자연어처리 모델을 활용한 카페 특성 분류

데이터기반의사결정 - 3팀
김경원 문수빈 최도운 하효흔 황규민

목차

1. 프로젝트 계획

   프로젝트 목표 및 가설

2. 데이터 수집 및 전처리 과정

   데이터 수집

   전처리 과정

3. 모델 적용

   키워드 추출 방식

   임베딩 기반 분류 방식

프로젝트 목표 & 가설

프로젝트 목표

사람들이 매번 카페를 갈 때 공부를 하러 가거나, 수다를 떨러 가는 등 다양한 목적이 있다. 하지만, 카페를 검색하면 다양한 특성의 카페들이 동시에 나오기에, 소비자가 직접 찾아야 한다. 이러한 문제를 해결하고자, 개별 카페들이 가진 고유한 특징을 자동으로 분류하는 시스템을 만드는 것.

이를 통해 사용자는 목적에 맞는 카페를 더 쉽게 탐색할 수 있고, 데이터 기반의 상권 분석에도 응용할 수 있다.

가설

1 . 카페의 리뷰 데이터는 그 카페의 특성을 반영할 것이다.

ex) 만약 디저트 맛집으로 유명한 카페가 있다면, 그 카페의 리뷰글에서 “디저트”라는 단어가 가장 많이 발견될 것이다.

2 . 각각의 카페는 유사한 리뷰들로 구성되어 있을 것이다.

ex) 공부하기 좋은 카페가 있다면, 다수가 비슷한 후기를 남겼을 것이다.

데이터 수집

데이터 수집 및 전처리

데이터 수집 : 네이버 지도 리뷰 웹 크롤링

사용 기술: Selenium 라이브러리

크롤링

Selenium이란?

사람처럼 웹사이트를 자동으로 클릭하고, 입력하고,
복사할 수 있게 해주는 파이썬 라이브러리

네이버 지도는 왜 Selenium이 필요한가?

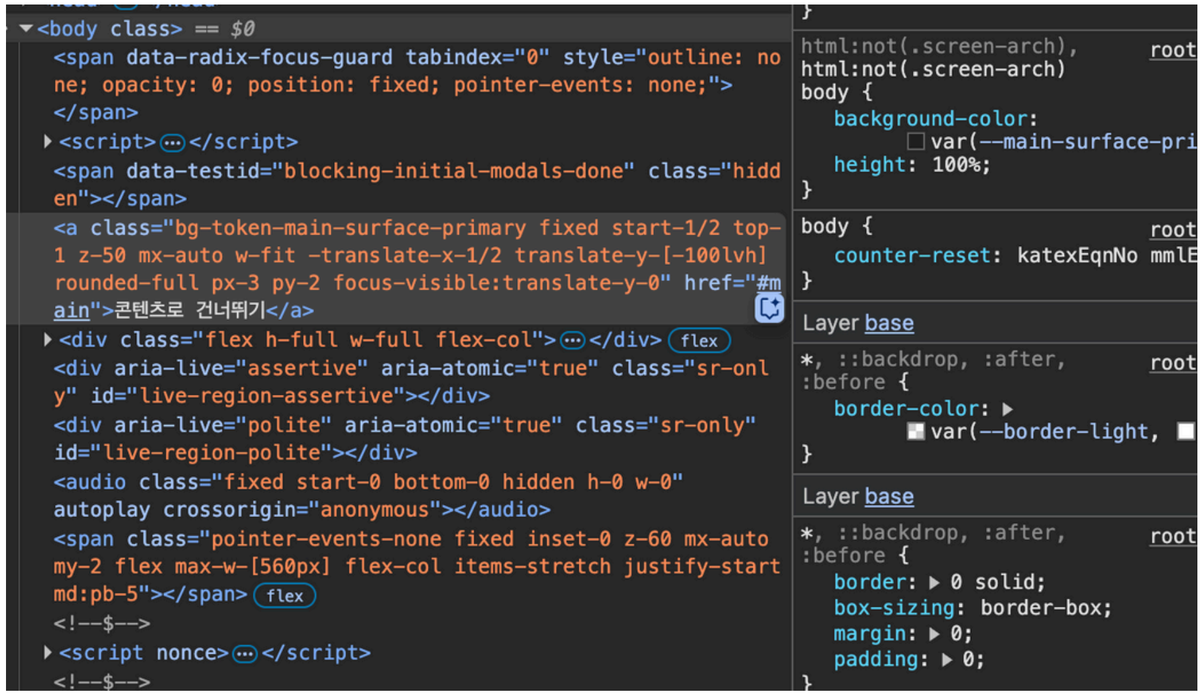
네이버지도 API에서 리뷰데이터는 제공하지 않음. 따라서,
사람처럼 행동하는 기능이 필요함.

