NAVER DataLab.

유행의 발견

우리는 어떻게 마라의 민족이 되었는가

| 박인서

네이버 검색 데이터 조회

유행은 어떻게 만들어지고 또 사라질까?
검색 데이터를 활용해 유행의 진화과정을 추적해본다.

NAVER DSC Internship

2020. 2. 5. (수)

주제 선정

“ 우리는 어떻게 마라의 민족이 되었는가 ”

: 마라탕 유행이 어떻게 만들어졌고, ‘문화’로까지 확산될 수 있었는지 검색 데이터를 통해 알아본다.

*사진 출처: 라화쿵부 공식홈페이지 (링크)

“마라 열풍” 개요

식품의약품안전처

요즘 인기 '마라탕'…식품위생법 위반 업체 37곳 적발
한국경제TV PICK | 2019.07.22. | 네이버뉴스
이번 점검은 지난 6월 3일부터 7월 5일까지 중국 사천지방을 판매하는 음식점이 대상이다. 점검 결과 식품위생령법을이 적발됐다. 주요 위반 내용은 수입...

위생 논란 점화
(2019.7.22)

*식약처에서 마라탕 전문점을 대상으로
위생법 위반업체 리스트 발표

유통업계에서도 마라맛 상품 출시

마라탕 전문점 줄지어 등장,
프랜차이즈화로 이어져

2016
2019
2020

데이터 소개

검색일자	검색어	QC (검색량)
2016-01-01	**2016새해맞이마라톤**	**2**
2016-01-01	델쿠마라디너가격	1
2016-01-01	마라도커피	3
2016-01-01	마라시fromytoy	4
2016-01-01	막스마라rail코트	1
…	…	…
2019-12-31	탕화쿵푸마라탕삼산점	5
2019-12-31	파로마라텍스모션베드	1
2019-12-31	피슈마라홍탕대전	3
2019-12-31	**행운에속지마라**	**124**

NAVER 검색 데이터

✓ 최근 4년 내

(기간: 2016-01-01~2019-12-31)

✓ “마라”를 포함한 모든 검색어

✓ 각각에 대한 일별 검색량

*전체 Row 수 약 *00만 개, QC 총합 *억 *천 회

*고유 검색어 수 약 *0만 개

Table of Contents

1. 키워드 추출하기
   - ex. 신촌마라탕 = 신촌 / 건대마라탕 = 건대
2. 키워드 군집화를 통해 이슈 발견하기
   - ex. 신촌, 건대 키워드는 '장소'라는 이슈에 해당
3. 이슈의 진화 과정을 시각화하기
   - ex. '장소' 이슈가 시간에 따라 어떻게 진화했는가

"마라" 관련 검색 유형

(1) 마라탕 포함

(ex. 신촌마라탕, 마라탕맛집)

: 차트 기반

(2) 마라탕 미포함

(ex. 마라샹궈, 마라농도)

: Suffix 기반

1. 키워드 추출하기: 마라탕 포함 vs. 미포함

(1) 마라탕 포함: 차트 기반

	검색어	검색량
1	마라탕만들기	
	건대마라탕	
	신룽푸마라탕	
	마라탕맛집	
	홍대마라탕	
6	마라탕재료	
	강남역마라탕	
	강남마라탕	
	마라탕칼로리	
	마라탕소스	

모든 검색어에 포함된 “마라탕” 제거

키워드
만들기
건대
신룽푸
맛집
홍대
재료
강남역
강남
칼로리
소스

키워드 약 70만 개

→ 상위 빈도 50개를 주요 키워드로 선택해 분석에 활용할 것

1. 키워드 추출하기: 마라탕 포함 vs. 미포함

(2) 마라탕 미포함: Suffix 기반

Idea: “마라” 관심단어들은 대부분 마라OO과 같은 패턴을 보인다.

⇒ “마라” 이후 최대 2개 음절까지를 추출한다.

