LDA & Market basket 기법 활용

고객 맞춤형 콘텐츠 제안

곽지훈 유한정

토픽 모델을 이용한 고객 Segmentation

기존의 거리기반 군집(k-means, hierarchical cluster...)은 데이터가 거대해 짐에 따라, 군집의 수가 증가함에 따라 성능이 현저히 떨어지는 단점을 지니고 있고 이를 보완하기 위한 토픽 모델 선택

LDA (Latent Dirichlet allocation)

* 주어진 문서 내 각 문서의 단어 출현빈도를 이용하여 어떤 주제들이 존재하는지 예측하는 확률적 토픽 모델
* 주로 텍스트 마이닝 기법으로 이용
* 본 분석은 문서를 각 고객으로 출현단어를 구매 상품으로 바꾸어 토픽 추출 후 고객 세분화
* R 패키지 lda 이용

김용대, 정구환 토픽모형을 이용한 빅데이터 기반 마이크로 세그멘테이션 방법론 연구(2016)

장바구니 분석(연관 분석)을 이용한 상품 Recommendation

* 이전 단계에서 세분화한 고객 집단의 구매 내역을 이용하여 상품 추천 로직 구현

Apriori Algorithm

* 연관분석의 대표적인 모델
* 지지도(support), 신뢰도(confidence), 리프트 (lift)
* 위 지표를 활용하여 빈발 항목 집합의 규칙 생성
* R 패키지 arules 이용

모델 생성 과정

STEP.1 세분화

업종구분

무의미한 연관규칙 피하기 위해

LDA모델 생성

아이템 중분류이용 토픽 정의

토픽부여

각 고객의 토픽 별 가중치(theta) 활용

고객 세분화

동일한 토픽이 선택 된 고객집단 구분

STEP.2 맞춤형 추천

장바구니 규칙 생성

세분화 된 고객 집단의 장바구니 이용

특정 고객 아이템 구매 빈도 추출

아이템 소분류 이용 장바구니 생성

상위 빈발 품목 기준 규칙 생성

특정 고객 추천을 위 한 새로운 규칙 생성

상품추천

LDA 모델 생성 과정(A02,대형마트) (*전처리 과정 소스 코드 내 삽입)

용어 설명

theta(θ): 해당 고객에 대한 각 Topic대한 가중치

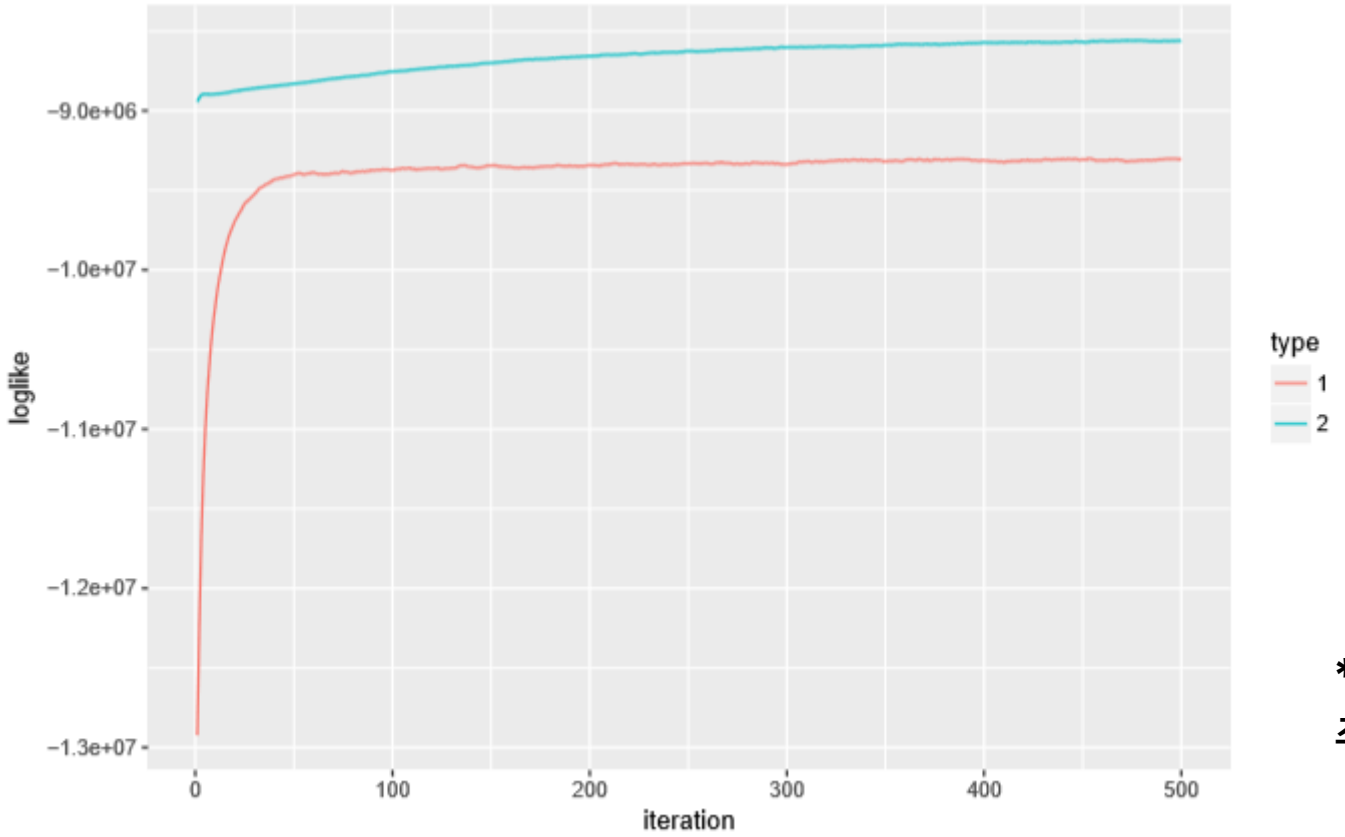
phi(ϕ): 각 Topic별 특정 아이템 구매 확률(?)