크롤링 방식

1. 필요 정보가 담긴 웹사이트 URL에 접근

2. 필요부분의 HTML 태그 찾기
   〈DIV CLASS="REVIEW"〉이 카페 너무 좋아요!〈/DIV〉

3. 태그 안에서 필요한 부분만 뽑기
   이 카페 너무 좋아요!



크롤링 진행 단계

1. 브라우저 열기
2. 장소 검색
3. 장소 상세 페이지 접속
4. 리뷰 탭 클릭
5. 스크롤 & 더보기
6. 리뷰 내용 긁어오기
7. 데이터 저장

수집 데이터 형태

데이터 크기 : 20000
카페 개수 : 210

카페명	URL	리뷰 내용
투또톤토	https://map.naver.com/p/search/건대입구....	분위기도좋고 커피가 맛있네요.
투또톤토	https://map.naver.com/p/search/건대입구....	디저트 종류 생각보다 많고, 다 맛나요
투또톤토	https://map.naver.com/p/search/건대입구....	빵은 맛있는데, 자리가 매우 불편함.
엠케이갤러리스튜디오	https://map.naver.com/p/search/건대입구....	재방문 하고싶은 추천카페!!
엠케이갤러리스튜디오	https://map.naver.com/p/search/건대입구....	분위기 있고 음악도 좋고 맛있어요

데이터 전처리(필터링)

부정 리뷰 필터링

"공부하기 시끄러워요" → "공부“ 라는 키워드가 추가되어서, “카공하기 좋은 카페“로 분류될 수 있음
하지만 공부하기에는 부적절한 카페라는 사실을 반영하지 못한다.