ex. **마라샹궈**: 신촌**마라샹궈**, **마라샹궈**맛집
마라: 막스**마라**, ~하지**마라**

	검색어	검색량
1	마라톤	
	마라	
	마라샹궈	
	마라톤대	
	마라룽샤	
6	마라도	
	마라이브	
	마라도나	
	마라훠궈	
	마라톤일	

키워드 약 15만 개

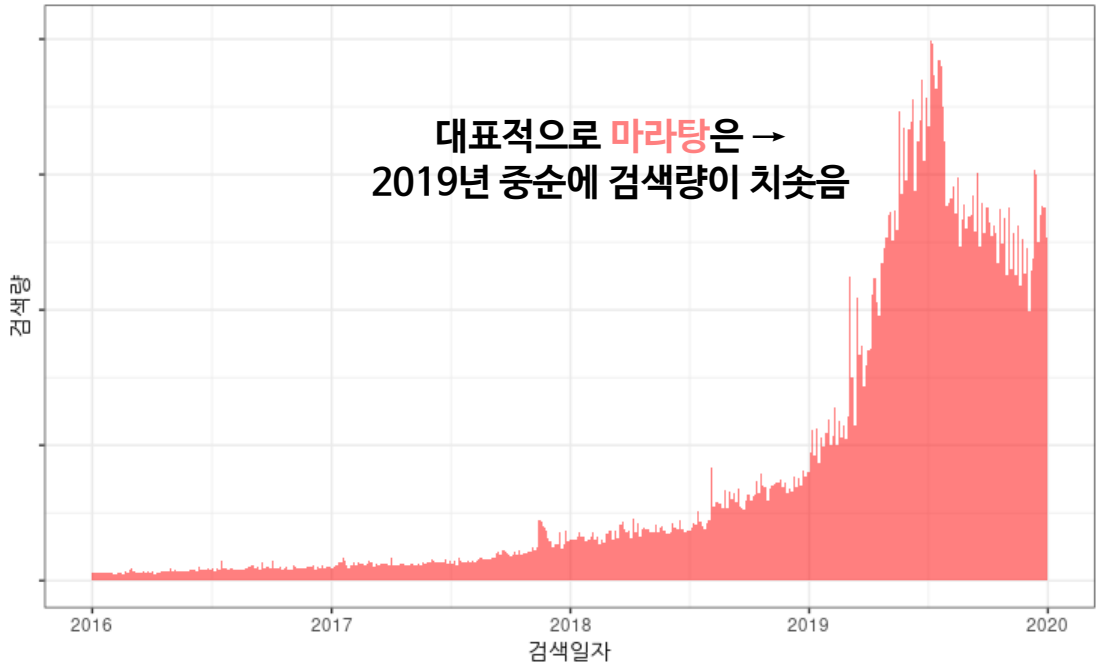
실제로 “마라”와 관련된 검색어의 선택이 관건

1. 키워드 추출하기: 마라탕 포함 vs. 미포함

(2) 마라탕 미포함: Suffix 기반

Q. 그렇다면 어떻게 “마라” 와 관련된 키워드들만 선택할 수 있을까?

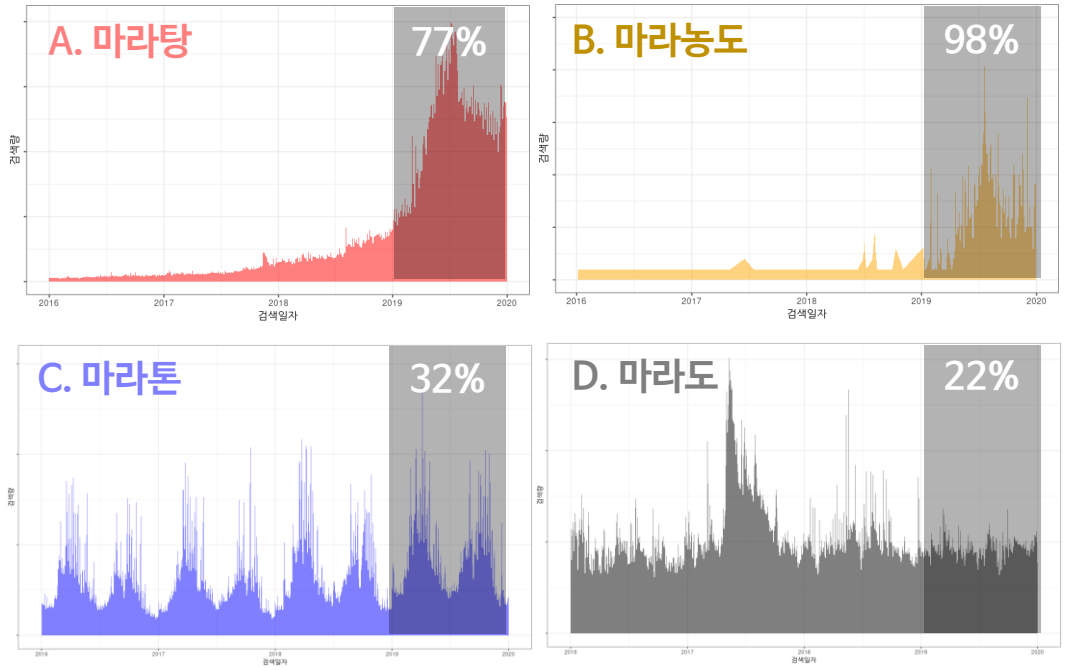
A. “마라” 관심단어들은 마라탕이 유행한 시기에 주로 검색되었을 것이다.

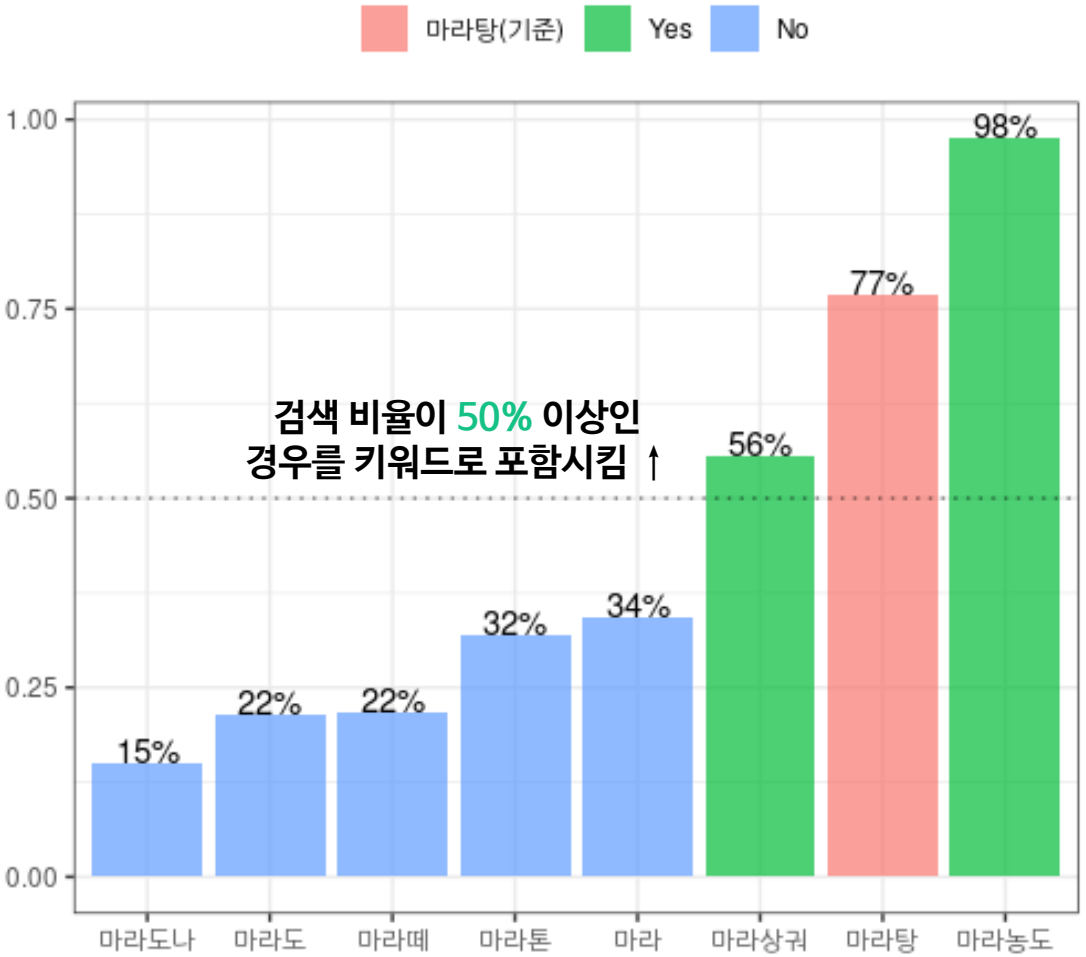


1. 키워드 추출하기: 마라탕 포함 vs. 미포함

(2) 마라탕 미포함: Suffix 기반

⇒ 유행시기(2019년) 中 검색량 비율 계산





1. 키워드 추출하기: 마라탕 포함 vs. 미포함

(2) 마라탕 미포함: Suffix 기반

	키워드	검색비율 (2019년)
1	마라샹궈	55%
	마라룽샤	53%
	마라상궈	52%
	마라소스	82%
	마라홍탕	84%
6	마라볶음	73%
	마라공방	96%
	마라치킨	77%
	마라떡볶	99%
	마라새우	59%

*마라탕 포함과 마찬가지로 상위 50개 선택함.

