

Datepop

매의 눈 프로젝트

프로젝트 진행 기간 : 2025.03.01~06.30



기다슬

이전 프로젝트

가망 매장("컨택하면 좋을 매장") 자동 관리 시스템 구축

- 기존 매장 검수 시스템 일부 자동화
- 인기도 예측 (datepop내 판매량 * 가격 -> 매출)
 - XGboost / Graph Neural Network / MLP
- 기존 매장과의 유사도 -> concise feature embedding + cosine similarity

이전 프로젝트

가망 매장("컨택하면 좋을 매장") 자동 관리 시스템 구축

- 기존 매장 검수 시스템 일부 자동화
- 인기도 예측 (datepop내 판매량 * 가격 -> 매출)
 - XGboost / Graph Neural Network / MLP
- 기존 매장과의 유사도 -> concise feature embedding + cosine similarity

가망 매장의 기준점이 정해지지 않아 초기엔 인기도를 바탕으로 선정을 하였지만,
실제 매장의 후보들을 선택하는 기준인 QC_score 목록이 존재함.

이를 통해 1차적인 필터링이 가능하도록 자동화 하는 것을 목표로 설계함.

새로운 프로젝트 목표

가망 매장("컨택하면 좋을 매장") 자동 관리 시스템 구축 재 설계

- 데이터 수집 구조 변경
- 데이터 분석
- QC-Score 기준점을 바탕으로 파이프라인 구축
- 테스트 및 평가

프로젝트 기대 효과

가망 매장("컨택하면 좋을 매장") 자동 관리 시스템 구축 재 설계

LLM 기반의 분석 파이프라인을 통해 네이버와 카카오맵의 분산된 데이터를 하나의 정량적 QC 스코어로 통합하여

자동화된 1차 필터링 시스템은 영업 및 QC팀의 수동적인 정보 분석 업무를 제거하고, 검증된 잠재 고객에게만

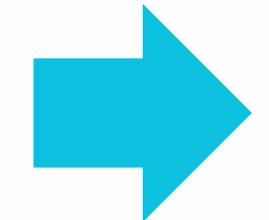
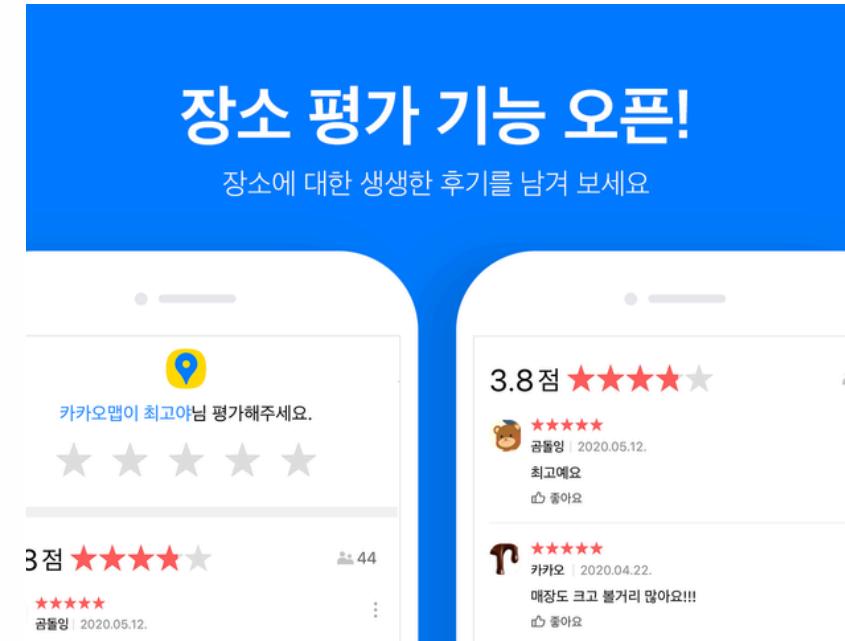
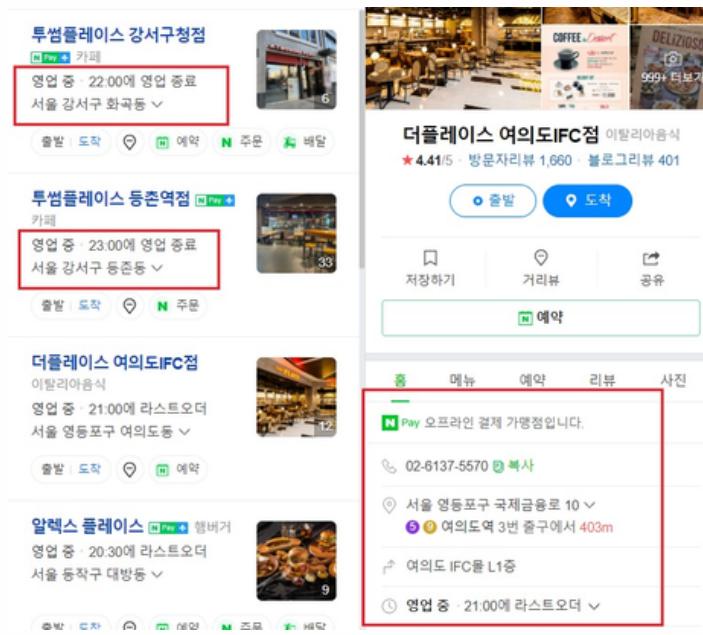
집중할 수 있는 환경을 제공하여 결과적으로 업무 리소스의 부담을 줄일 수 있을 것으로 예상됨.

