R Programming

Lesson 5

Insight campus

MARKET BASKET ANALYSIS

Association Rules

STEPS OF MARKET BASKET ANALYSIS

1. Data Acquisition
2. Obtain Association Rules
3. Visualization

GROCERIES DATA

- data(Groceries) {arules}
- K = 169 (items, columns), N = 9,835 (transactions, rows)
- Michael Hahsler, Kurt Hornik, and Thomas Reutterer (2006) Implications of probabilistic data modeling for mining association rules. In M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nuernberger, and W. Gaul, editors, From Data and Information Analysis to Knowledge Engineering, Studies in Classification, Data Analysis, and Knowledge Organization, pages 598--605. Springer-Verlag.

WHAT IS MARKET BASKET ANALYSIS

- 시장바구니 분석(유사성 분석 또는 연관성 분석)은 어떠한 품목이 같이 판매되는가와 동시에 주문이나 판매되는 품목은 무엇인가에 대한 답을 찾는다.
- 분석 결과에 근거하면 해당 상품을 같은 곳에 진열할 수 있고, 교차 판매와 공동마케팅 프로모션, 제품 번들링 등을 기획할 수 있다.
- 시장바구니 분석용 데이터를 준비하기 위해 구입수량을 이진표시기로 변환한다. 한 아이템을 구매하면 쇼핑 목록에서 해당 열이 1이 되며, 입력 행의 나머지 열은 0으로 설정한다.
- 시장 바구니 분석에 대한 입력 데이터는 희소이진행렬(sparse binary matrix)이며 행과 열이 1과 0으로 이루어져 있다.
- 아이템의 구매 수량 N개 대신 1로 표시하는 이유는 이진 행렬이 실제 구매량의 값을 갖는 행렬보다 분석이 용이하기 때문이다.

SUPPORT

- 지지도(Support): A가 포함된 거래수 / 전체 거래수 -> 상대적 빈도 파악

$$supp(A) = \frac{A\text{를 구매한 건수}}{\text{전체 구매 건 수}}$$

CONFIDENCE

- 신뢰도(Confidence): “A한 사람이 B하더라”라고 말할 수 있는 비율

$$conf(A \Rightarrow B) = \frac{A, B\text{를 동시에 구매한 건수}}{A\text{를 구매한 건수}} = P_T(B|A)$$