N = 15만

↓

1) 4년간 100회 이상 검색됨 → *N* = 1만

2) 키워드 자체로 검색어가 아닌 경우 제거 → *N* = 8천

3) 유행시기(2019년) 검색량 비율 50% 이상 → *N* = 1,532

4) 3음절 키워드와 통합

- 3음절 키워드가 포함된 4음절 키워드인 경우 통합 (ex. 마라반맛 〈 마라반)
- 단, 3음절에 비해 4음절 키워드의 검색량이 더 많다면 4음절로 통합
  (ex. 마라샹 〈 마라샹궈, 마라라 〈 마라라면)

↓

N = 997

2. 키워드 군집화를 통해 이슈 발견하기

Q. 어떻게 두 키워드가 유사하다고 판단할 수 있을까?

A. 두 키워드가 사용된 검색 기록으로 코퍼스(corpus)를 만든다.이것이 해당 검색어의 맥락이 된다!

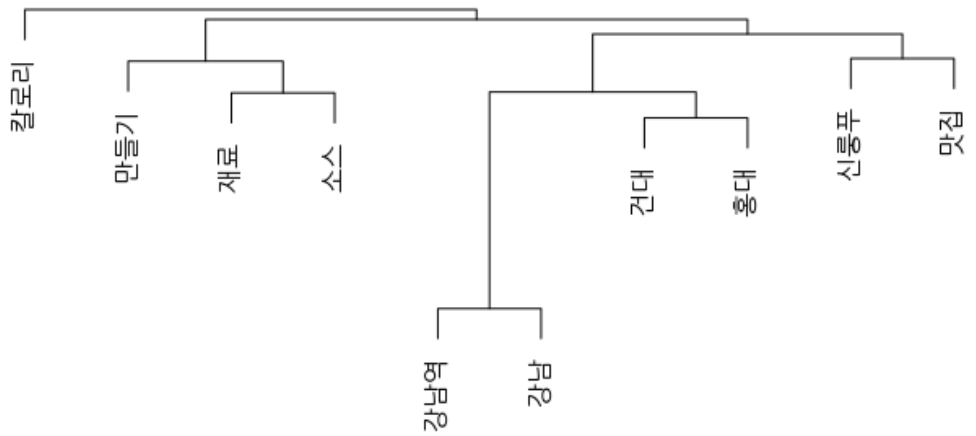
검색어 군집화 Steps

1. 키워드별 **DTM**을 만든다.
   - (키워드를 포함한 검색어) x (검색횟수)쌍들로 [키워드별 코퍼스] 생성
   - Bag of Words 이용, 비슷한 단어들이 출현한 키워드가 유사하다고 본다.

2. DTM에 기반하여 키워드 간 **거리**를 계산한다.
   - 맨해튼 거리의 변형인 캔버라 거리(Canberra distance) 이용

3. 키워드 간 거리로 **클러스터링**을 한다.
   - Hierarchical, K-Means clustering 등 이용 가능

ex. Top 10 (마라탕 포함)



2. 키워드 군집화를 통해 이슈 발견하기

Step 1. 키워드별 DTM을 만든다. (DTM: Document-Term Matrix, 문서-단어 행렬)

(예시) 2016년 1월 키워드셋 中

① 신촌 ② 호탕(상호명) ③ 건대 ④ 만들기

검색일자	검색어	검색량
2016-01-01	신촌마라탕시간	2
2016-01-01	신촌마라탕맛집	7
2016-01-01	신촌호탕마라탕	3
2016-01-01	신촌호탕마라탕가격	2
2016-01-01	호탕마라탕위치	3
2016-01-01	호탕마라탕배달	3
2016-01-01	건대마라탕맛집	8
2016-01-01	건대마라탕배달	5
2016-01-01	마라탕만들기재료	2
2016-01-01	마라탕만들기꿀팁	5

Idea: 비슷한 단어들이 함께 등장하는 키워드는 서로 유사하다.

(ex. 신촌과 건대는 '맛집' 단어 빈출)

신촌 시간 → "마라탕"을 공백으로 바꾸고

신촌 시간 신촌 시간 → 검색량(x2)만큼 반복한다.

⇒ 결과적으로 신촌 관련 코퍼스에서 "시간"이라는 단어가 2회 등장하도록 한다.

2. 키워드 군집화를 통해 이슈 발견하기

Step 1. 키워드별 DTM을 만든다. (DTM: Document-Term Matrix, 문서-단어 행렬)

Term

Document (|D| = 4)

문서 (키워드)	신촌	시간	맛집	호탕	가격	위치	배달	건대	만들기	재료	꿀팁
신촌	14	2	7	5	2	0	0	0	0	0	0
호탕	5	0	0	11	2	3	3	0	0	0	0
건대	0	0	8	0	0	0	5	13	0	0	0
만들기	0	0	0	0	0	0	0	0	7	2	5

⇒ 단어의 수(12개)만큼 차원에 흩뿌려진 벡터로 표현되었다.

∴ 키워드간 유사성 = 벡터간 유사성을 계산하는 문제다.

BoW 방식을 사용한 이유?

단어 임베딩 방법 中 BoW(Bag of Words) 방식을 사용한 것에 불과하다.
이는 검색어별 검색(등장) "횟수"라는 점에서 논리가 일치한다!

2. 키워드 군집화를 통해 이슈 발견하기

Step 2. DTM에 기반하여 키워드 간 거리를 계산한다.

- **TF-IDF (Term Freq. - Inverse Document Freq.)**
  - 지나치게 자주 등장하는 단어는 낮은 가중치를 줌
  - $tf\text{-}idf(t,d,D) = tf(t,d) \times idf(t,D)$

    Where,

    $idf(t,D) = \log\left(\frac{|D|}{df(t)}\right)$

    $|D|$ = number of all documents

    $df(t)$ = Number of documents containing the term.