→ 따라서 감정 분석 모델을 사용하여, 각 리뷰에 대한 분석을 수행한 후 부정적인 리뷰는 미리 제거하고 분류 정확도 높임

감정 분석 작동 방식

SANGRIMLEE/BERT-BASE-MULTILINGUAL-CASED-NSMC

*문장을 벡터화 후, 딥러닝 기반 분류기를 통해 감정을 예측함

입력 문장	분류 결과
“공부하기 좋은 카페예요“	긍정 87%, 부정 5%
“공부하기 시끄러워요“	부정 93%, 긍정 7%

✓ 긍정 확률이 높으면 "좋은 리뷰"
✓ 부정 확률이 높으면 "필터링 대상"으로 구분

최종 데이터 형태

긍정적인 리뷰만 모아둔 리뷰 모음

투또톤토	분위기도좋고 커피가 맛있네요.휘낭시에 말고 다른 디저트도 먹어보고 싶네요(*´`*) 라떼랑 휘낭시에 너무 맛있어요 단골이에요!!! 커피도 맛나고 사장님고 친절하세요! 매장도 쾌적하고 좋아요 자주올게요~ N번째 방문 리뷰 입니당 여태까지 방문해서 먹어 본 음료와 디저트 모두. 다. 맛있었어요!!ㅜㅜ 원래 카페에서 휘낭시에 안시켜먹는 사람인데 여기 휘낭시에 미쳤습니다. 겉바속쫀 짱이에요.. 자꾸 생각나서 먹으러 갈 수 밖에 없는 곳.. 커피 원두 세 개 중에 고를 수도 있고 좋아요!! 카페가 조용하고 깔끔하네요 ㅎㅎ 마감 전에 와가지고 사람 별로 없고 책 읽기가 넘 좋은 환경이네여! 그리고 여기 치즈케이크진짜 맛있네여 !! 사장님도 마니 친절하시구 ㅜㅜ 담에 무조건 친구 델고 올거에여!! 아 벌써 크리스마스 분위기라녀 매장 깨끗하고 조용해서 좋아요 커피랑 디저트도 맛있어요 :) 카페 너무 예쁘면서 아늑해요 구조도 특이하고,, 인테리어하신분 센스쟁이시네요 산미있는 커피로 시켰는데 산미 강한편입니다! 휘낭시에 맛있어요 언제와도 조용한 곳입니다. 시끌벅적한 카페를 좋아하지 않아서 건대에 몇 안되는 좋아하는 곳 중 하나입니다. 입구는 작은데 안에 들어가니 엄청 넓네요^^
엠케이갤러리스튜디오	재방문 하고싶은 추천카페!! 분위기 있고 음악도 좋고 맛있어요 .....

분류와 모델적용

분석 방식

리뷰 분석 2가지 방식

문장 임베딩 기반 분류

Ko-Sentence-BERT와 같은 사전학습 모델을 이용:

- 통합된 리뷰 문단 → 고차원 임베딩 벡터로 변환
  - 토크나이징: 문단을 토큰(단어 또는 하위 단어) 단위로 분해함. ["말차", "빙수", "는", "위", "에", "연유", "가", "뿌려져", "있어서", "달콤한", ...]
  - Transformer Encoding: 각각의 토큰에 대해 벡터(예: 768차원)를 생성
- 이 벡터들은 카페 리뷰의 주제적 의미를 반영
- 이후 KMeans, Spectral Clustering 등 다양한 알고리즘을 적용해 자동 분류

키워드 기반 분류

특정 키워드(예: "조용하다", "공부하기", "풍경", "데이트")가 등장하면 해당 카페를 특정 테마로 분류.

- KoNLPy로 리뷰에서 명사/형용사 추출
- Apache Spark로 워드 카운트 (빈도 기반 분석)
- Apache Pig로 테이블 조인 및 최종 분류
- 키워드 빈도수에 가중치 부여 → 테마 분류

문장 임베딩 기반 분류

임베딩 모델(SNUNLP/KR-SBERT-V40K-KLUENLI-AUGSTS) 사용

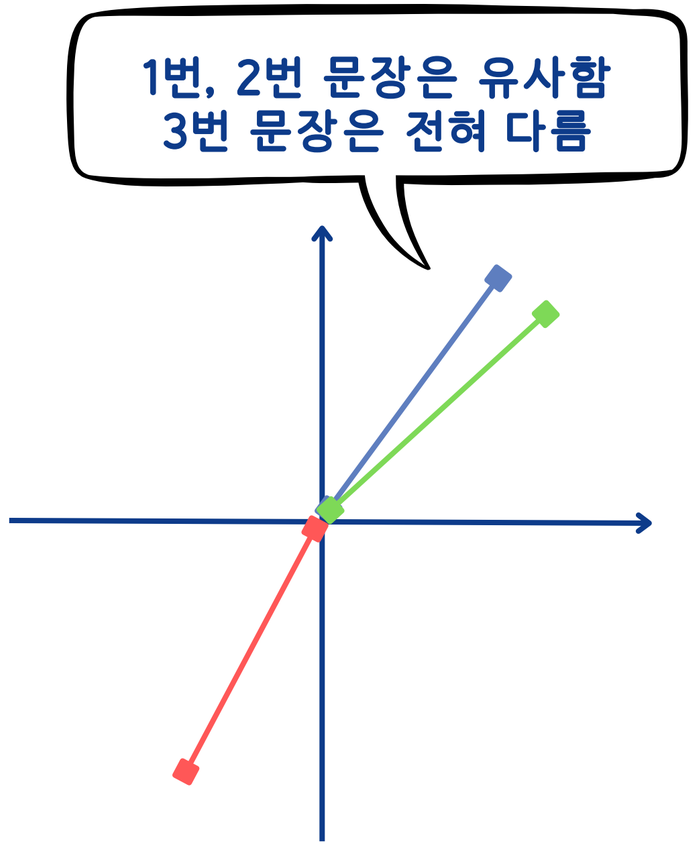
BERT : 구글이 만든 자연어 처리(NLP) 모델

BERT는 문장을 벡터화하는 데 활용될 수 있는 도구이자 모델임

"이 카페 분위기 너무 좋아요"