lift: Topic에서 각 상품의 확률(phi)을 쇼핑 쇼핑자료 전체에서 해당 상품이 차지하는 비율로 나눈 값

(토픽은 품목에 대한 확률 분포를 나타내기 때문에 lift값이 높은 값으로 Topic 이름 설정)

LDA 모델 생성 과정(A02,대형마트)

Parameter :(alpha = 1/k, beta = 1/w, iteration =500, frequency value = ID : PD_M_NM**)**
(K=토픽 수(10), W=총 아이템 수, PD_M_NM=상품 중분류,ID=고객ID**)** *** 고객 세분화는 아이템 중분류 이용**



*** 이후 모델결과로 사용할 theta(θ),phi(ϕ)**
추정을 위한 로그우도 수렴 여부 판단

LDA 모델 생성 과정(A02,대형마트)

theta matrix(A02 구매 고객 수: 13538, 토픽 수: 10)

	1.냉장햄세트-열대과일선물세트	2.전기요/장판-온라인저장과일	3.온라인돼지고기-온라인양곡	4.마른안주선물세트-롤러보드	5.해물선물세트-테이블	6.스포츠브랜드편집-원재료	7.온라인양념/루리체소-온라인잎/쎌러드체소	8.주방가전-기타구색생선	9.주방소형가전-기타	10.이동통신-주유소
1	0.0	0.00000000	0.0000000	0.00000000	0.0000000	0.0000000	0.000000000	0	0	1
2	0.0	0.01992032	0.0000000	0.00000000	0.9760956	0.0000000	0.003984064	0	0	0
3	0.0	0.10810811	0.0000000	0.05405405	0.8378378	0.0000000	0.000000000	0	0	0
4	0.0	0.04867257	0.7699115	0.00000000	0.0000000	0.1814159	0.000000000	0	0	0
6	0.0	0.00000000	0.0000000	0.03377386	0.9647577	0.0000000	0.001468429	0	0	0
7	0.7	0.00000000	0.0000000	0.00000000	0.0000000	0.0000000	0.300000000	0	0	0

(행: ID, 열: 토픽) (13,538 x 10) *lift 기준 상위 두가지 품목으로 토픽 이름 설정

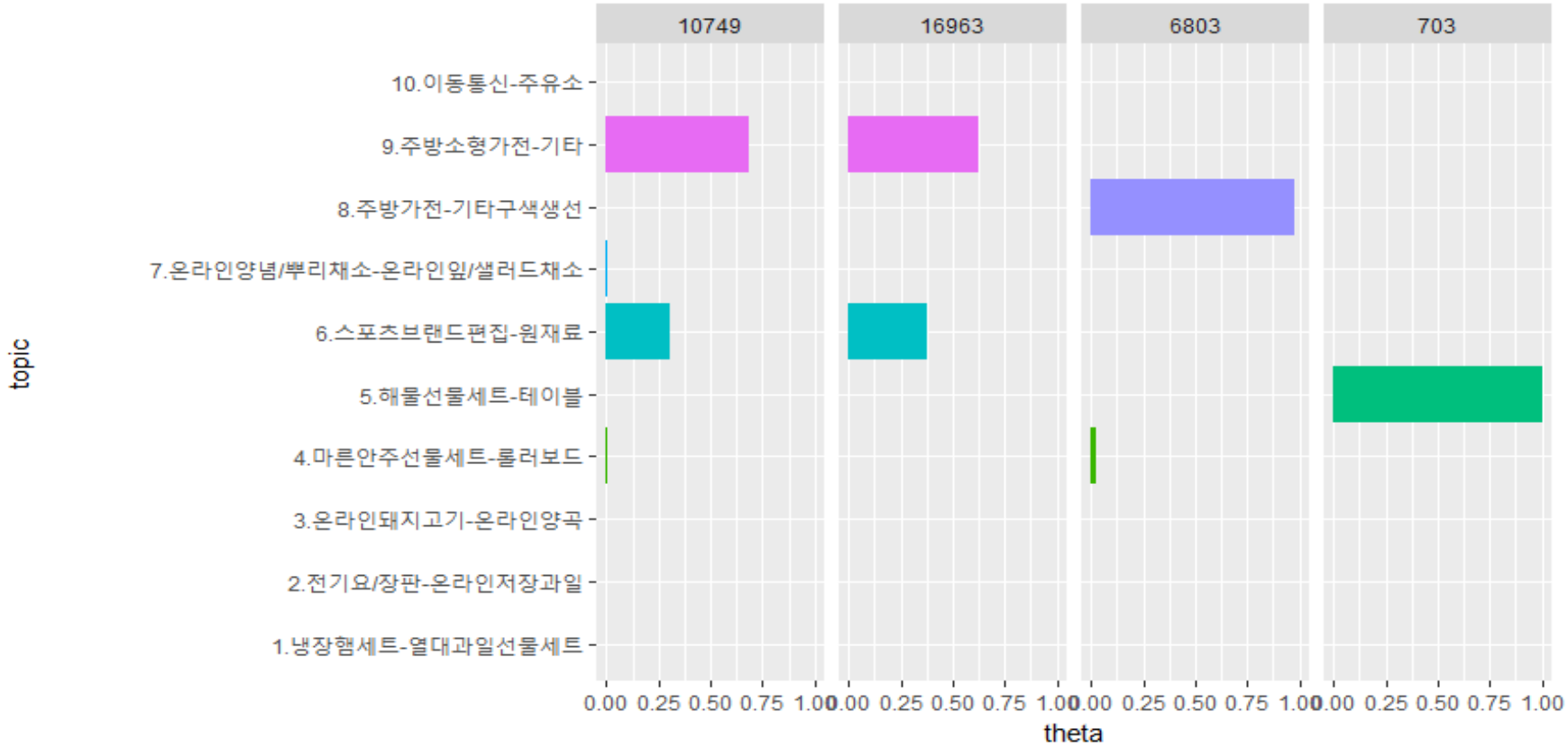
LDA 모델 생성 과정(A02,대형마트)