- **캔버라 거리 (Canberra distance)**
  - 맨해튼 거리(L1 distance)의 변형
  - 순위(ranked list)를 비교할 때 주로 쓰임
  - $distance(x-y) = \sum \frac{|x_i - y_i|}{|x_i| + |y_i|}$

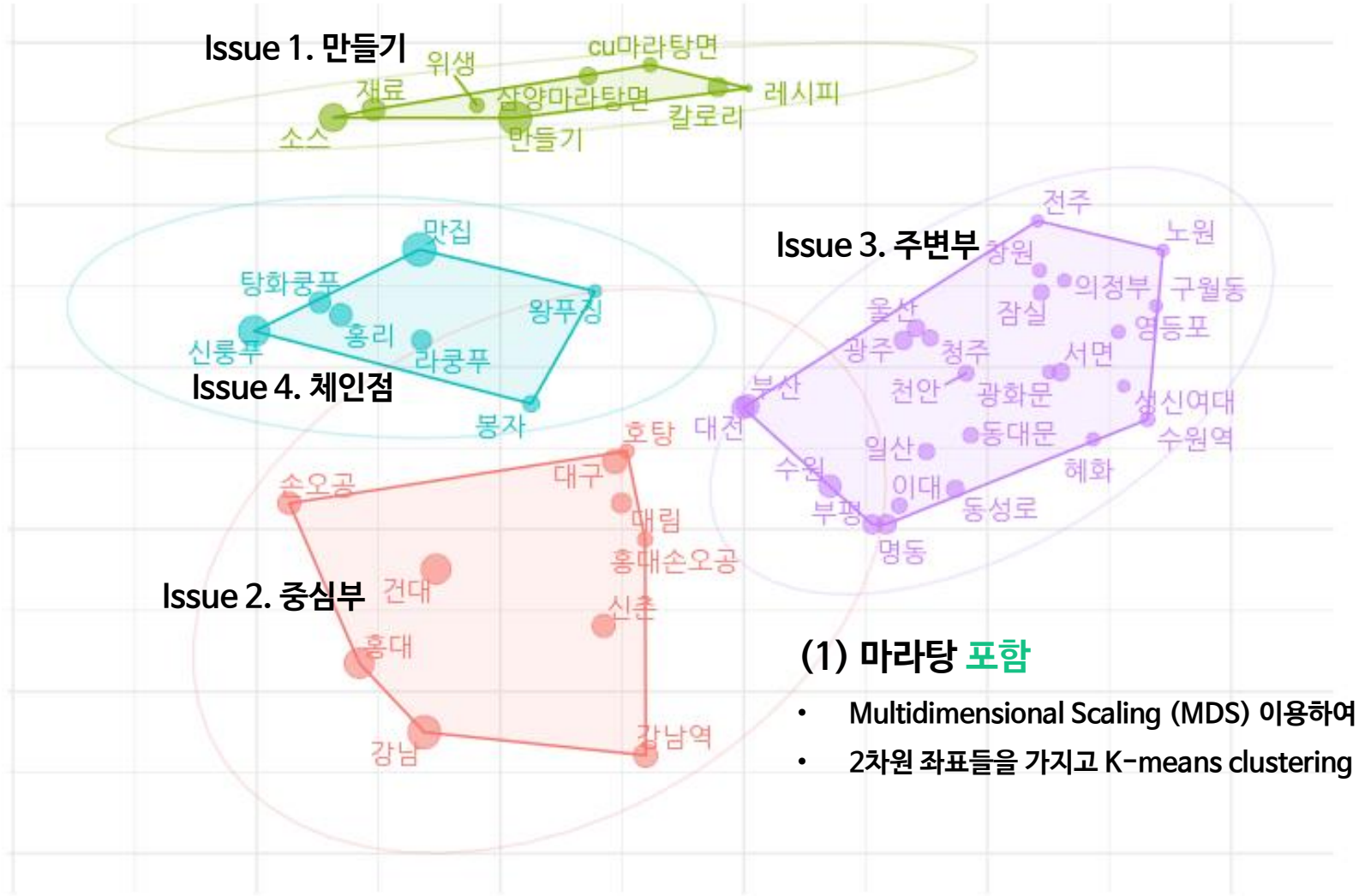
(예시) 키워드별 Distance Matrix

	신촌	호탕	건대	만들기
신촌	0	7.62	9.53	11
호탕	.	0	9.82	11
건대	.	.	0	11
만들기	.	.	.	0

⇒ **위 거리 행렬을 이용해 키워드들을 클러스터링한다.**

2. 키워드 군집화를 통해 이슈 발견하기

Step 3. 키워드 간 거리로 클러스터링을 한다.



(1) 마라탕 포함

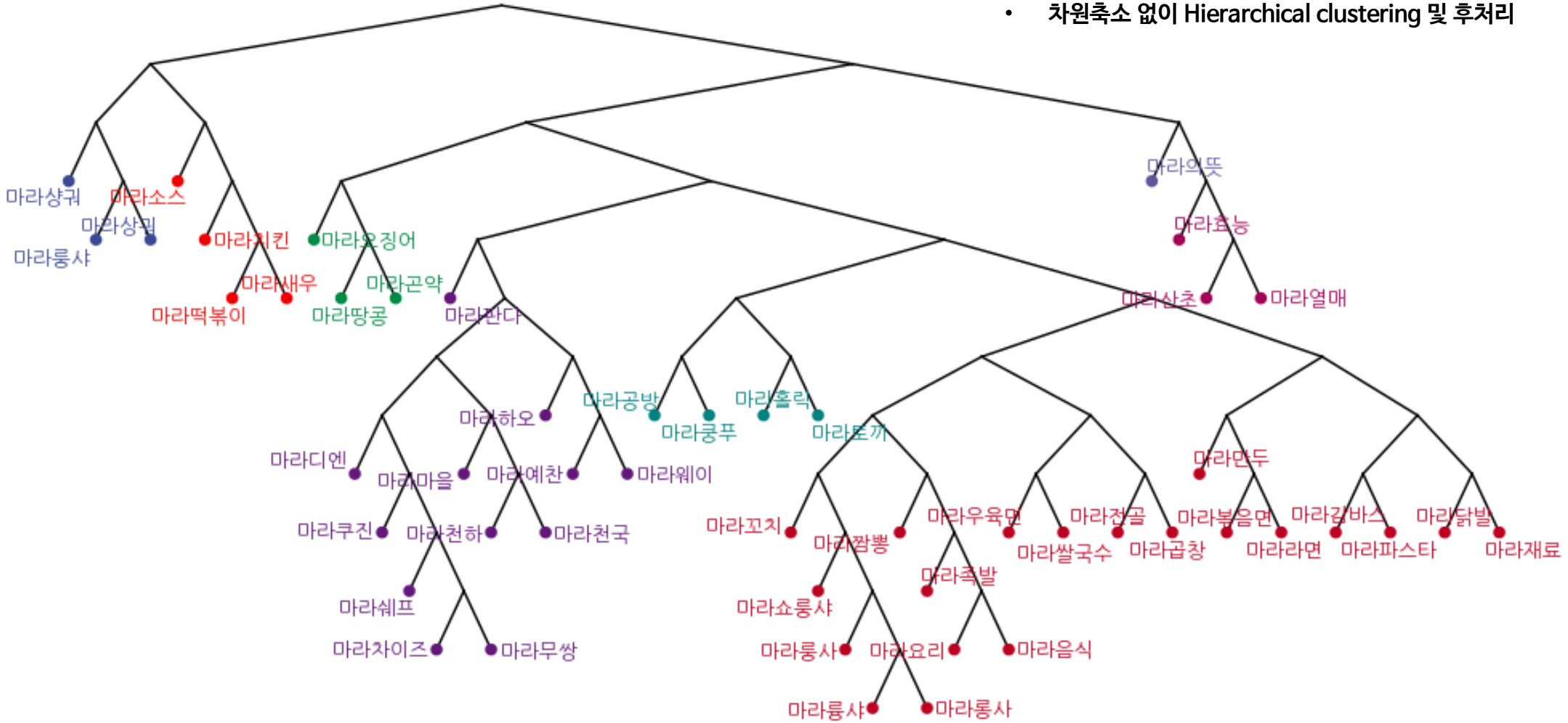
- Multidimensional Scaling (MDS) 이용하여 2차원으로 축소
- 2차원 좌표들을 가지고 K-means clustering