Process



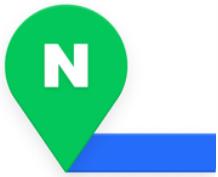
주간 미팅 단위의 피드백 루프(Loop)를 기반으로 애자일 방법론을 적용함.

Pipeline



Gemini

LLM을 통한 점수 산출



네이버 맵



카카오 맵

네이버 맵에서 수집한 '정량적' 매장 정보와 카카오 맵의 '정성적' 리뷰 데이터를 통합하여 QC_Score를 산출함.
특히 LLM을 활용하여, 단순 평점만으로는 파악하기 어려운 리뷰의 잠재적 의미와 맥락을 해석하는 의미론적 분석을 수행하도록 설계함.

데이터 수집 - 네이버 지도(1)

네이버 지도 검색 결과

검색어: 가로수길 맛집

지도 화면: 가로수길 맛집 목록과 각 맛집의 상세 정보(예약, 리뷰, 사진 등)를 표시합니다.

상세 맛집 정보 화면 예시:

- 우살롱**: 서울 강남구 압구정역 근처에 위치한 한우 전문 레스토랑입니다. 프라이빗 & 프리미엄 우살롱은 데이트, 회식, 접대, 동호회를 위한 압구정 한우 레스토랑으로 최고의 모임 공간과 음식을 제공해 드립니다.
- 우살롱**: 최고급 1++ 한우와 독창적인 음식 그리고 와인/위스키를 함께 즐길 수 있는 압구정 소고기 맛집입니다. 모두 단독룸으로 구성되어 있으며 등심, 채끝, 꽃등심, 안심, 살치살, 새우살, 안창살 모두 서버가 구워드
- 어썸로즈 가로수길점**: 서울 강남구 압구정역에 위치한 한우 전문 레스토랑 우살롱입니다. 프라이빗 & 프리미엄 우살롱은 데이트, 회식, 접대, 동호회를 위한 압구정 한우 레스토랑으로 최고의 모임 공간과 음식을 제공해 드립니다.

DataLab. (2025.01.21. 업데이트)

테마키워드

- 분위기: 분위기 좋은, 맛있는, 다양한, 데이트하기 좋은, 가성비 좋은
- 인기토픽: 레스토랑, 레드와인, 스테이크샐러드스타, 반려동물, 생화
- 찾는목적: 기념일, 나들이, 소개팅, 특별한날, 카메라
- 수요일, 시간대별 검색 인기도: 일, 월, 화, 수, 목, 금, 토

연령별 / 성별 검색 인기도

10대 20대 30대 40대 50대 60대
남자 여자

데이터랩은 네이버 내/외부의 빅데이터 분석을 통해 생성한 정보입니다.