BERT로 임베딩
→ [0.12, -0.02, 0.48, ..., 0.91] (768차원)

단어가 정확히 일치하지 않아도, 의미가 비슷하면 벡터 거리가 가까움



“이 카페는 공부하기 정말 좋아요.”

“조용하고 집중하기 좋은 분위기예요.”

“디저트가 너무 맛있어서 깜짝 놀랐어요.”

모델 적용 가설

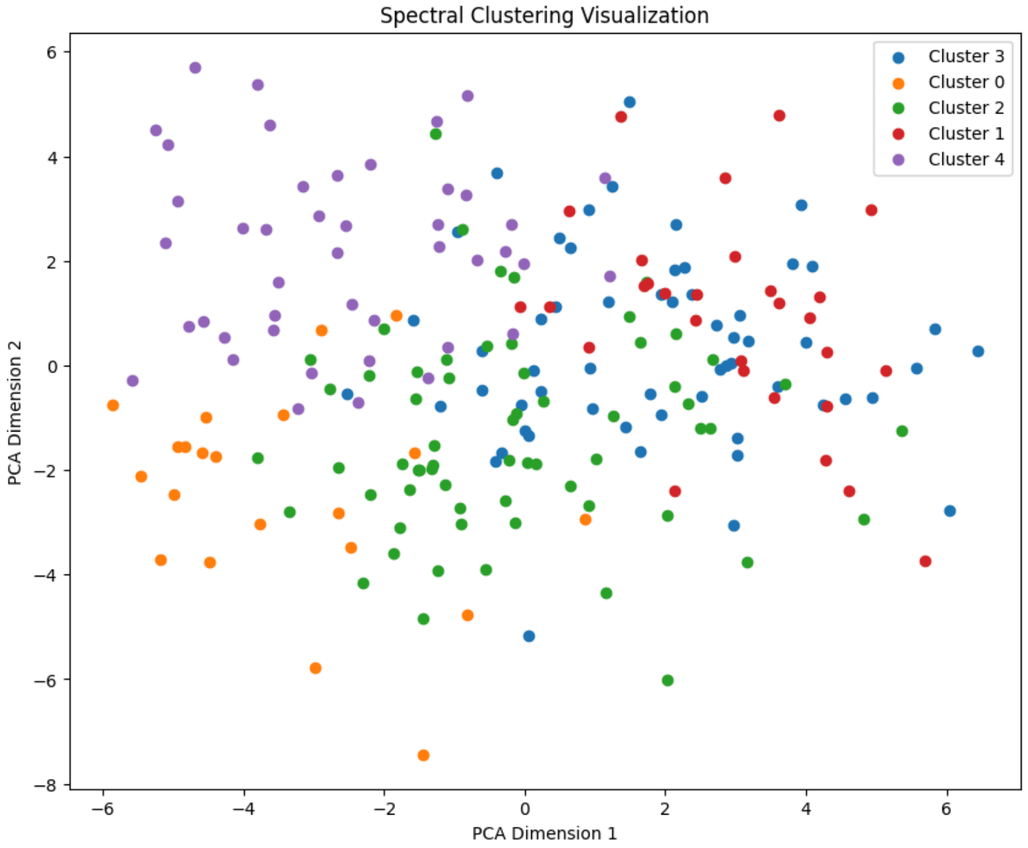
임베딩 기반 클러스터링은 키워드 방식보다 문장의 의미적 유사도를 포착하는 데 유리하며, 리뷰의 뉘앙스와 맥락을 고려한 테마 분류가 가능하다.