2. 키워드 군집화를 통해 이슈 발견하기

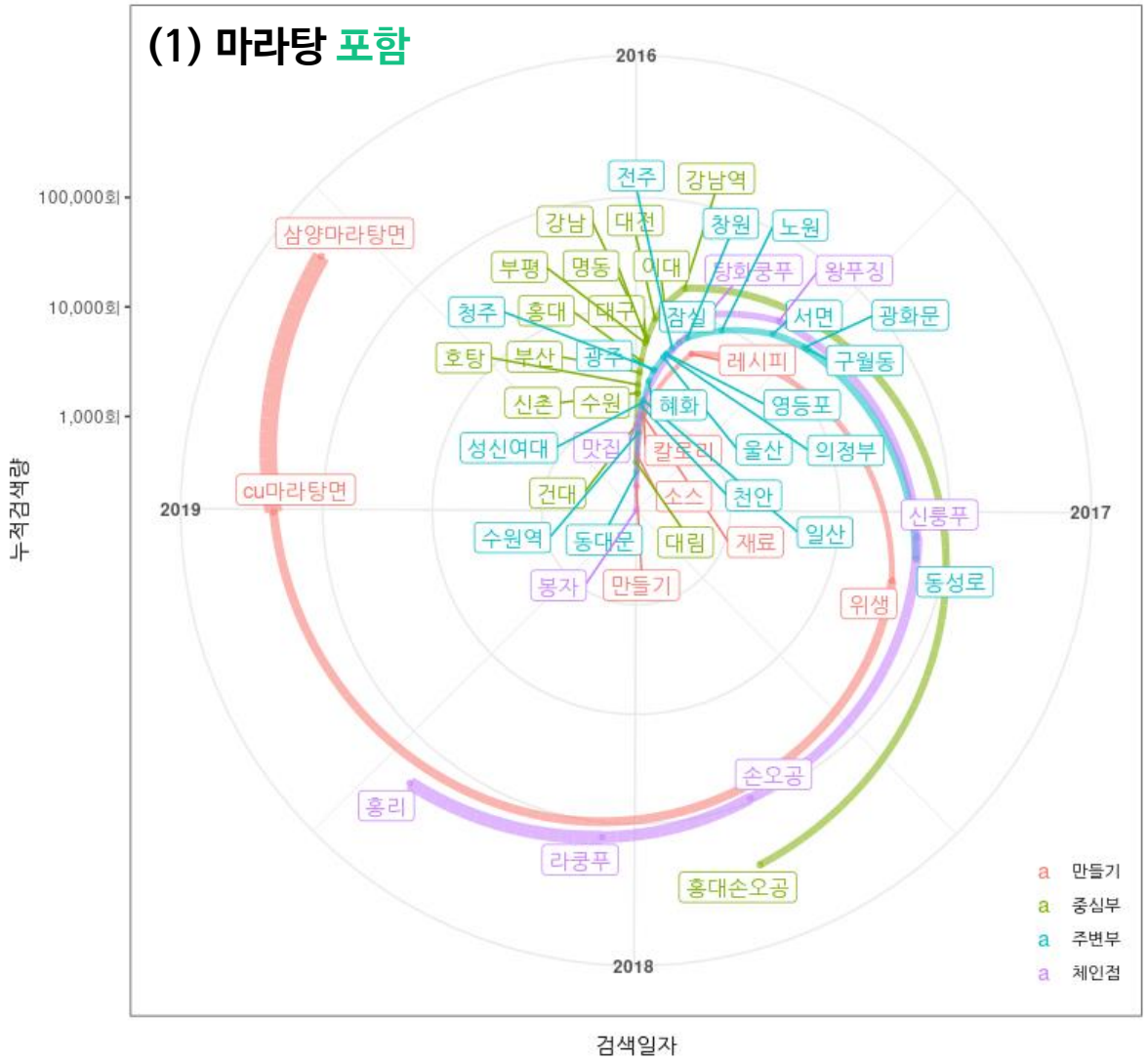
Step 3. 키워드 간 거리로 클러스터링을 한다.