데이터 수집 - 네이버 지도(2)

The screenshot displays a restaurant's menu and review section. On the left, there are five menu items with their names and prices: 어복쟁반 (80,000원), 물냉면 (16,000원), 들기름 비빔면 (16,000원), 제육 (25,000원), and 내장 (25,000원). Each item has a small image of the dish. To the right, a list of reviews is shown, each with a rating icon, the review text, and the number of likes (e.g., 19, 10, 9, 6, 6). Below the reviews, a button says "리뷰 클렌징 시스템 작동중입니다". At the bottom, there is a profile for a user named "다산양양" with 1,503 reviews and 503 photos, and a link to their Instagram account.

+ 메뉴 정보, 리뷰 정보(최신순 5개)

기본 정보

- naver_id : 네이버 장소 ID
- search_word : 검색어
- name : 매장 이름
- category : 카테고리
- new_store : 신규 매장 여부
- address : 주소
- phone : 전화번호
- gps_latitude : 위도
- gps_longitude : 경도
- naver_url : 네이버 지도 URL

리뷰 및 인기도 정보

- visitor_review_count : 방문자 리뷰 수
- blog_review_count : 블로그 리뷰 수
- review_category : 리뷰 키워드 및 개수
 - 디저트가 맛있어요
 - 특별한 메뉴가 있어요
 - 커피가 맛있어요
 - 음료가 맛있어요
 - 매장이 청결해요
- review_info : 개별 리뷰 정보 (리뷰별 반복)
 - date : 날짜
 - comment : 내용

부가 정보

- instagram_link : 인스타그램 링크
- instagram_post : 인스타그램 게시물 수
- instagram_follower : 인스타그램 팔로워 수
- theme_mood : 테마 (분위기)
- theme_topic : 테마 (주제)
- theme_purpose : 테마 (목적)
- distance_from_subway : 지하철역과의 거리 (m)
- distance_from_subway_origin : 지하철역 출구 기준 상세 정보
- on_tv : TV 방송 여부
- parking_available : 주차 가능 여부
- seoul_michelin : 서울 미쉐린 가이드 선정 여부

고객 분석 정보

- age_2030 : 20-30대 방문 비율
- gender_balance : 성별 균형 여부
- gender_male : 남성 방문 비율
- gender_female : 여성 방문 비율
- running_well : 활발히 운영 중인지에 대한 지표

메뉴 정보

- menu_list : 메뉴 목록 (메뉴별 반복)
 - name : 메뉴 이름
 - intro : 메뉴 소개
 - price : 가격
 - isRepresentative : 대표 메뉴 여부

데이터 수집 - 카카오 맵

A screenshot of the Kakao Map application showing search results for "음식점" (Restaurant). The map displays several food establishments in the area around Seongnam Station. A specific restaurant, "밀밭칼국수", is highlighted with a red circle. Other visible locations include "GONGSKIN" (an advertisement for a 900won event), "상가", "푸리닭치킨", and "메가 MGC 카페". The interface includes a search bar, filters for walking routes, buses, and subways, and a sidebar with a GONGSKIN advertisement.