따라서, 이 프로젝트에서는 문장 임베딩 기반 클러스터링 방식이 더 효과적일 것이라는 가설을 설정

클러스터링 알고리즘 비교

알고리즘	원리	적합성
KMeans	데이터를 k개의 군집으로 나누고, 각 군집 중심과의 거리를 최소화	복잡한 패턴이나 경계가 명확하지 않을 경우 한계 있음
Spectral Clustering	데이터 간 유사도 행렬을 분해해서 저차원 공간으로 변환 후 클러스터링	의미 기반 문장 임베딩처럼 거리보다 유사도가 중요한 경우에 적합
Hierarchical Clustering	데이터를 하나씩 군집으로 시작 → 유사한 군집끼리 병합 또는 분할하며 계층 트리 구조	리뷰 수가 적고 계층 구조를 보고 싶을 때 적합

클러스터링 결과 - 1



CLUSTER 0

너무, 케이크, 좋아요, 맛있어요, 레터링, 케이크가, 케이크도, 진짜, 케이크를, 정말

CLUSTER 1

좋아요, 맛있어요, 너무, 넓고, 커피, 카페, 좋고, 커피도, 분위기, 스타벅스

CLUSTER 2

좋아요, 맛있어요, 너무, 커피, 가성비, 친절해요, 맛있고, 커피가, 진짜, ㅎㅎ

CLUSTER 3

너무, 맛있어요, 좋아요, 카페, 커피, 좋고, 맛있고, 분위기도, 분위기, 커피도

CLUSTER 4

맛있어요, 너무, 좋아요, 진짜, 맛있고, 와플, 정말, 맛있었어요, 에그타르트, 요거트

불용어가 너무 많음

클러스터링 결과 – 2 (불용어 제거)

Cluster 0 : “분위기 좋은 카페” – 커피, 커피도, 분위기도, 분위기, 넓고, 좋은, 디저트도, 디저트

분석: 전반적으로 커피와 디저트가 모두 좋고, 공간이 넓으며 분위기도 괜찮은 곳에 대한 이야기.

Cluster 1 : “작업하기 좋은 카페” – 커피, 넓고, 있어서, 커피도, 좋은, 분위기, 있는, 집중하기

분석: 집중하기 좋은 넓은 공간의 카페. 공부나 작업용 카페를 찾는 리뷰의 묶음.

Cluster 2 : “가성비 맛집 카페” – 커피, 커피가, 가성비, 친절해요, 빙수, 커피도, 아메리카노, 매장이

분석: 가성비 좋고, 직원 친절하며 커피와 다양한 음료가 있는 곳.

Cluster 3 : “단골 디저트 카페” – 베이글, 자주, 디저트, 맛있어서, 아이스크림, 사장님도, 젤라또

분석: 자주 가는 단골 느낌의 디저트 중심 카페. 사장님 언급도 있는 걸 보면 친근한 분위기.

Cluster 4 : “디저트 천국” – 밀크티, 맛있게, 요거트, 맛있었어요, 맛있어서, 에그타르트, 먹고, 있어서

분석: 다양한 디저트나 음료의 맛에 집중된 리뷰. 밀크티, 요거트, 에그타르트 등 특정 메뉴 언급이 많음.