(2) 마라탕 미포함

- 차원축소 없이 Hierarchical clustering 및 후처리



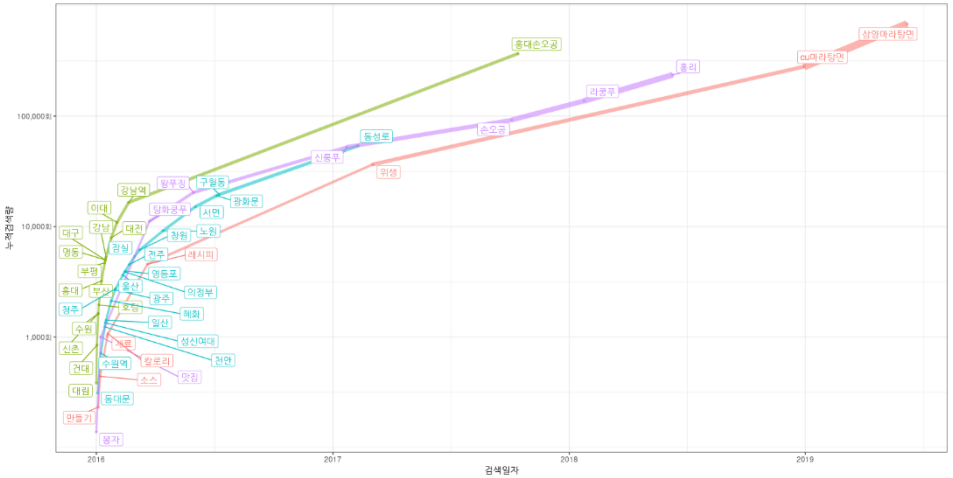
3. 이슈의 진화 과정을 시각화하기



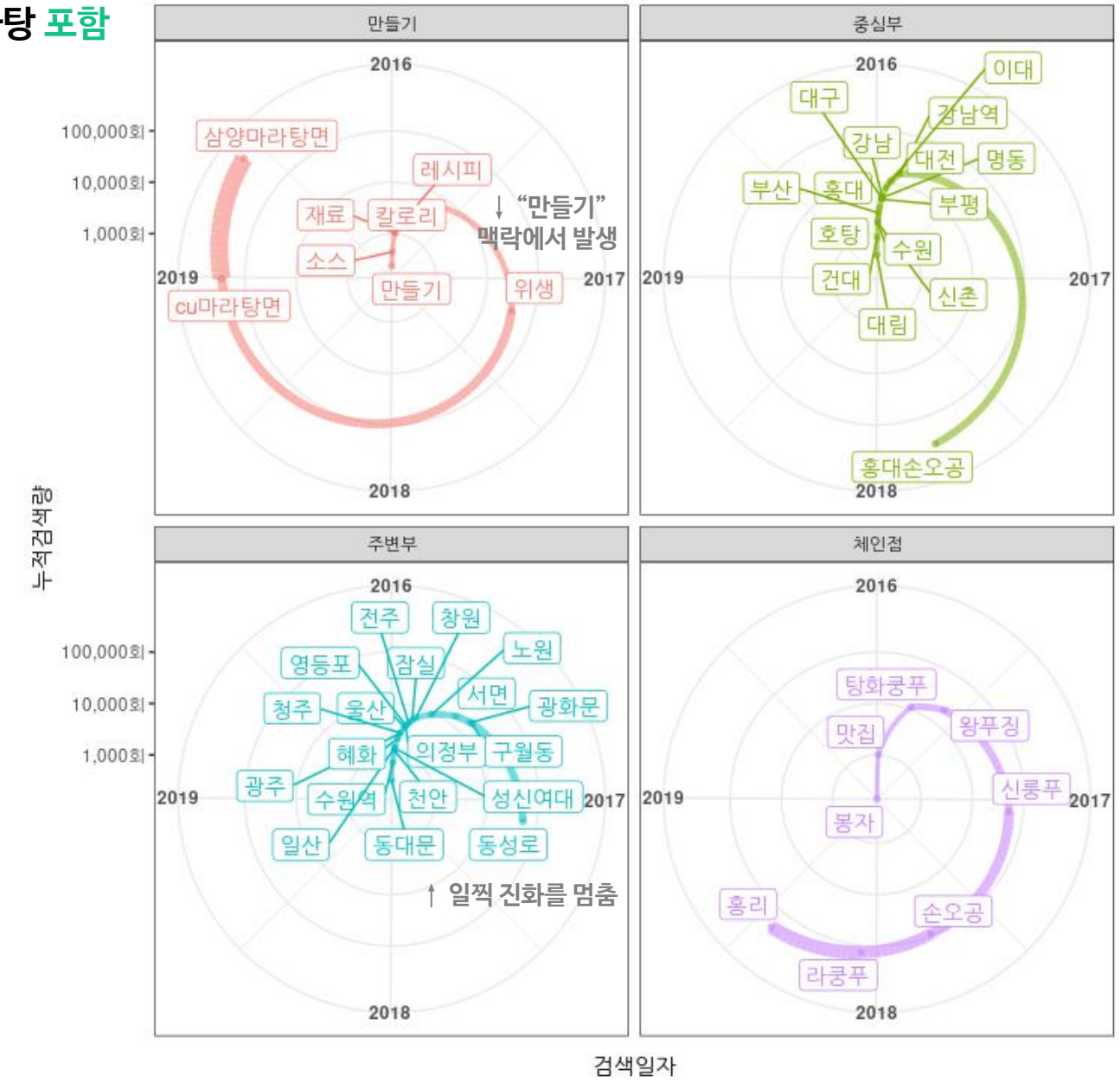
시각화 로직

- 2016년에서 **시계방향**으로 시간적 흐름을 나타냄
- “마라탕”을 포함한 상위 검색량 **50개** 키워드들 대상
- 이슈에 따라 다른 색으로 연결됨
- 누적검색량이 **100회 이상**인 일자에 해당 키워드가 ‘등장’
- 중심에서 **멀수록**, 선의 두께가 **두꺼울수록** 많이 검색된 것

cf. 극좌표를 이용하지 않는 경우?