LIFT

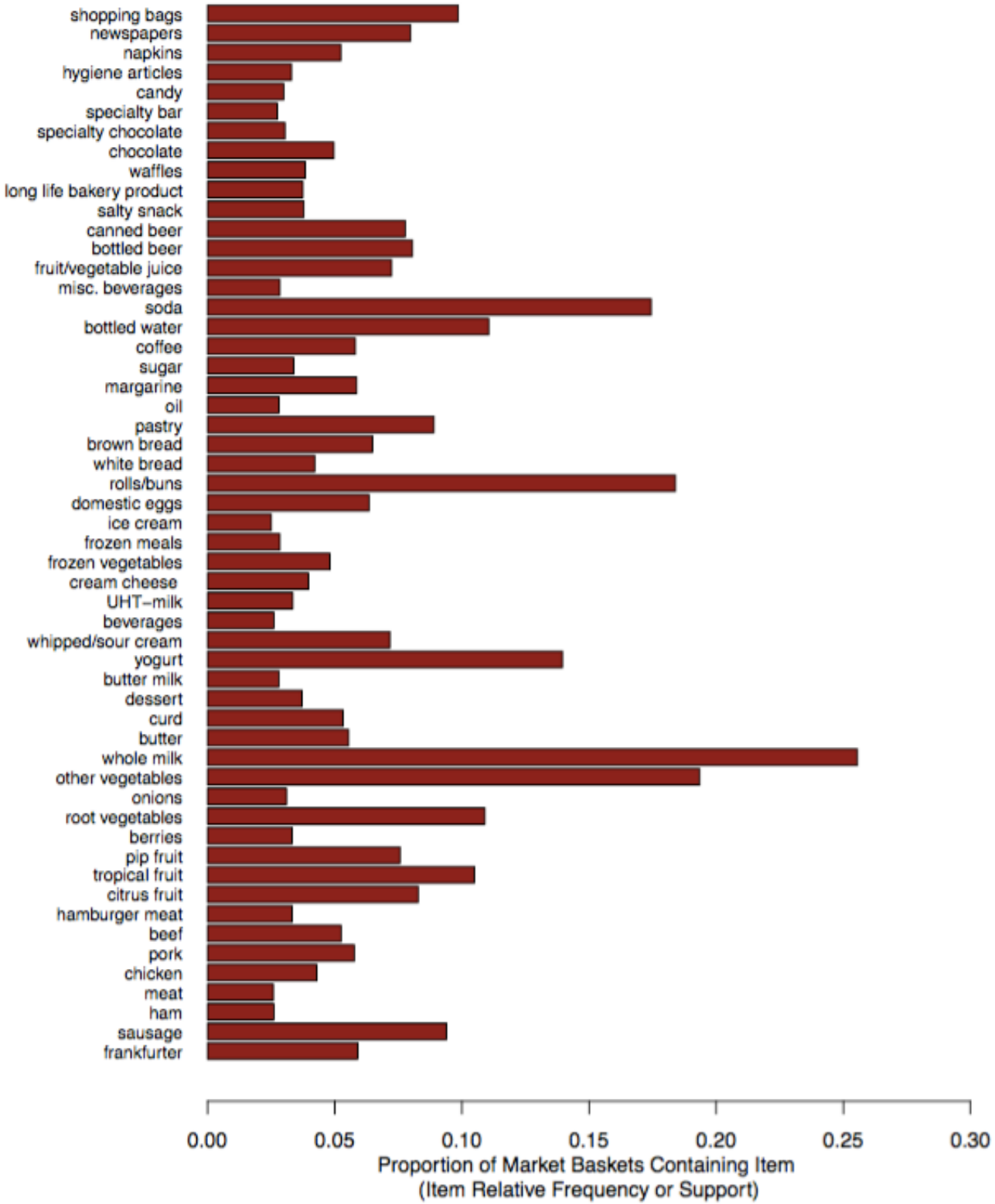
- 향상도(Lift): A가 주어지지 않았을 때의 품목 B의 확률에 비해 A가 주어졌을 때의 품목 B의 확률의 증가 비율

$$lift(A \Rightarrow B) = \frac{A, B\text{를 동시에 구매한 건수}}{A\text{를 구매한 건수}} \Big/ B\text{를 구매한 건수}$$

