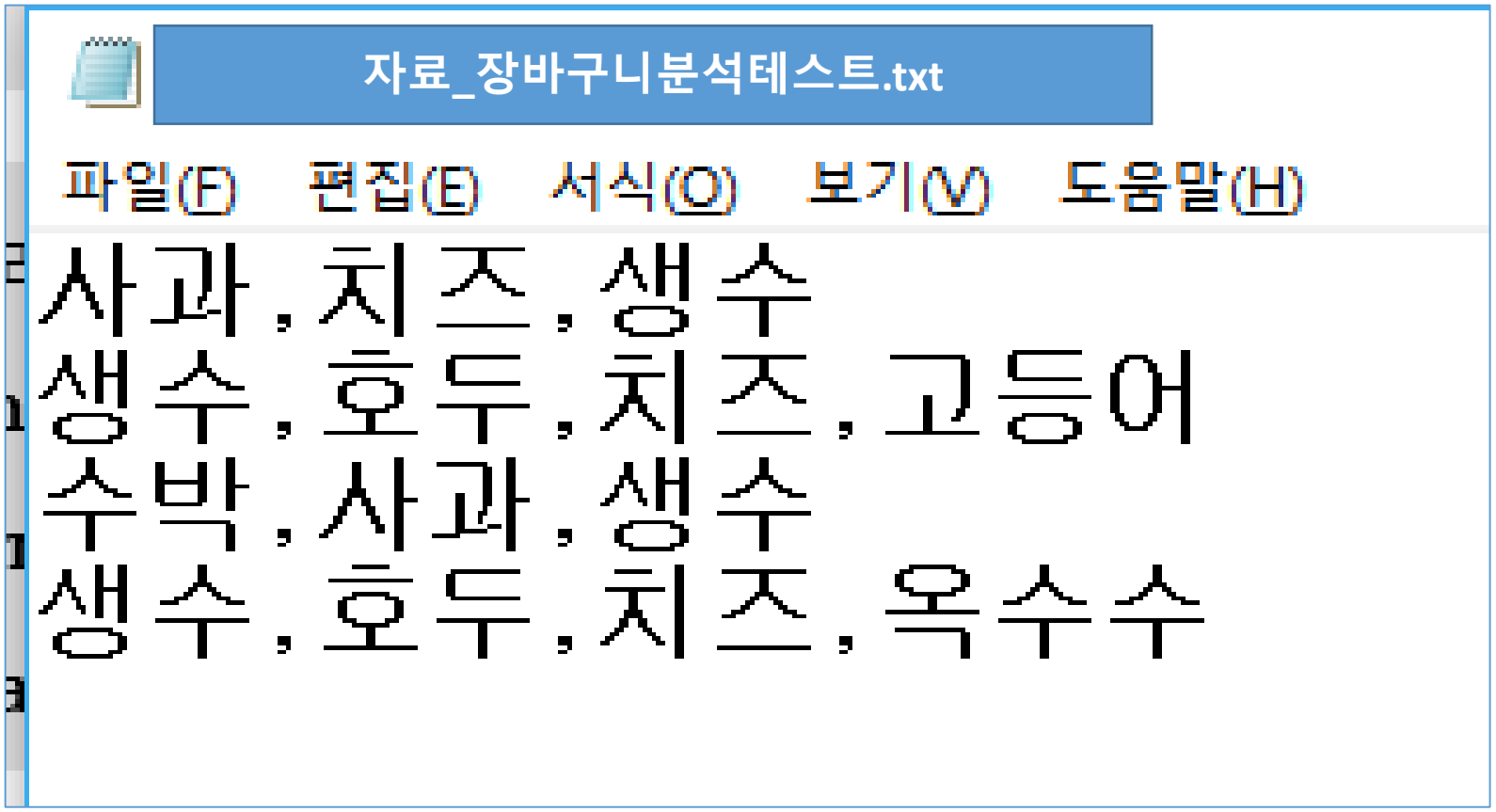
항목별 지지도[Support]