Topic matrix(theta matrix 변환)

	1.냉장햄세트-열대과일세트	2.전기요/장판-온라인저장과일	3.온라인돼지고기-온라인양곡	4.마른안주선물세트-롤러보드	5.해물선물세트-테이블	6.스포츠브랜드편집숍-원제	7.온라인양념/부리체소-온라인잎/셀러드체소	8.주방가전-기타구색생선	9.주방소형가전-기타	10.이동통신-주유소
1	0	0	0	0	0	0	0	0	0	1
2	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	1	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0
7	1	0	0	0	0	0	1	0	0	0
8	0	0	0	1	0	0	0	0	1	0

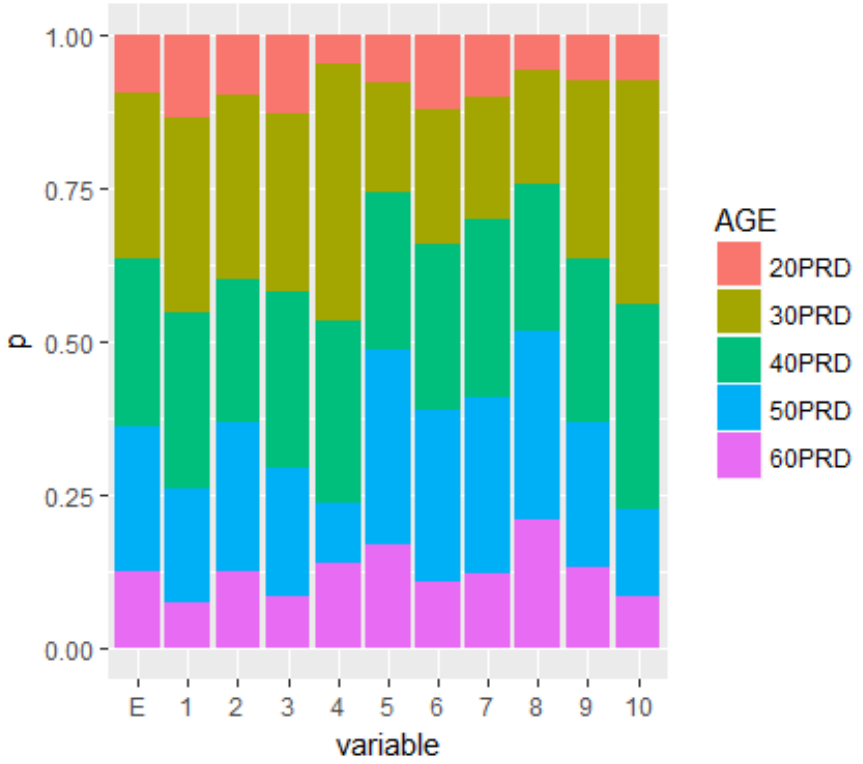
(theta>=0.2이상 토픽 선택) *각 ID에 선택된 TOPIC 1 else 0

각 고객의 토픽 별 가중치

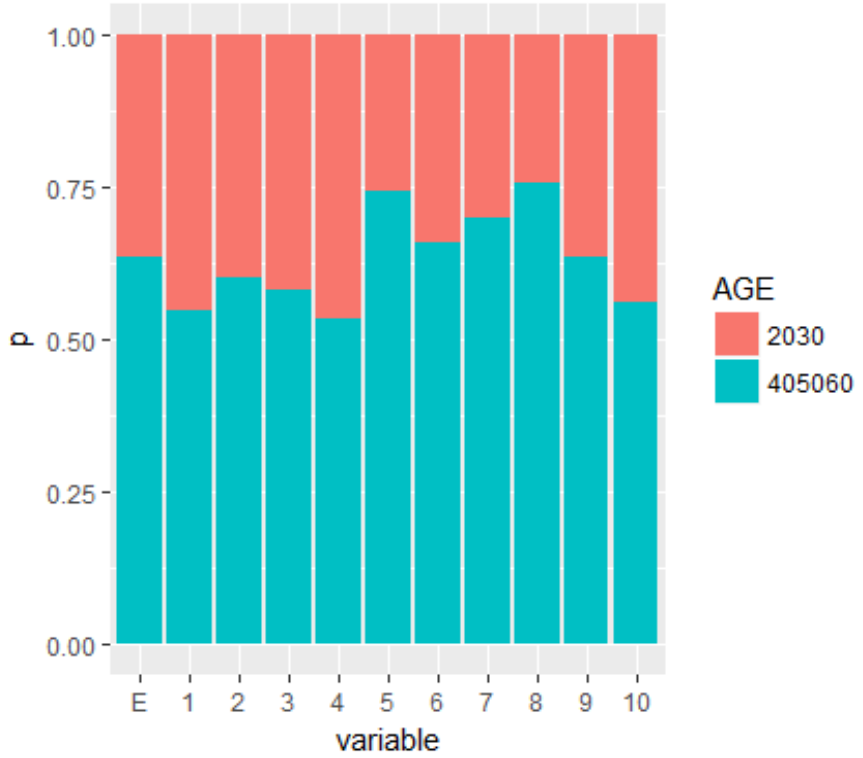


* 6, 9번의 토픽이 선택된 10749고객과 16963 고객은 유사한 itemset의 구매 행위를 했을 것으로 예측 가능
* 위 고객 집단(6,9번 토픽)을 중심으로 분석 진행