$$lift(A \Rightarrow B) = \frac{P_T(B|A)}{P_T(B)}$$

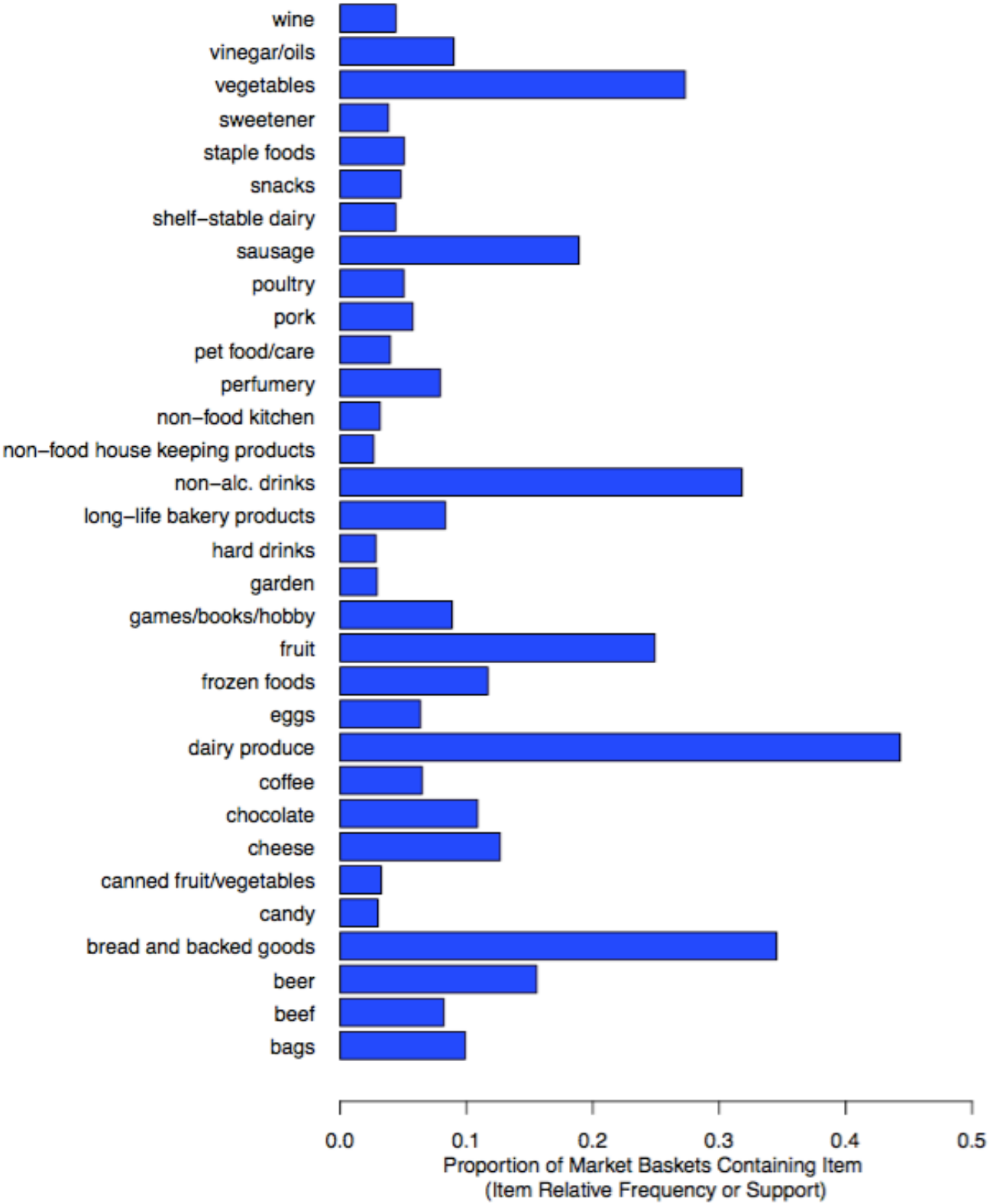
Insight campus

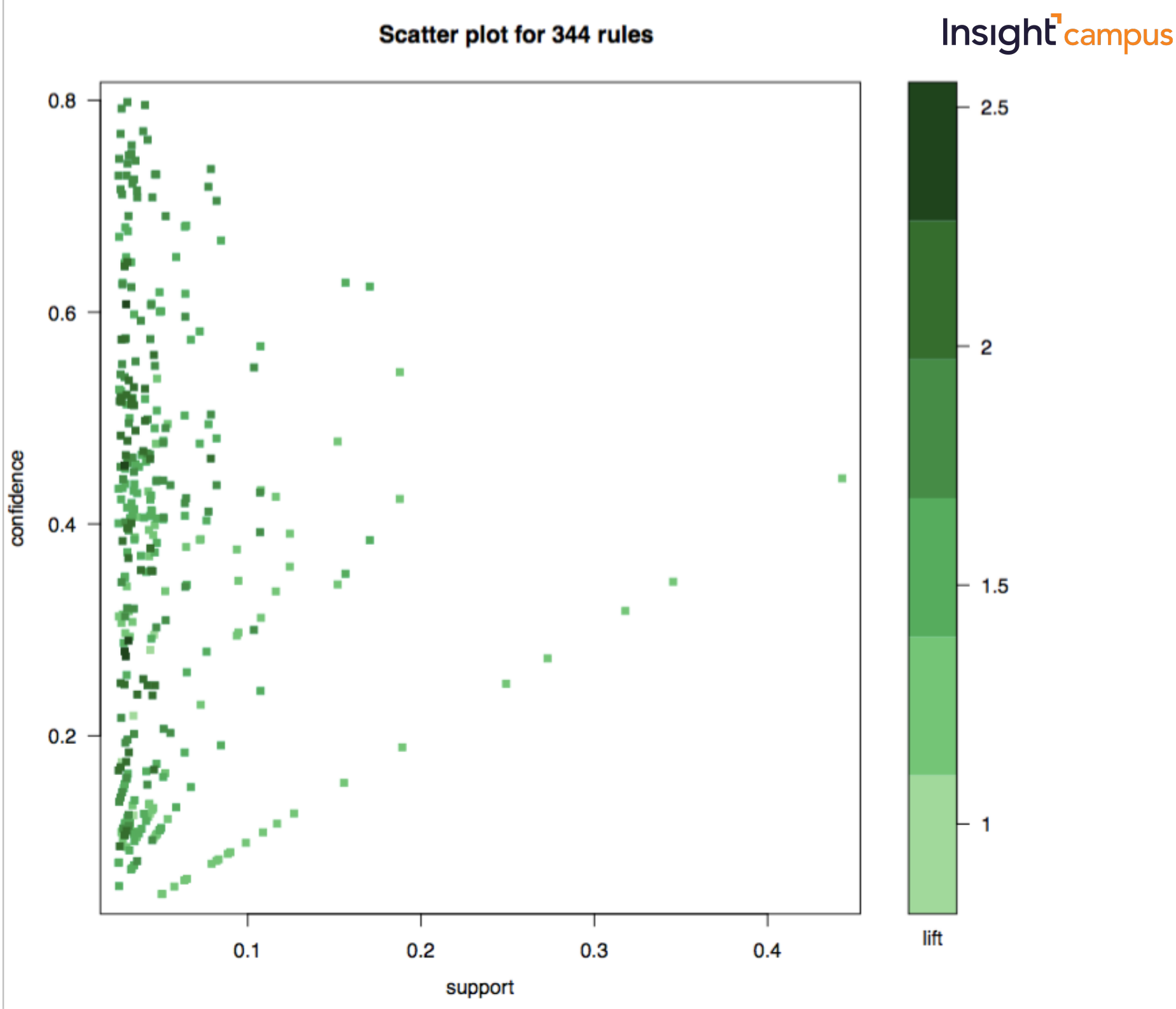
frequency of support > 0.025

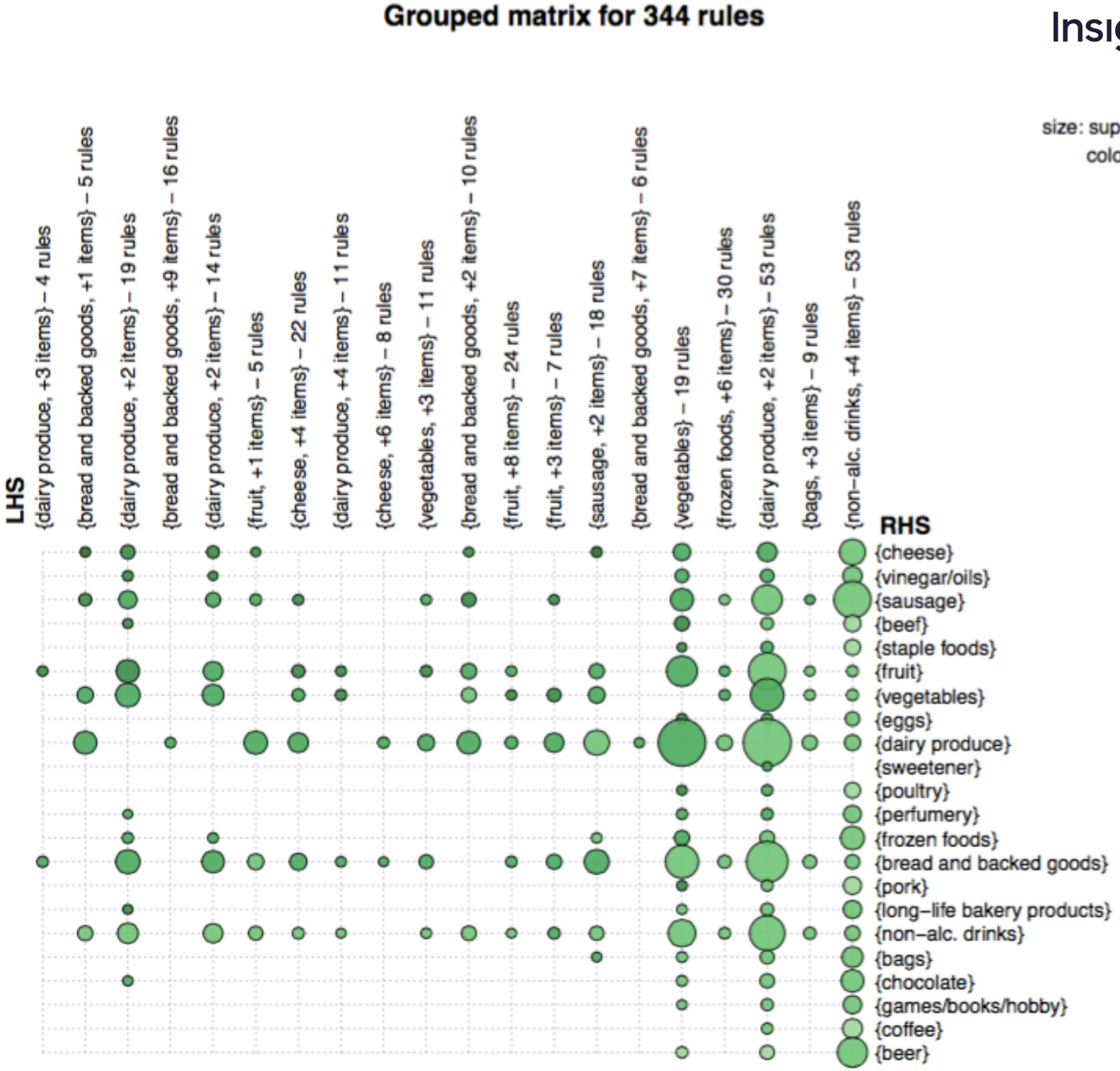


Insight campus

55 distinct levels

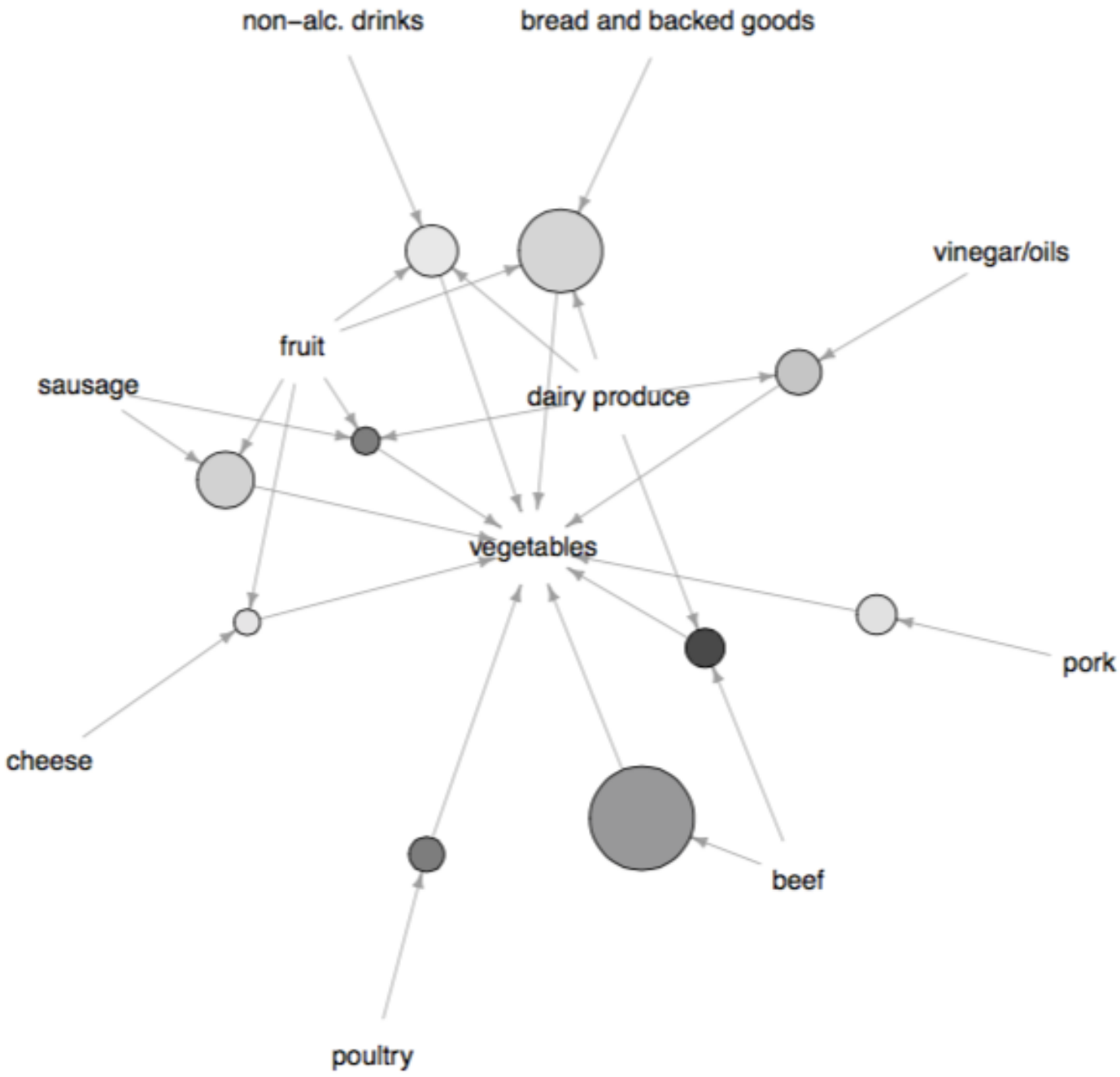



Scatter plot for 344 rules
Insight campus
confidence
0.8
0.6
0.4
0.2
0.1
0.2
0.3
0.4
support
2.5
2
1.5
1
lift


Grouped matrix for 344 rules
size: support
color: lift