4.5 ★★★★☆

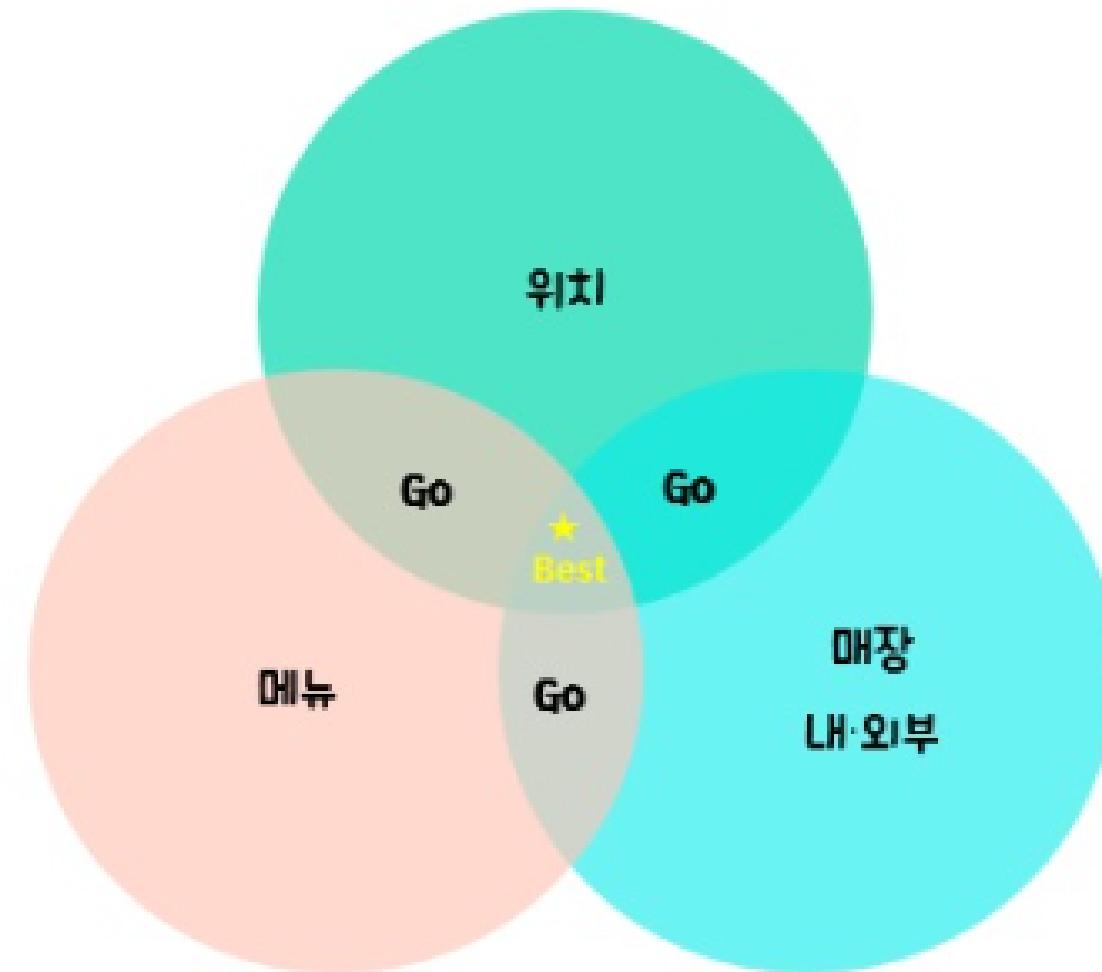
후기 38



카카오맵 상세 데이터

- `kakao_score` : 카카오맵의 전체 별점 (5점 만점)
- `kakao_review` : 카카오맵의 전체 후기 개수
- `kakao_taste` : 맛에 대한 평가 점수
- `kakao_value` : 가격 대비 만족도(가성비)에 대한 평가 점수
- `kakao_kindness` : 서비스 및 친절도에 대한 평가 점수
- `kakao_mood` : 매장 인테리어나 분위기에 대한 평가 점수
- `kakao_parking` : 주차 편의성에 대한 평가 점수