각 토픽 별 성비



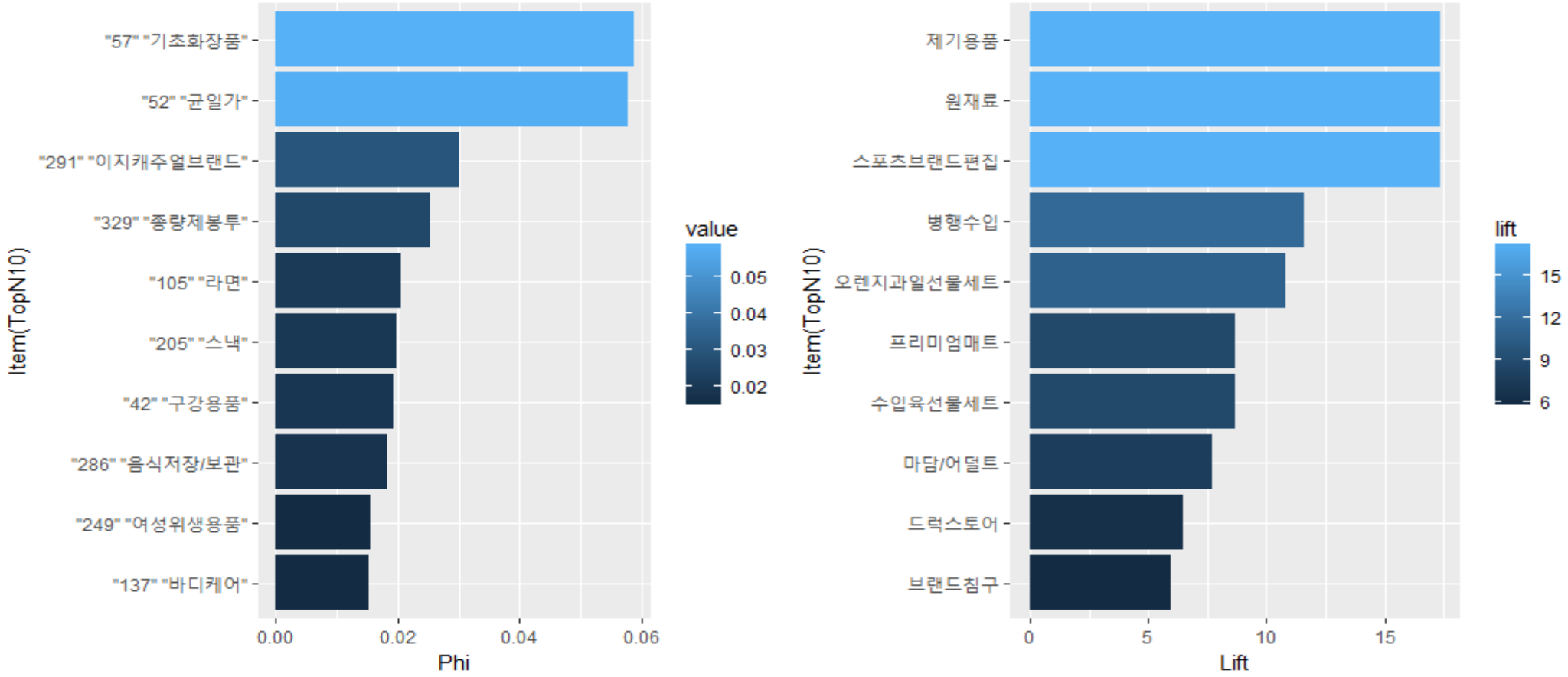
E: 해당 업종 전체 비율



6,9번 토픽은 상대적으로 젊은층 보다는 중장년층의 비중이 다소 높음

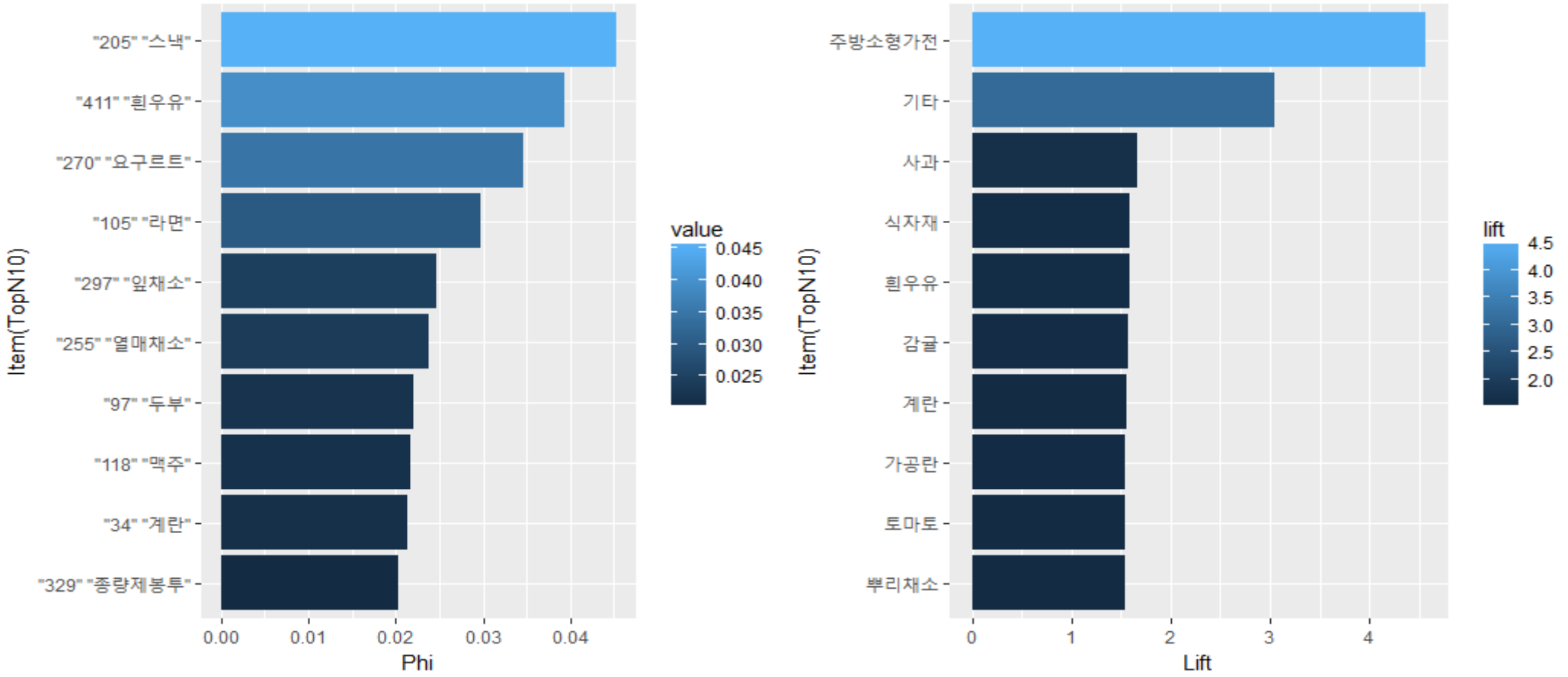
토픽 내 item 가중치(phi) & lift (*상위 10개 item)