Cluster 5 : “프랜차이즈 만족형” – 커피, 가성비, 스타벅스, 넓고, 매장이, 매장, 커피가, 굳굳

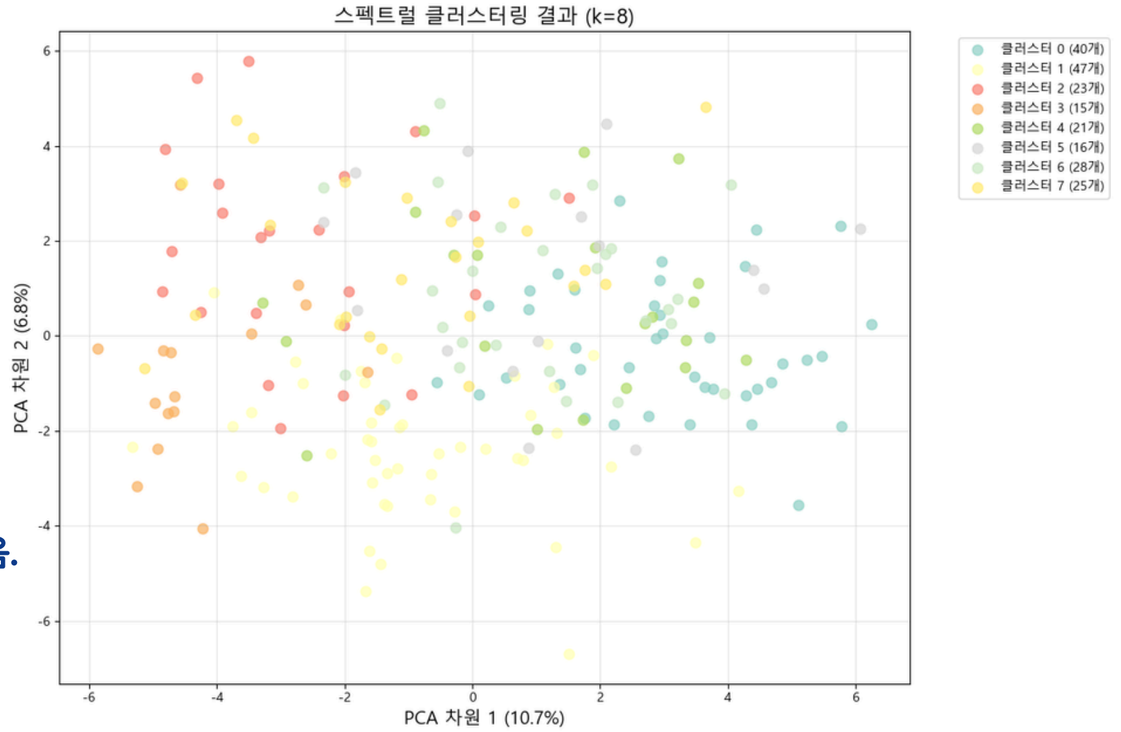
분석: 프랜차이즈나 대형 카페에 대한 리뷰일 가능성이 높고, 가격 대비 만족도가 높다는 뉘앙스.

Cluster 6 : “식사 겸용 이국풍 카페” – 빵이, 베트남, 샌드위치, 샐러드, 빵도, 친절해요, 종류가

분석: 식사 대용 메뉴(샌드위치, 샐러드 등)가 있는 이국적인 느낌의 카페.

Cluster 7 : “디자인 케이크 전문” – 케이크, 케이크가, 케이크도, 레터링, 예쁘게, 디자인도

분석: 케이크, 특히 디자인 케이크(레터링, 예쁜 비주얼 등)에 대한 언급이 중심.



단점

우리가 원하는 카테고리로 분류를 하기 어렵다.

문장 임베딩 기반 클러스터링은 데이터의 의미적 유사성을 바탕으로 자동으로 묶는 방식이기 때문에

우리가 *명확하게 원하는 카테고리(ex. 데이트카페, 카공카페, 뷰맛집)와 정확히 일치하지 않을 수 있음.

리뷰 간 의미가 너무 다양하면 클러스터가 뭉개질 수 있음

문장 임베딩 기반 클러스터링은 데이터의 의미적 유사성을 바탕으로 자동으로 묶는 방식이기 때문에

우리가 명확하게 원하는 카테고리와 정확히 일치하지 않을 수 있음.