(1) 마라탕 포함



시각화 의도

1. 각 이슈별로 진화해온 과정

- "만들기" 이슈의 경우 재료, 칼로리 등을 거쳐 위생, 마라탕면 등으로 진화를 거듭함

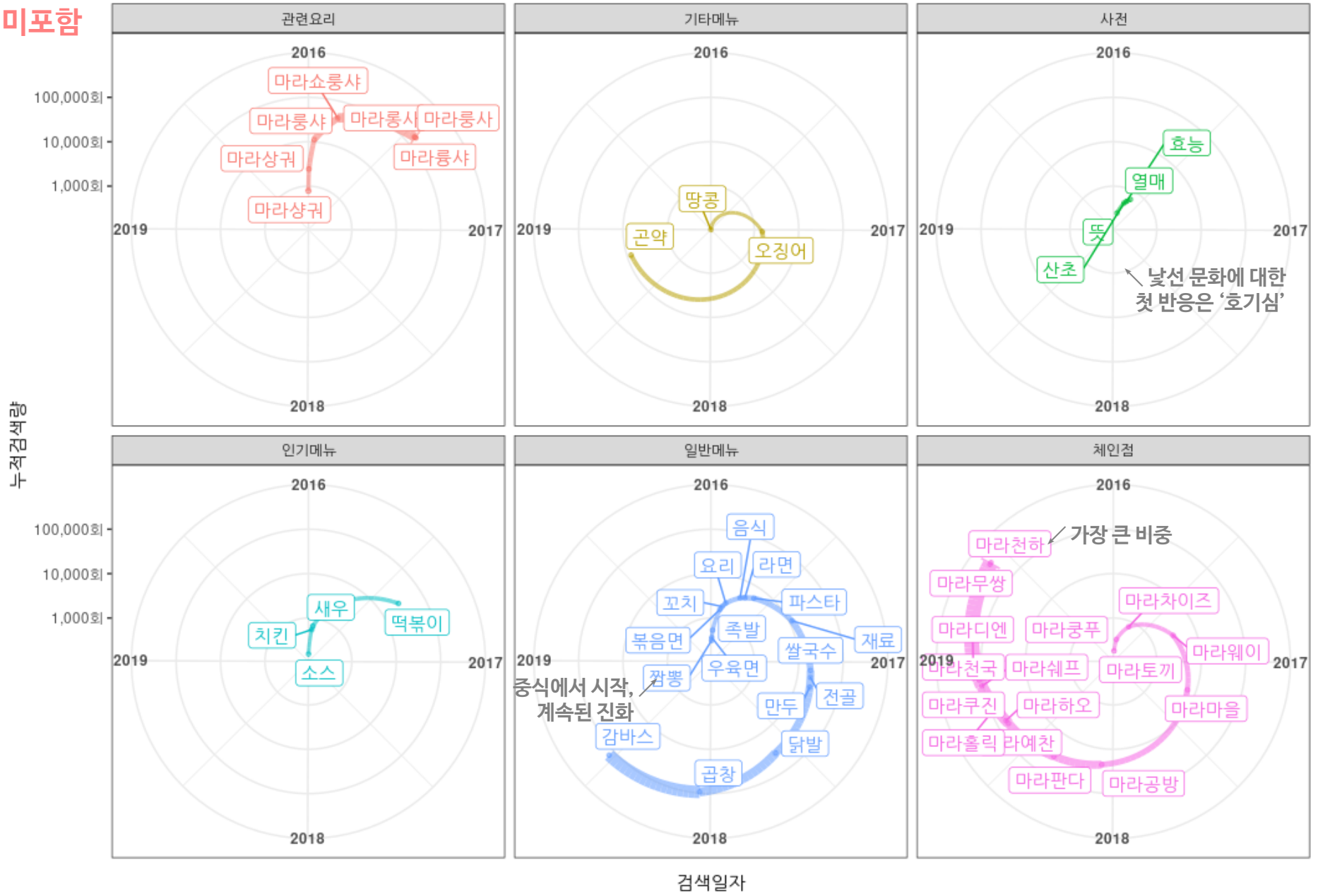
2. 진화의 속도

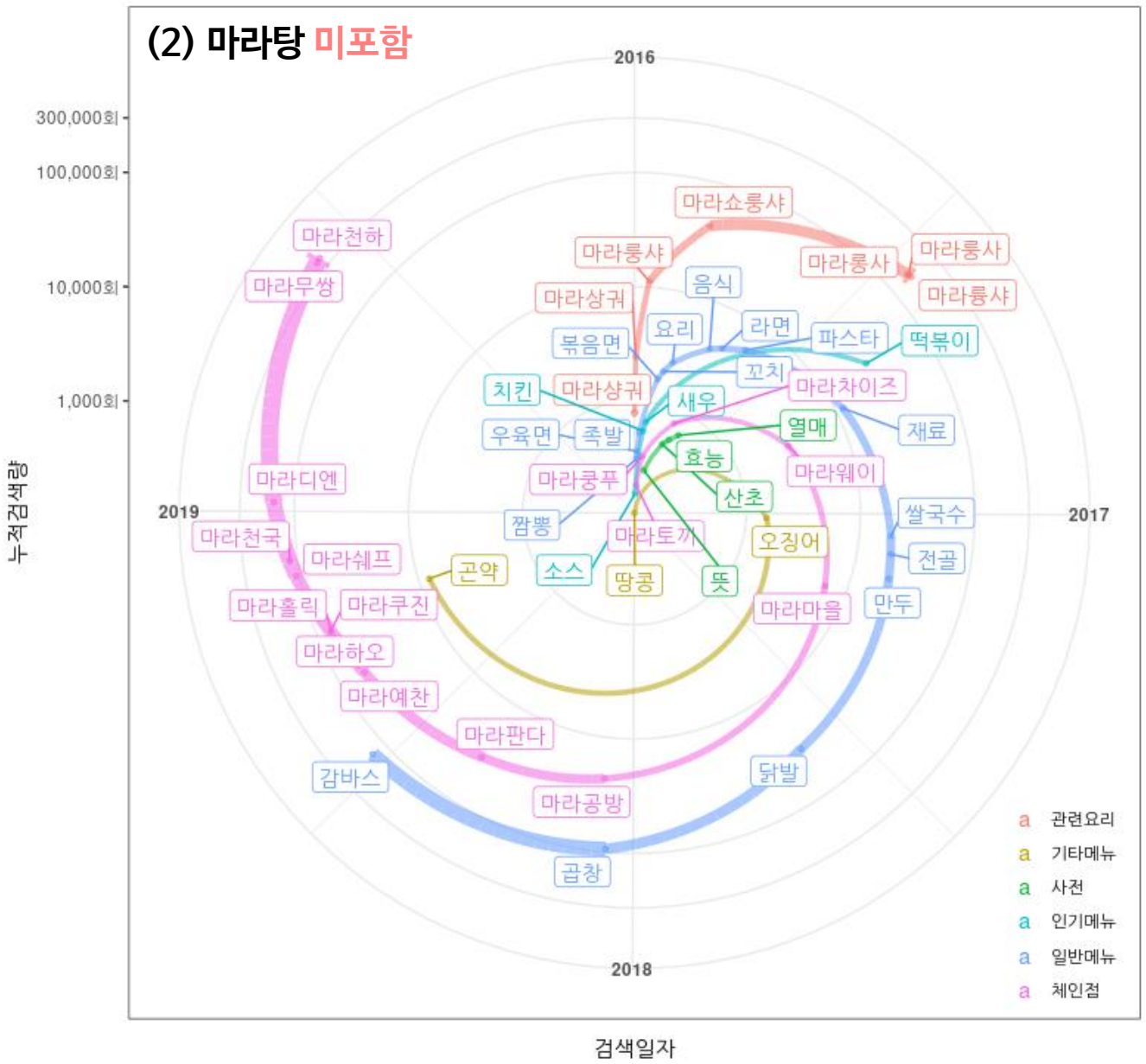
- "주변부" 이슈의 경우 빠르게 다양한 검색어들이 나타났지만 2017년 이후로는 진화를 멈춤