Topic6

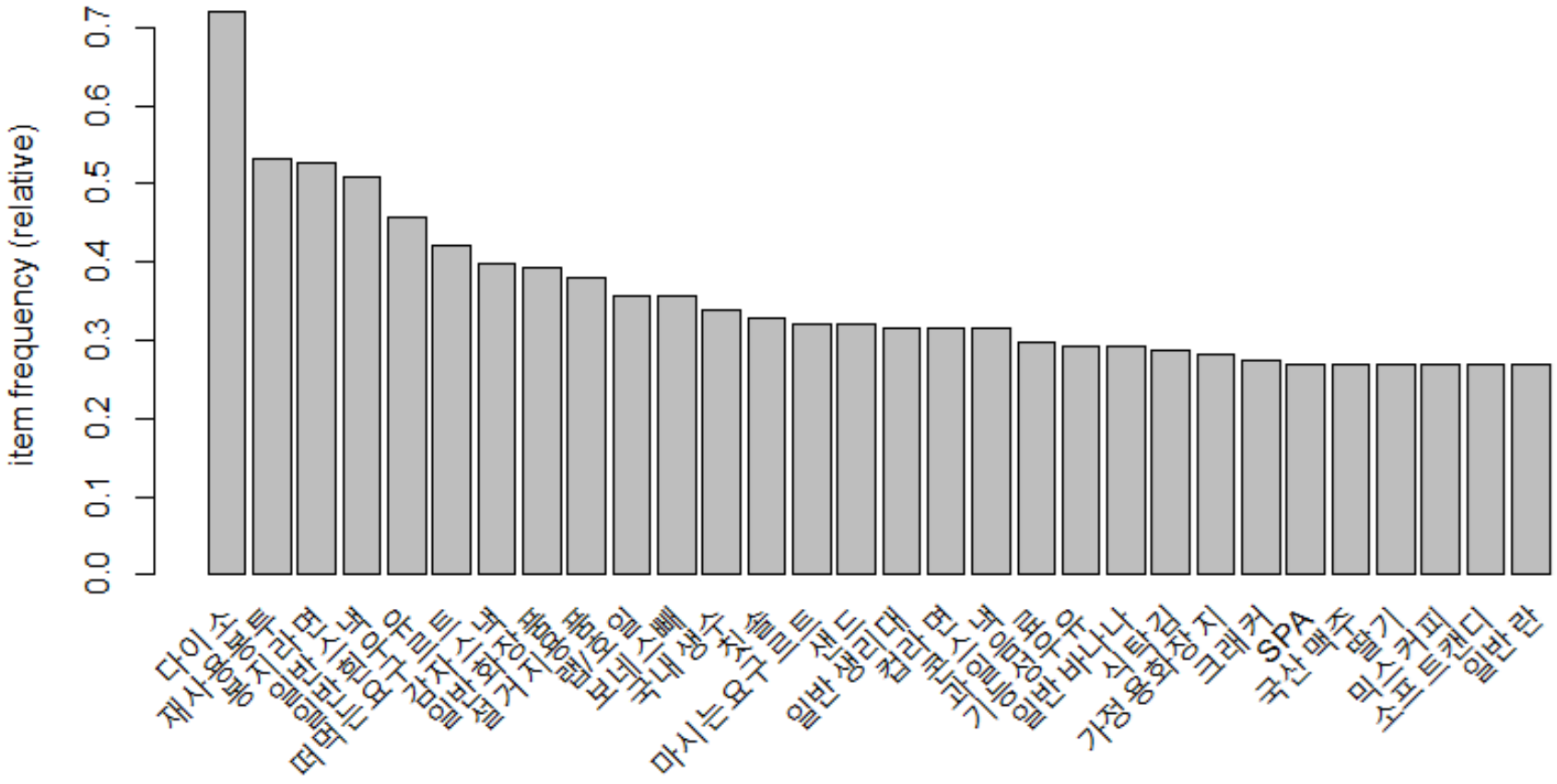


토픽 내 item 가중치(phi) & lift (*상위 10개 item)