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

1. txt 자료 준비

/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



연관성 분석 – R(연관분석할수 있게 트랜잭션으로 읽기)

```
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 최소값을 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  1.000000 2
[2]  {}           => {호두} 0.50    0.5000000  1.000000 2
[3]  {}           => {치즈} 0.75    0.7500000  1.000000 3
[4]  {}           => {생수} 1.00    1.0000000  1.000000 4
[5]  {사과}       => {생수} 0.50    1.0000000  1.000000 2
[6]  {생수}       => {사과} 0.50    0.5000000  1.000000 2
[7]  {호두}       => {치즈} 0.50    1.0000000  1.333333 2
[8]  {치즈}       => {호두} 0.50    0.6666667  1.333333 2
[9]  {호두}       => {생수} 0.50    1.0000000  1.000000 2
[10] {생수}       => {호두} 0.50    0.5000000  1.000000 2
[11] {치즈}       => {생수} 0.75    1.0000000  1.000000 3
[12] {생수}       => {치즈} 0.75    0.7500000  1.000000 3
[13] {치즈,호두}  => {생수} 0.50    1.0000000  1.000000 2
[14] {생수,호두}  => {치즈} 0.50    1.0000000  1.333333 2
[15] {생수,치즈}  => {호두} 0.50    0.6666667  1.333333 2
```

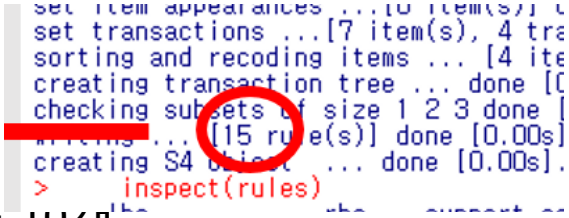
지지율 | 신뢰도 | 향상도

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 10개
```

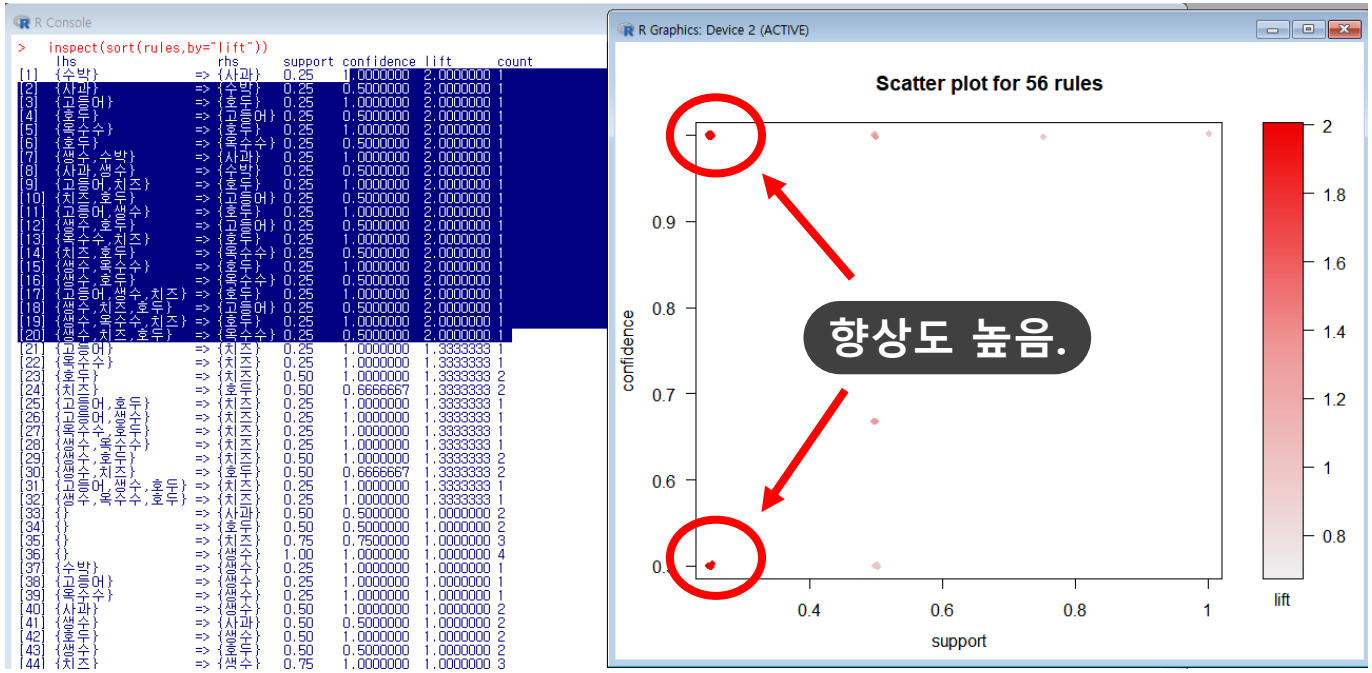


아래내용은 선택사항임. 필수작업 아님