클러스터링 결과를 “왜 이렇게 묶였는지” 사람이 직관적으로 해석하기 어렵다

임베딩 후 클러스터링된 결과는 각 군집의 중심이 고차원 벡터이기 때문에

→ “왜 이 카페들이 같은 군집으로 묶였는가?”를 사람 눈으로 바로 이해하기 어렵다

키워드 분석

1. 리뷰에서 형태소 분석기로 명사와 형용사를 추출
2. 키워드를 정의 (커피 맛집 → 진하다, 향, 드립커피 등 포함)
3. 리뷰에서 제일 많이 나온 키워드를 토대로 테마 분류 (진하다, 깊다, 고소하다 → 커피맛집으로 분류)
4. 각 테마별로 키워드 등장 순위를 정하고 5등까지 출력

```
theme_keywords = {
    "☕ 커피 향 가득한 카페": [
        "진하다", "깊다", "고소하다", "향기롭다", "향", "바디감", "농도", "묵직하다", "원두",
        "핸드드립", "에스프레소", "로스팅", "싱글오리진", "추출", "드립", "콜드브루",
        "플랫화이트", "전문적", "수준", "퀄리티", "전문점", "정성"
    ],
    "🍰 디저트 천국": [
        "케이크", "디저트", "마카롱", "초코", "크림", "빵", "와플", "젤라또", "타르트", "빙수"
    ],
    "💸 실속 있는 선택": [
        "가성", "가성비", "혜자", "합리적", "알찬", "푸짐", "구성", "가격대비", "저렴", "싸다", "자주", "단골",
        "테이크아웃", "빠르다", "편의성", "간단", "프렌차이즈", "체인점", "메가", "컴포즈", "이디야", "포즈", "빽"
    ],
    "💻 집중하기 좋은 곳": [
        "공부", "노트북", "작업", "조용", "집중", "콘센트", "와이파이", "자리", "타이핑", "앉다"
    ],
    "💖 함께 오기 좋은 공간": [
        "데이트", "연인", "커플", "애인", "로맨틱", "기념일", "남자친구", "여자친구", "사진찍기좋다", "설레다", "분위기", "좋은시간"
        "달달하다", "감성", "감성있다", "따뜻하다", "분위기좋고", "함께"
    ]
}
```

🍰 디저트 천국
1. 릴리베이커리
2. 세이케이크
3. 카페 그자체 베이커리베이커리
4. 레스트인무드
5. 레브 디저트카페

💻 집중하기 좋은 곳
1. MouseRabbit카페
2. 카공족 어린이대공원역점스터디카페
3. 앤딩스터디카페 자양점
4. 카페온더플랜 건대점카페,디저트
5. 할리스 건대입구점카페

💸 실속 있는 선택
1. 컴포즈커피 건대스타시티몰점카페
2. 컴포즈커피 건대양꼬치거리점
3. 컴포즈커피 자양신양초교점
4. 빽다방 광진화양삼거리점테이크아웃커피
5. 고망고 건대점

💖 함께 오기 좋은 공간
1. 플레어 에스프레소 바
2. 엠케이갤러리스튜디오
3. 브루크 성수점
4. 심리카페멘토 건대점
5. 구르다방

☕ 커피 향 가득한 카페
1. 리틀히포 로스팅랩카페,디저트
2. 에스프레소 바 시 자양점
3. 건대커피 랩
4. 최가커피 자양점
5. 코에오 COEO카페,디저트

어떤 방식이 더 정확한가?

문장 임베딩 기반 분류의 한계

- 리뷰의 유사도를 판단한다는 점에서, 우리가 유도한 종류의 분류가 잘 진행되지 않음
- 의미적 유사성과 실제 카페 테마 분류가 일치하지 않는 경우가 많음

키워드 기반 분류의 우수성

- 원하는 목적에 맞는 키워드의 빈도수 분석을 통해 좀 더 정확한 분류가 가능
- 실제 사용자가 카페를 선택할 때 고려하는 요소들과 직접적으로 연결됨

결론 도출

- 단순한 키워드 검색을 넘어, 리뷰 속에 담긴 문맥과 분위기를 정략적으로 이해하여 카페를 의미적으로 분류하는 시도
- 텍스트 임베딩과 클러스터링을 활용한 이 방법은 소비자 취향을 반영한 맞춤형 탐색 서비스나 상권 분석 도구로 확장 가능함
- 임베딩 방식과 키워드 방식 중 임베딩 방식이 더 좋을 것이라는 가설을 세웠으나, 결과값이 좋지 않았음
    - 그래서 키워드 방식을 사용했더니 좋은 결과값이 나왔음

확장 가능성

자동 분류기 생성 (지도 학습으로 확장)