Topic9



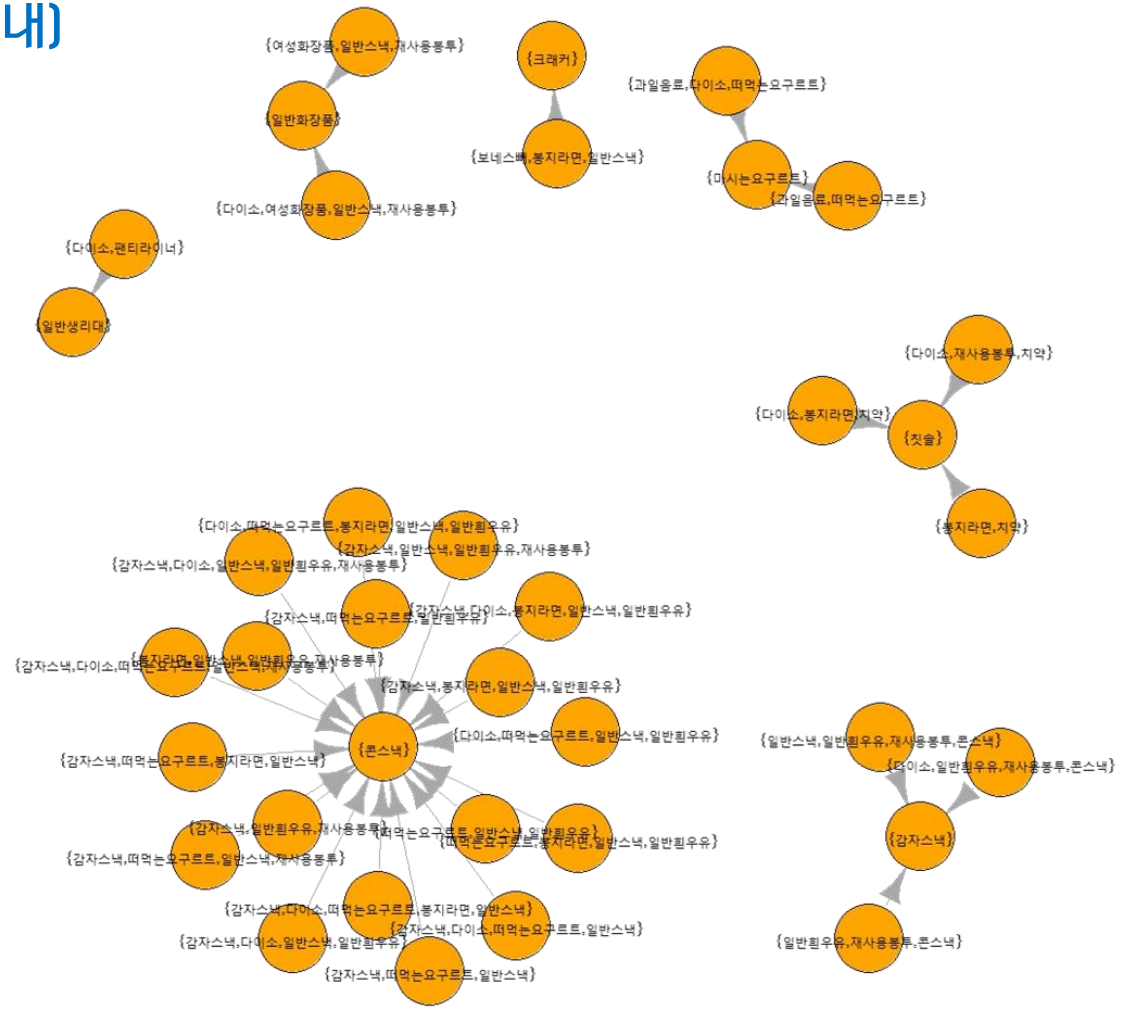
장바구니 분석 (6,9번 토픽이 동시에 선택된 고객 집단 내)



*단일 구매 상품 빈도 Top30

장바구니 분석 (6,9번 토픽이 동시에 선택된 고객 집단 내)

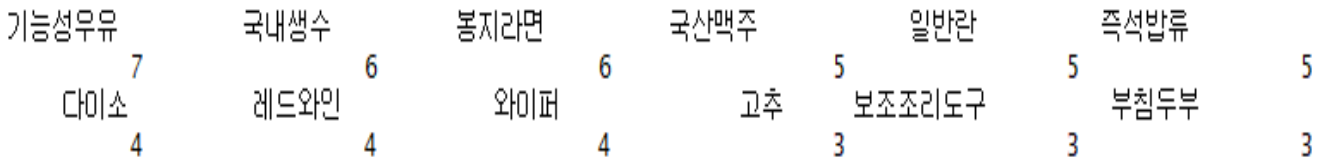
	lhs		rhs	support	confidence	lift
1302	{보네스빼,봉지라면,일반스낵}	=>	{크래커}	0.1520468	0.8125000	2.956117
2279	{감자스낵,떠먹는요구르트,봉지라면,일반스낵}	=>	{콘스낵}	0.1637427	0.8750000	2.770833
2261	{감자스낵,일반스낵,일반흰우유,재사용봉투}	=>	{콘스낵}	0.1578947	0.8709677	2.758065
2333	{떠먹는요구르트,봉지라면,일반스낵,일반흰우유}	=>	{콘스낵}	0.1578947	0.8709677	2.758065
2532	{감자스낵,다이소,봉지라면,일반스낵,일반흰우유}	=>	{콘스낵}	0.1520468	0.8666667	2.744444
2338	{다이소,떠먹는요구르트,일반스낵,일반흰우유}	=>	{콘스낵}	0.1754386	0.8571429	2.714286
2256	{감자스낵,봉지라면,일반스낵,일반흰우유}	=>	{콘스낵}	0.1637427	0.8484848	2.686869
2284	{감자스낵,떠먹는요구르트,일반스낵,재사용봉투}	=>	{콘스낵}	0.1578947	0.8437500	2.671875
1827	{떠먹는요구르트,일반스낵,일반흰우유}	=>	{콘스낵}	0.1812865	0.8378378	2.653153
2289	{감자스낵,다이소,떠먹는요구르트,일반스낵}	=>	{콘스낵}	0.1695906	0.8285714	2.623810
1188	{다이소,봉지라면,치약}	=>	{칫솔}	0.1578947	0.8437500	2.576451
148	{다이소,팬티라이너}	=>	{일반생리대}	0.1520468	0.8125000	2.572917
1768	{감자스낵,떠먹는요구르트,일반흰우유}	=>	{콘스낵}	0.1520468	0.8125000	2.572917
2347	{봉지라면,일반스낵,일반흰우유,재사용봉투}	=>	{콘스낵}	0.1520468	0.8125000	2.572917
1783	{감자스낵,떠먹는요구르트,일반스낵}	=>	{콘스낵}	0.1754386	0.8108108	2.567568
1485	{과일음료,다이소,떠먹는요구르트}	=>	{마시는요구르트}	0.1637427	0.8235294	2.560428
2266	{감자스낵,다이소,일반스낵,일반흰우유}	=>	{콘스낵}	0.1695906	0.8055556	2.550926
1777	{감자스낵,일반흰우유,재사용봉투}	=>	{콘스낵}	0.1637427	0.8000000	2.533333
483	{봉지라면,치약}	=>	{칫솔}	0.1637427	0.8235294	2.514706
1191	{다이소,재사용봉투,치약}	=>	{칫솔}	0.1578947	0.8181818	2.498377
751	{과일음료,떠먹는요구르트}	=>	{마시는요구르트}	0.1637427	0.8000000	2.487273
1087	{여성화장품,일반스낵,재사용봉투}	=>	{일반화장품}	0.1520468	0.9285714	2.369936