```
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")
```

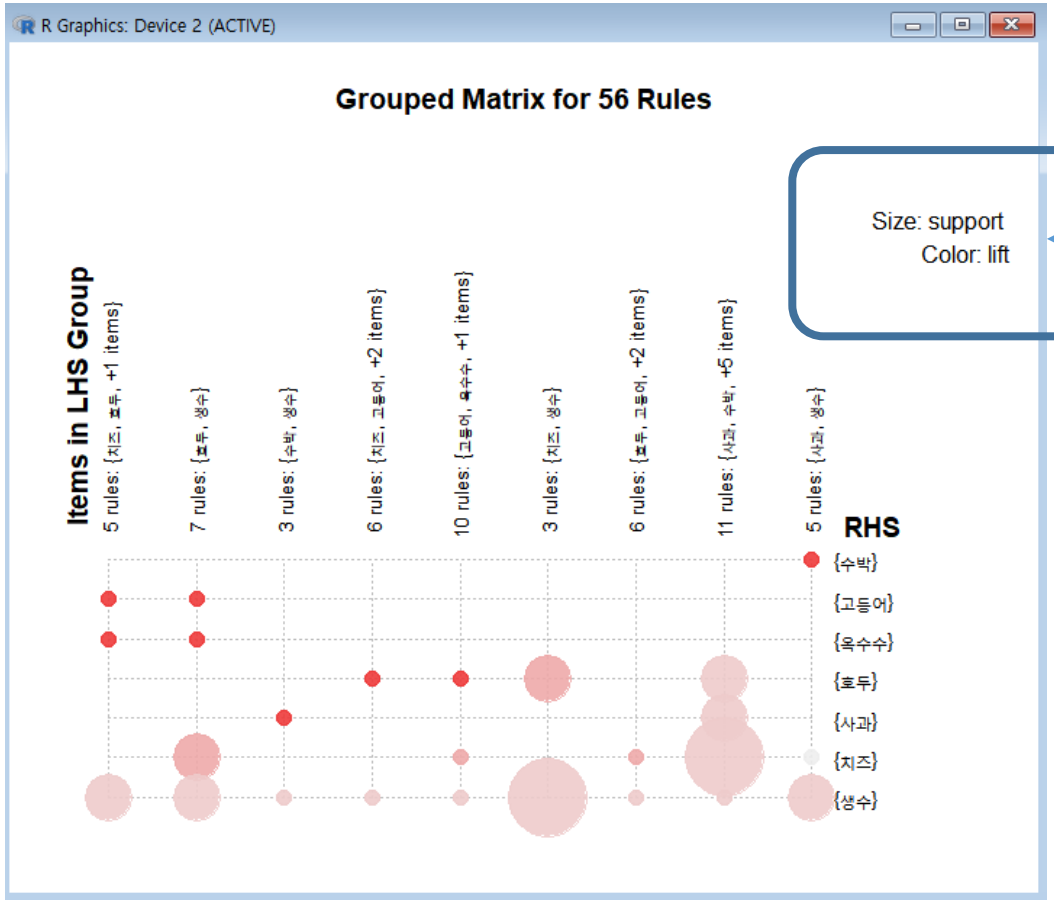
연관성분석-시각화

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

# 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

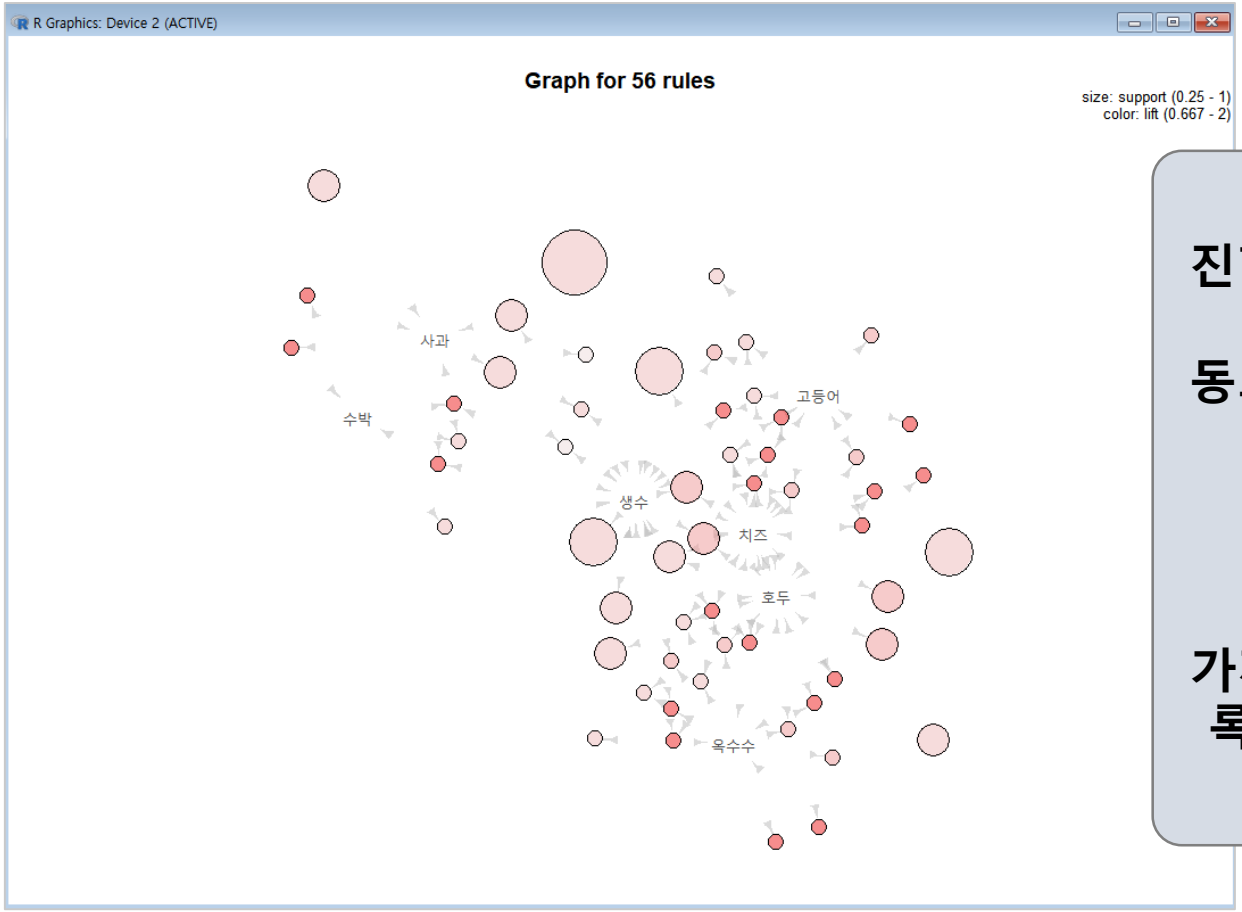
plot(rules,method="grouped")