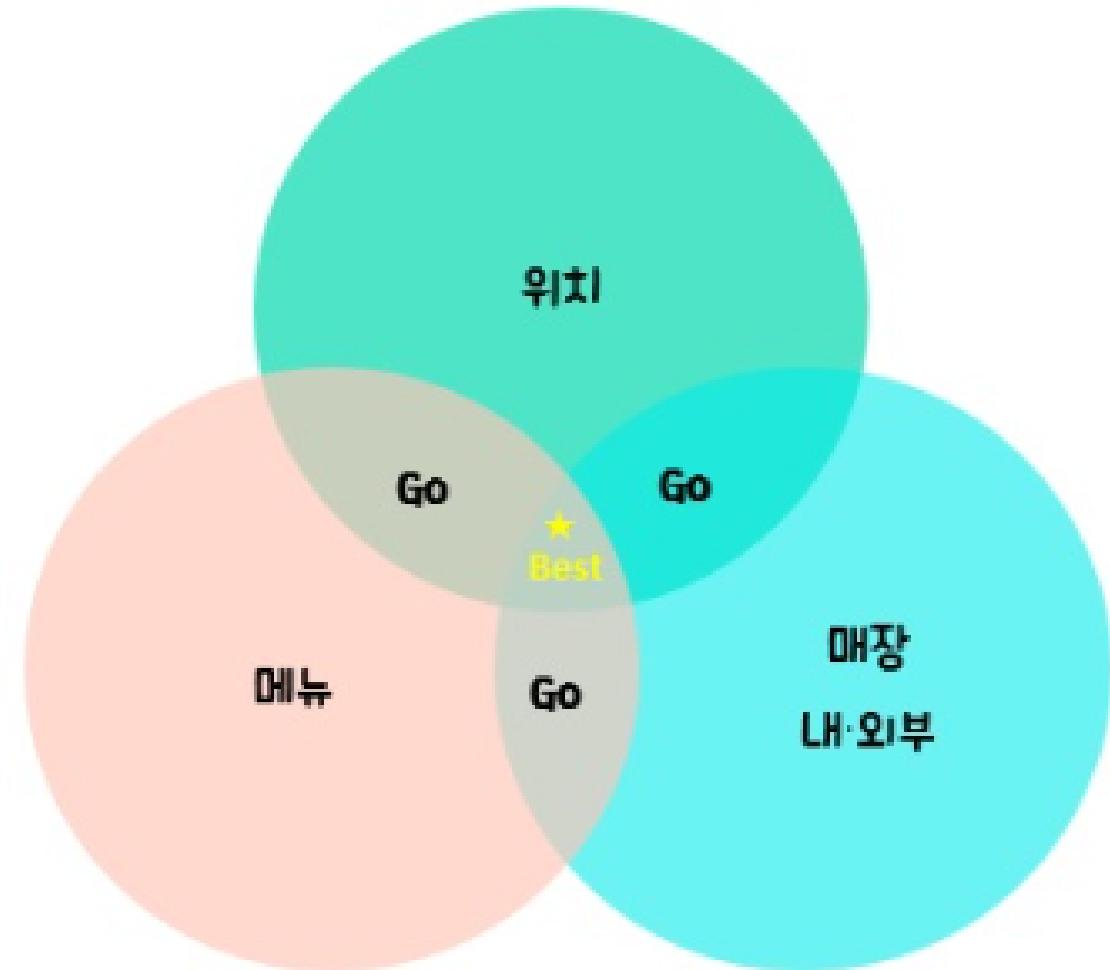
QC_Score 매칭



- 위치
 - 서울, 경기/광역시 : 1~5점으로 구분 됨.
- 메뉴
 - 맛집/놀거리/카페/술집 : 1~5점으로 구분 됨.
- 매장 내 외부
 - 인테리어 평가는 별도의 모델 개발이 필요하여, 효율성을 고려해
추후에 개발하는 방향으로 포함하지 않음.

$$(위치 + 메뉴 + 매장 내 외부 / 3) + total = QC_Score 산출$$

QC_Score 매칭



이번 분석의 범위는 **서울 지역**으로 한정했으며, LLM의 메뉴 카테고리 매칭 성능을 높이기 위해 우선 '**맛집**'을 대상으로 테스트를 진행하였고, Total 부분에 대한 점수는 가능한 요소만 추가하여 점수 산출에 사용함.

메뉴 카테고리의 판단 정확도를 위해 의미론적 해석도 가능한 LLM(Gemini)를 활용하여 판단함.

점수 산출 시 **산출 근거**를 포함하여 전달하고,
이는 '1차 필터링' 자료로 활용하여 후속 작업을 위한 가이드라인을 제시함.

QC_Score 매칭 - 위치(1)