3. 진화의 강도

- 선의 굵기는 이슈별 누적검색량을 나타냄
- "만들기" 이슈가 가장 길고도 강도 높게 진화

(2) 마라탕 미포함


관련요리
기타메뉴
사전
인기메뉴
일반메뉴
체인점
2016
2017
2018
2019
100,000회
10,000회
1,000회
누적검색량
검색일자
마라쇼룽샤
마라룽샤
마라롱사
마라룽샤
마라상궈
마라룽샤
마라샹궈
땅콩
곤약
오징어
효능
열매
뜻
산초
낯선 문화에 대한
첫 반응은 '호기심'
새우
치킨
떡볶이
소스
음식
요리
라면
꼬치
파스타
족발
볶음면
재료
쌀국수
짬뽕
우육면
전골
만두
감바스
닭발
곱창
중식에서 시작,
계속된 진화
마라천하
가장 큰 비중
마라무쌍
마라차이즈
마라디엔
마라쿵푸
마라웨이
마라천국
마라쉐프
마라토끼
마라쿠진
마라하오
마라마을
마라홀릭
라예찬
마라판다
마라공방



시각화 의의

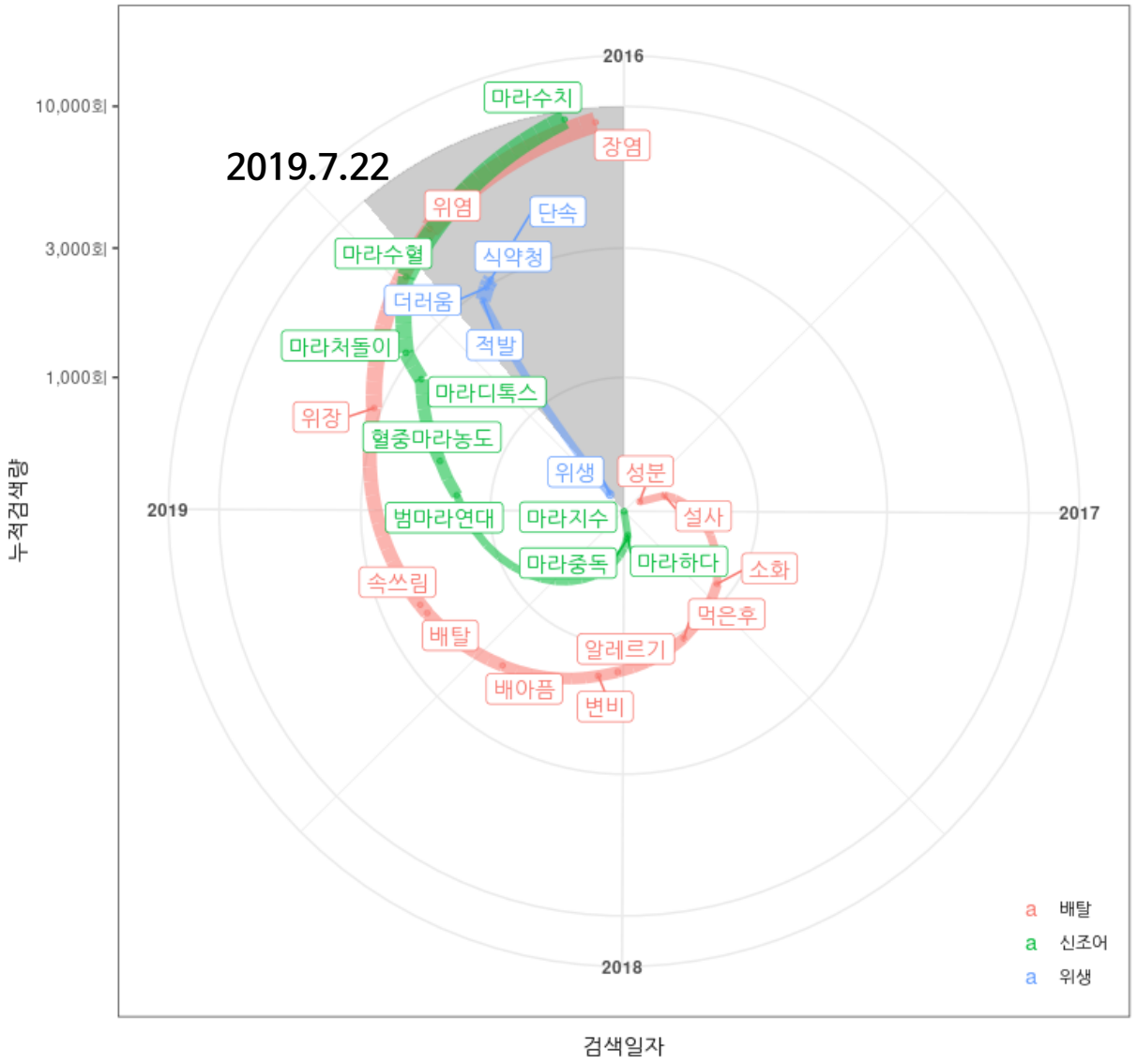
- 상위 50개 키워드들을 나타냄으로써, 마라 관련 검색어들이 주로 **어떤 키워드들로 구성**되는지 안다.
- 중심에서 멀어지는 회오리선과, 누적검색량 100회 이상일 경우 나타나는 라벨을 통해 어떤 이슈가 **진화하는 과정을 구체적인 키워드로** 살펴본다.
- 마라탕 외에도 **다른 유행**에 대해서도 적용 가능하다.

∴ 클러스터링 방법으로 유행을 요약하였다.

+ 관심 이슈 시각화

- '마라문화'의 일부인 각종 **신조어**들과, **위생**논란이라는 위기, 그리고 **배탈** 이슈가 어떻게 진화했는지 비교
- 앞서 임베딩한 결과에 거리(Canberra distance) 대신 **유사도(Cosine similarity)**를 적용, 관심 키워드에 대한 유사 단어들을 추출한다.
- $$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$ (최소 0, 최대1)

① 마라배탈 [배탈]	마라설사(.766) 마라배아픔(.562)
② 마라농도 [신조어]	마라혈중농도(.613) 마라처돌이(.134)
③ 마라위생 [위생]	마라식약청(.765), 마라적발(.438)



유행의 발견

우리는 어떻게 마라의 민족이 되었는가

| 박인서

감사합니다.