진한 빨간색일수록 향상도가 높은 자료임.
동그라미가 클수록 지지도

(많이 나온 빈도수가 높은 자료임)

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



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

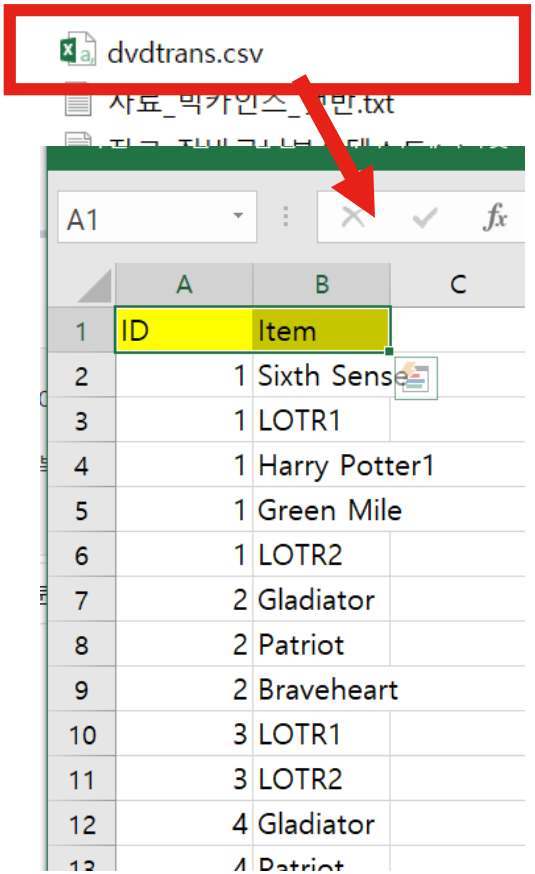
(많이 나온 빈도수가 높은 자료임)

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



좌측의 데이터 프레임을
우측의 구조로 변경하는 작업이 필요함.

	A	B
1	장바구니번호	구매내역
2	1	넥타이
3	1	셔츠
4	1	양말
5	2	양말
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

	A	B	C	D
1	넥타이	셔츠	양말	
2	양말	벨트	장갑	셔츠
3	지갑	넥타이	셔츠	
4	양말	벨트	장갑	바지
5				

연관성분석-데이터 프레임 형식일때

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

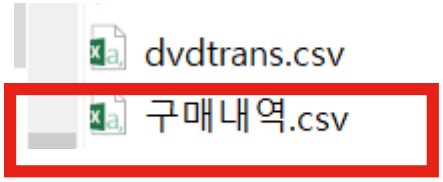
메모리 지우기
연관분석 패키지설치