(최소지지도=0.2,최소신뢰도=0.8, 중복규칙 제거,lift기준 상위 30개 set 관계도)

장바구니 분석 (선택된 고객 집단 내 10749고객의 추천상품)

해당 고객 구매 아이템 빈도



상위 빈도 품목 4가지(기능성우유, 국내생수, 봉지라면, 국산맥주)를 포함한 거래만 추출

	lhs		rhs	support	confidence	lift
31	{국내생수,국산맥주,기능성우유,봉지라면}	=>	{과일음료}	0.05847953	0.9090909	3.048128
23	{국내생수,국산맥주,봉지라면}	=>	{과일음료}	0.08187135	0.8750000	2.933824
16	{국내생수,국산맥주,기능성우유}	=>	{과일음료}	0.05847953	0.8333333	2.794118
15	{국내생수,국산맥주,기능성우유}	=>	{컵라면}	0.05847953	0.8333333	2.638889
29	{국내생수,국산맥주,기능성우유,봉지라면}	=>	{컵라면}	0.05263158	0.8181818	2.590909
21	{국내생수,국산맥주,봉지라면}	=>	{컵라면}	0.07602339	0.8125000	2.572917
30	{국내생수,국산맥주,기능성우유,봉지라면}	=>	{마시는요구르트}	0.05263158	0.8181818	2.543802
22	{국내생수,국산맥주,봉지라면}	=>	{마시는요구르트}	0.07602339	0.8125000	2.526136
24	{국내생수,국산맥주,봉지라면}	=>	{떠먹는요구르트}	0.07602339	0.8125000	1.929688
26	{국내생수,국산맥주,봉지라면}	=>	{재사용봉투}	0.08771930	0.9375000	1.761676
17	{국내생수,국산맥주,기능성우유}	=>	{재사용봉투}	0.06432749	0.9166667	1.722527
32	{국내생수,국산맥주,기능성우유,봉지라면}	=>	{재사용봉투}	0.05847953	0.9090909	1.708292
7	{국내생수,국산맥주}	=>	{재사용봉투}	0.10526316	0.9000000	1.691209
9	{국산맥주,봉지라면}	=>	{재사용봉투}	0.18128655	0.8857143	1.664364
19	{국산맥주,기능성우유,봉지라면}	=>	{일반스낵}	0.08771930	0.8333333	1.637931
12	{기능성우유,봉지라면}	=>	{일반스낵}	0.16959064	0.8285714	1.628571
25	{국내생수,국산맥주,봉지라면}	=>	{일반스낵}	0.07602339	0.8125000	1.596983
6	{국내생수,국산맥주}	=>	{일반스낵}	0.09356725	0.8000000	1.572414
5	{국산맥주,기능성우유}	=>	{다이소}	0.12280702	0.9545455	1.327051
28	{국내생수,기능성우유,봉지라면}	=>	{다이소}	0.10526316	0.9473684	1.317073
20	{국산맥주,기능성우유,봉지라면}	=>	{다이소}	0.09941520	0.9444444	1.313008
13	{기능성우유,봉지라면}	=>	{다이소}	0.19298246	0.9428571	1.310801
27	{국내생수,국산맥주,봉지라면}	=>	{다이소}	0.08771930	0.9375000	1.303354
11	{국내생수,기능성우유}	=>	{다이소}	0.14035088	0.9230769	1.283302
2	{기능성우유}	=>	{다이소}	0.26900585	0.9200000	1.279024
18	{국내생수,국산맥주,기능성우유}	=>	{다이소}	0.06432749	0.9166667	1.274390

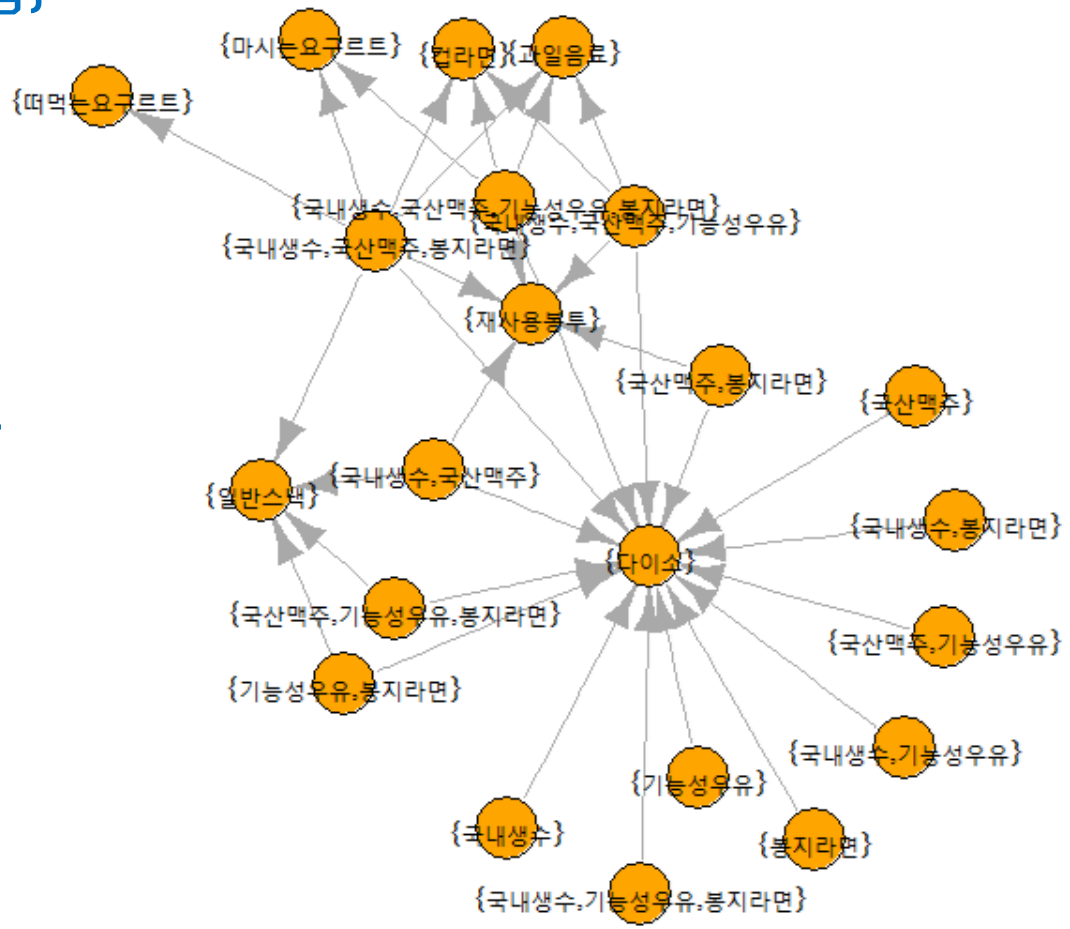
(최소지지도=0.05,최소신뢰도=0.8, 중복규칙 제거,lift기준 상위 30개 set 관계도)