- 클러스터링 결과를 라벨처럼 활용하면, 이후 새로운 카페 리뷰를 입력했을 때 테마를 예측할 수 있는 지도 학습 모델로 확장 가능

상권 분석 및 트렌드 추적

- 시간에 따른 리뷰 변화, 인기 테마, 지역별 특성 분석 등 마케팅/프랜차이즈 전략에 응용 가능

비즈니스 인사이트

소비자 의사 결정 지원 관점

1. 소비자는 더 이상 리뷰를 일일이 읽을 필요 없음

2. 의도에 맞는 '정확한' 카페 선택 가능

3. 추천 알고리즘의 기반이 될 수 있음

4. 리뷰는 부정적일지라도 '의미 있는 분류 정보'

5. 소비자 맞춤형 카페 지도 제작 가능

소비자 인식 기반의 테마별 시장 세분화

임베딩을 통해 자연스럽게 소비자가 인식한 테마별로 카페를 군집화할 수 있음.

단순히 '카페'가 아니라,

→ 데이트 목적, 공부 목적, 인스타 감성, 조용한 공간, 가성비 등

소비자 시선에서 분리된 시장 세그먼트를 파악 가능.

이는 BI 관점에서 고객 중심 카테고리 전략 수립에 바로 활용 가능함.

위치 기반 전략 도출 (GIS 데이터와 융합 가능)

클러스터링 결과를 지역 데이터와 결합하면,

→ 예: “강남 지역에는 ‘카공 중심’ 카페가 밀집되어 있다”

→ “홍대 지역에는 ‘디저트 감성’ 카페 클러스터가 주를 이룬다”

이는 상권분석 및 신규 매장 입지 전략 수립에 유용함.

프랜차이즈나 커피 브랜드 본사에서 지역별 테마 카페 전략을 수립하는 데 핵심 자료가 됨.

브랜드 전략 방향성 제시

예: A 카페가 원래는 '카공' 브랜드로 인식되길 바랐지만, 클러스터 결과는 '디저트 중심' 클러스터에 속한다면,

브랜드 전략과 소비자 인식 간 괴리가 있는 것임.

이를 바탕으로,

브랜딩 방향 수정 (광고, 인테리어, 메뉴 전략 등)

혹은 소비자 인식 전환 마케팅 실행 가능.

맞춤형 추천 알고리즘 개발의 전초작업

클러스터링을 통해 소비자 리뷰 기반 카페 특성 벡터를 확보하면,

향후 개인화 추천 시스템에서 매우 강력한 feature로 활용 가능함.

예: “카공 목적 + 노이즈 민감”한 사용자에게는, 해당 클러스터에 속한 카페를 자동 추천.

한계점

1. 데이터 총량 부족
   "데이터 크롤링에 시간이 오래 걸림
2. 리뷰의 비핵심 문장까지 과도하게 반영
   "근처에서 친구랑 만났는데, 오다가 비가 왔어요. 그래도 카페 분위기는 조용했어요."
3. 모델은 사람처럼 '의도'를 파악 불가
   모든 단어, 문장을 동등하게 취급하거나, 가중치를 부여하더라도 맥락 해석이 제한됨
4. 도메인 불일치 문제
   임베딩 모델은 일반 뉴스 / 백과 문장으로 학습됨, 하지만 실제 리뷰는 말투가 다르고 짧거나 줄임말이 많음
5. 문장과 테마의 표현 방식이 너무 달라 유사도 계산 왜곡 가능
   테마 문장은 보통 한 두 단어로 표현, 반면 리뷰는 다양한 요소가 뒤섞임

최종 결론

- 임베딩 방식은 의미적 유사도를 계산할 수 있어 이론상으로는 고급 방식이지만,
- 실전에서는 짧고 목적이 뚜렷한 리뷰 분류에는 오히려 키워드 방식이 더 효과적일 수 있음.
- 특히 이번 프로젝트처럼 주제 분류 기준이 명확하고, 사용자의 기대 방향이 뚜렷한 경우,
- 키워드 방식이 정확하고 직관적이며 컨트롤하기도 좋음

THANK YOU