LHS
{dairy produce, +3 items} – 4 rules
{bread and backed goods, +1 items} – 5 rules
{dairy produce, +2 items} – 19 rules
{bread and backed goods, +9 items} – 16 rules
{dairy produce, +2 items} – 14 rules
{fruit, +1 items} – 5 rules
{cheese, +4 items} – 22 rules
{dairy produce, +4 items} – 11 rules
{cheese, +6 items} – 8 rules
{vegetables, +3 items} – 11 rules
{bread and backed goods, +2 items} – 10 rules
{fruit, +8 items} – 24 rules
{fruit, +3 items} – 7 rules
{sausage, +2 items} – 18 rules
{bread and backed goods, +7 items} – 6 rules
{vegetables} – 19 rules
{frozen foods, +6 items} – 30 rules
{dairy produce, +2 items} – 53 rules
{bags, +3 items} – 9 rules
{non–alc. drinks, +4 items} – 53 rules
RHS
{cheese}
{vinegar/oils}
{sausage}
{beef}
{staple foods}
{fruit}
{vegetables}
{eggs}
{dairy produce}
{sweetener}
{poultry}
{perfumery}
{frozen foods}
{bread and backed goods}
{pork}
{long–life bakery products}
{non–alc. drinks}
{bags}
{chocolate}
{games/books/hobby}
{coffee}
{beer}

Graph for 10 rules

size: support (0.027 – 0.046)
color: lift (1.899 – 2.225)



WORD CLOUD

Text Mining

GLOSSARY

텍스트 마이닝(text mining)	텍스트 데이터베이스에서 좀 더 효율적인 방법으로 유용한 정보를 탐색하는 방법
단어줄기(stem word)	동사 등에서 변하는 어미를 제거한 단어
불용어(stop word)	문서 중에 나타나는 빈도는 높으나 의미가 없는 단어를 말하며, 예를 들면 관사, 또는 '것' 등
코퍼스(corpus)	문서들의 집합
워드 클라우드(word cloud)	텍스트 데이터베이스의 한 문서에서 단어의 출현빈도를 사용하여 많이 나타난 단어일수록 더 큰 글자로 화면 가운데 배치하여 단어 구름을 만들어 시각화하는 그림