```
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 연관되어묶여있는지 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
2	1	홍길동	새우깡
3	2	홍길동	맛동산
4	3	홍길동	맥주
5	4	일지매	짱구
6	5	일지매	감자강
7	6	강감찬	감자깡
8	7	강감찬	새우깡
9	8	전우치	자갈치
10	9	전우치	맛동산
11	10	홍길동	짱구
12	11	어우동	빠다코코낫
13	12	어우동	맛동산
14	13	강감찬	포카칩
15	14	강감찬	맥주
16	15	김유신	자갈치
17	16	김유신	짱구
18	17	김유신	맛동산
19	18	홍길동	맛동산
20	19	전우치	초코침쿠키
21	20	강감찬	크라운산도

구매항목을 고객명으로 나누면

홍길동 고객
새우깡, 맛동산,맥주, 맛동산

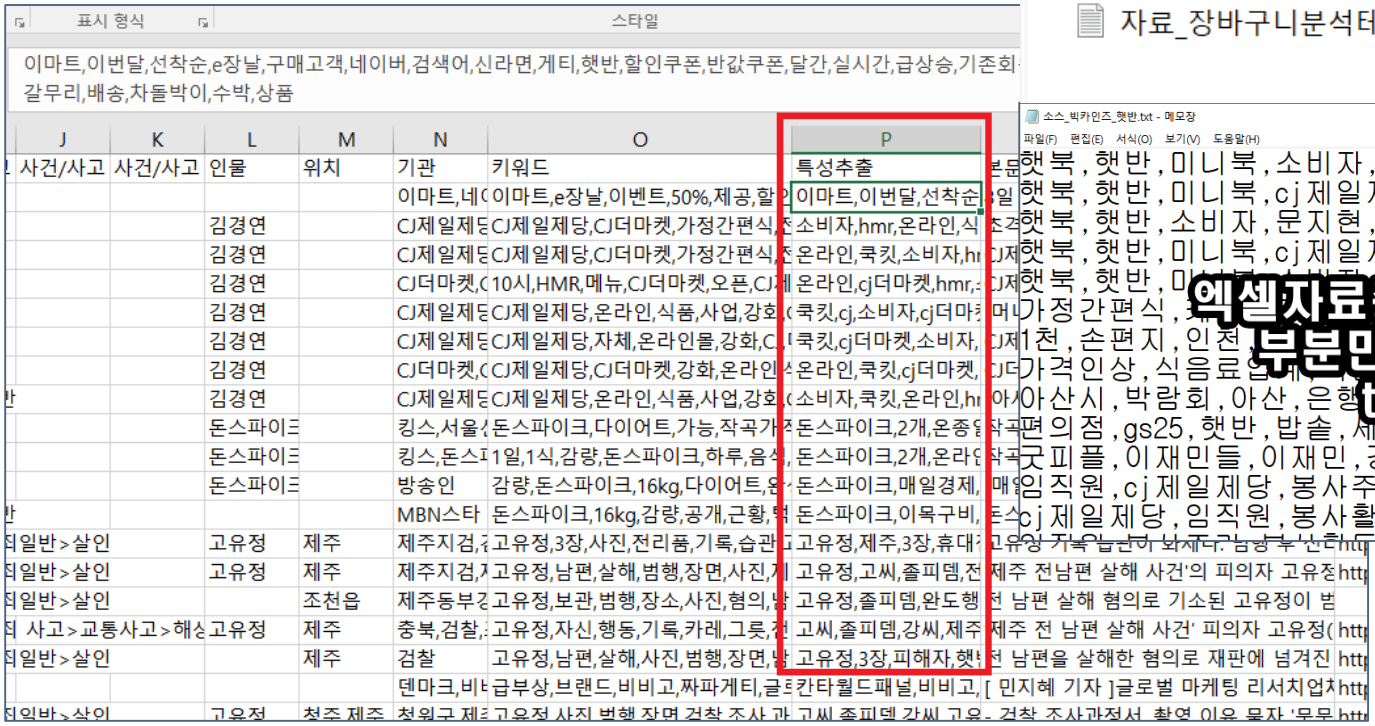
➔ 으로 맛동산 항목이 중복됨.
➔ 연관분석은 중복데이터가 있으면 에러

위의 사이트를 참조하여서 반드시 실습

연관분석을 이용한 텍스트 분석

자료확인: '자료_빅카인즈_햇반.txt'

빅카인즈(신문기사 분석 플랫폼)에서 단어가 다 분리된 자료를 이용하여서도 연관분석 가능함.
이자료는 '햇반' 키워드로 자료 다운로드한 자료의 P열 특성추출 은 Q열의 본문자료를 단어단위로 쉼표분리한 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



연관성분석-시각화 (텍스트를 이용한)

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