장바구니 분석 (선택된 고객 집단 내 10749고객의 추천상품)

lift기준 상위 30개 rule에서 아이템 추출

과일음료, 컵라면, 마시는 요구르트, 떠먹는 요구르트
일반스낵, 다이소

10749 고객에게 위 4가지 상품을 추천한다면
기존 구매 빈도가 높은 상품을 구매 하면서 함께
구매할 확률이 매우 높아질 것이다.



상위 빈도 품목 4가지(기능성우유, 국내생수, 봉지라면, 국산맥주)를 포함한 거래만 추출

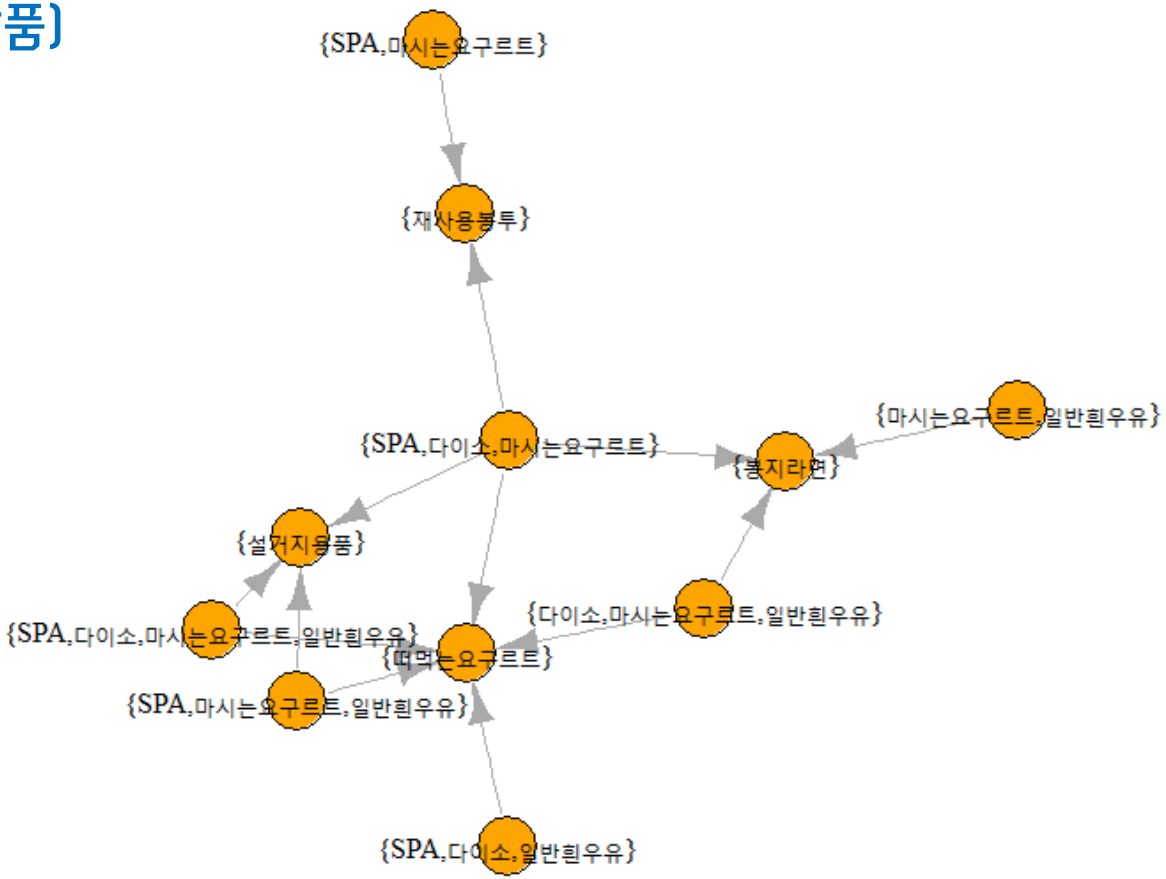
(최소지지도=0.05,최소신뢰도=0.8, 중복규칙 제거,lift기준 상위 30개 set 관계도)

장바구니 분석 (선택된 고객 집단 내 16963고객의 추천상품)

lift기준 상위 30개 rule에서 아이템 추출

설거지용품, 떠먹는 요구르트, 봉지라면

116963 고객에게 위 4가지 상품을 추천한다면
기존 구매 빈도가 높은 상품을 구매 하면서 함께
구매할 확률이 매우 높아질 것이다.



상위 빈도 품목 4가지(일반 흰우유, 마시는 요구르트, 다이소, SPA)를 포함한 거래만 추출
(최소지지도=0.05,최소신뢰도=0.8, 중복규칙 제거,lift기준 상위 30개 set 관계도)