TEXT MINING PROCESS

1	텍스트 데이터베이스에 크기가 서로 다른 문서들이 있을 때 이 문서들의 집합인 코퍼스를 생성한다.
2	코퍼스 안에 있는 모든 단어의 정제작업을 한다. 단어줄기 추출**(word stemming)**. 명사의 경우에는 복수를 나타내는 어미를 제거하고, 동사의 경우에는 변하는 어미를 제거하고 어간만 추출한다.
3	언어의 특성에 따른 추가 정제작업을 한다. 예를 들어 영어는 모든 문자를 소문자로 바꿀 수 있다.
4	전치사, 관사, 접속사 등과 관심의 대상이 되지 않는 불용어를 제거한다.
5	정제된 코퍼스 안에 있는 모든 단어의 출현빈도를 조사하여 중요 단어를 추출한다.
6	중요 단어를 큰 글자나 색, 회전 등으로 구별하여 화면 가운데에 배치하는 워드 클라우드를 만든다.

DOWNLOAD RJAVA

- rJava가 설치되어 있지 않으면 오류가 발생한다.
- Java 설치하기
- http://www.java.com/en/download/manual.jsp

INSTALL PACKAGES

- # install.packages("rJava")
- # install.packages("tm")
- # install.packages("wordcloud")
- # install.packages("KoNLP")
- # install.packages("SnowballC")
- library(tm) # Text Mining package
- library(wordcloud) # Word Clouds
- library(KoNLP) # Korean NLP package (한국어 형태소 분석)
- library(SnowballC) # stemmers
- library(RColorBrewer) # ColorBrewer Palettes

ENGLISH TEXT MINING

코퍼스 생성

- obama <- Corpus(DirSource(“ ”))
- names(obama)
- summary(obama)
- str(obama)

- # 코퍼스 내용 확인
- inspect(obama)
- writeLines(as.character(obama[[1]]))

전처리 # PRE-PROCESSING

- getTransformations()
- ?content_transformer

- delete <- content_transformer(function(x, pattern) {gsub(pattern, " ", x)})
- obama <- tm_map(obama, delete, “,")

- # 공백 제거
- obama <- tm_map(obama, stripWhitespace)

- # 구두점 제거
- obama <- tm_map(obama, removePunctuation)

전처리 # PRE-PROCESSING

- # 소문자화, 대문자를 구별할 필요가 있을 때는 생략
- obama <- tm_map(obama, content_transformer(tolower))

- # 숫자 제거
- obama <- tm_map(obama, removeNumbers)

- # 불용어 제거
- obama <- tm_map(obama, removeWords, stopwords("english"))

- # 단어줄기 추출
- obama <- tm_map(obama, stemDocument)

전처리 # PRE-PROCESSING

- # R 버전 3.1.1 이후에서는 'inherits(doc, "TextDocument")는 TRUE가 아닙니다' 라는 에러가 나타날 수 있다.
- # 이때는 다음 명령어를 사용해야 한다.
- obama <- tm_map(obama, PlainTextDocument)
- inspect(obama)

워드 클라우드 생성

- wordcloud(obama, scale = c(4,0.5), max.words = 100,
  random.order = FALSE,
  rot.per = 0.35, colors = brewer.pal(8, “Set2"))

- rot.per: 수직방향으로 회전되어 배열되는 단어의 비율
- colors: brewer.pal 함수를 사용해 RColorBrewer의 팔레트를 사용할 수 있음
- random.color = T : 스크립트를 실행할 때마다 단어의 색이 변함
- random.order = F : 빈도가 큰 단어를 중앙에 두도록 함
- min.freq: 해당 빈도 이상의 단어만 나타도록 함
- scale: 글자 크기 c(MAX, MIN)

불용어 제거 및 수작업

- # gsub 함수를 사용한 단어 편집
- obama <- tm_map(obama, content_transformer(gsub), pattern = "countri", replacement = "country")

- # 사용자가 불용어 제거
- obama <- tm_map(obama, removeWords, "whether")

- # 여러 개의 불용어 제거
- eliminate <- c("whether", "less", "may", "shall", "across")
- obama <- tm_map(obama, removeWords, eliminate)