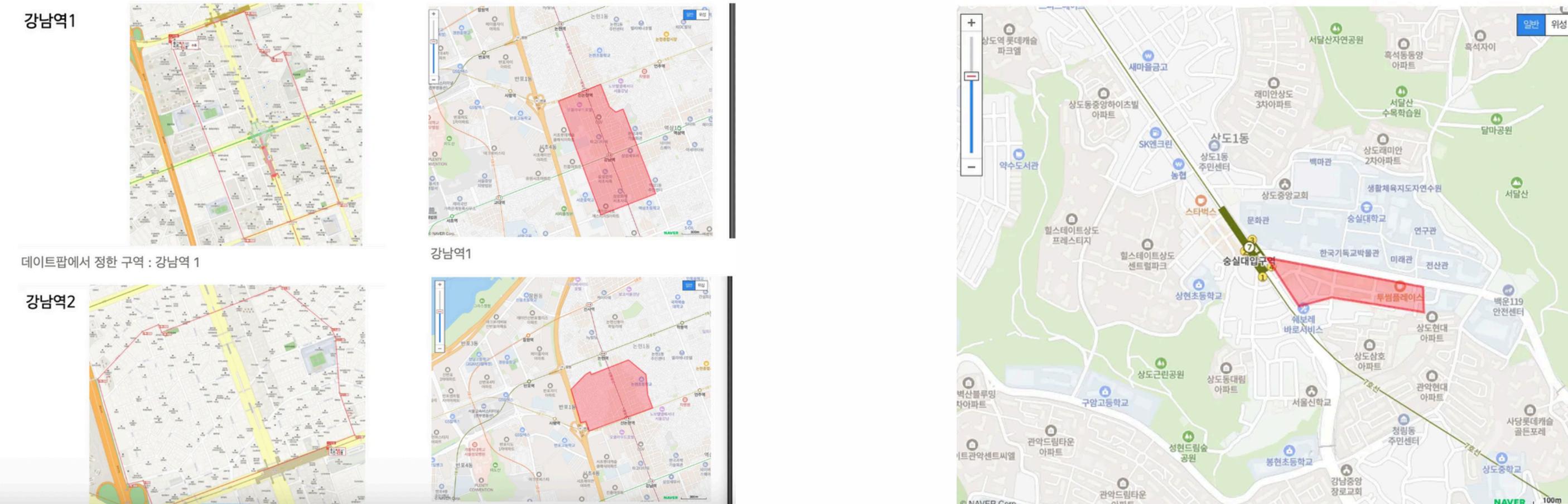
서울				
5점	4점	3점	2점	1점
앱 내 지역구로 포함되지 않은 핫플레이스 (ex.홍대-연남 등)	앱 내 지역구에 포함되어 있지만 대학가, 대형몰 입점 지역 및 신규 핫플레이스/데이트 장소	역에서 도보 15분 이내 (네이버상 카페거리, 로데오거리 가산점 0.5점 부여)	역에서 도보 25분 이내 (네이버상 카페거리, 로데오거리 가산점 0.5점 부여)	역에서 버스 환승 필요 (네이버상 카페거리, 로데오거리 가산점 0.5점 부여)

5점 : Polygon 사용 - gps_latitude, gps_longitude (포함 여부 확인)

4점 : Polygon or Keyword 사용 - gps_latitude, gps_longitude, address (포함 여부 확인)

1~3 점 : 거리 기반 - distance_from_subway (거리 기반)

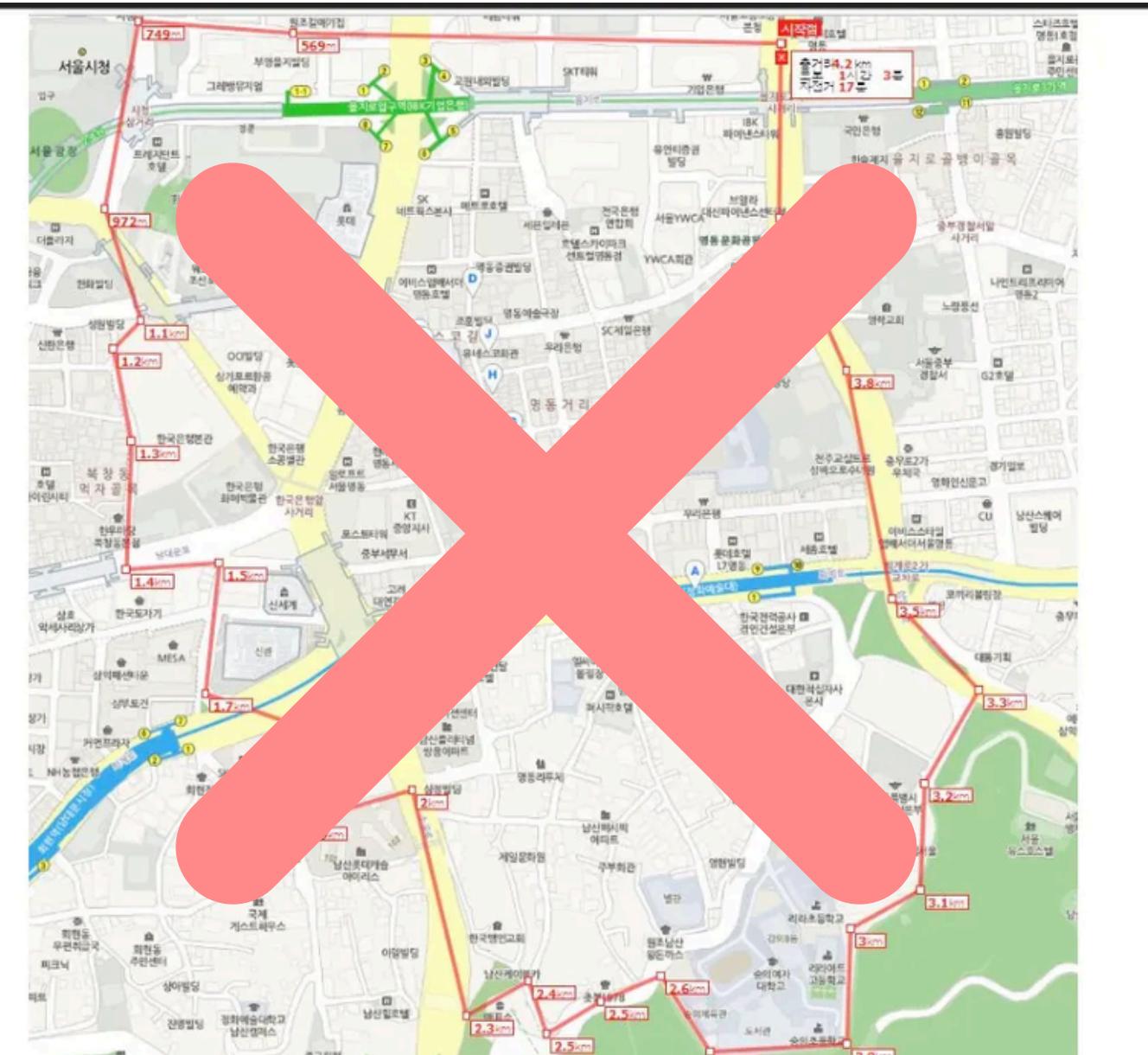
QC_Score 매칭 - 위치(2)



5점 : **Polygon 사용** - `gps_latitude, gps_longitude` (포함 여부 확인)

4점 : **Polygon or Keyword 사용** - `gps_latitude, gps_longitude, address` (포함 여부 확인)
=> 각 영역마다 Polygon을 만들어 경도, 위도 좌표를 확인하여 영역 확인

QC_Score 매칭 - 위치(2)



5점 : Polygon 사용 - gps_latitude, gps_longitude (포함 여부 확인)

4점 : Polygon or Keyword 사용 - gps_latitude, gps_longitude, address (포함 여부 확인)

=> 4점 중 코엑스, 익선동, 샤로수길 등 새로운 위치에 대한 폴리건은 존재하지 않아 키워드로 확인함.

QC_Score 매칭 - 위치(3)

점수	distance_from_subway	QC 기준표
3점	거리 900m 이하	15분 이내
2점	900m ~ 1000m	25분 이내
1점	1000~1500m	역에서 버스 환승 필요
0점	gps 없는 경우 산출 x	x

1~3 점 : 거리 기반 - **distance_from_subway** (거리 기반)

=> **distance_from_subway**로 얻은 거리 값으로 도보를 대략해서 예측하여 판단하게 함.

QC_Score 매칭 - 메뉴(1)

1. 초기 유사도 기반 매칭 - Category, Name, Theme_topic

정확 일치 매칭 (1순위) - 매장의 카테고리 정보가 score_mapping의 항목과 100% 정확하게 일치하는 경우, 최우선으로 매칭합니다.

부분 포함 매칭 (2순위) - 매장의 카테고리가 score_mapping 항목의 일부를 포함하는 경우 (예: '중식당'과 '중식'), 차순위로 매칭합니다.

유사도 기반 추론 (3순위)- theme_topic이 없을 경우: category와 name을 조합하여 RapidFuzz 라이브러리로 유사도를 계산하고, 임계값(70점) 이상일 때 매칭합니다.

다양한 유사도 평가:

- 필드 간 비교: category와 theme_topic 간의 포함 관계나 부분 유사도(partial_ratio)를 평가합니다.
- theme_topic 활용: theme_topic이 score_mapping 항목을 직접 포함하거나, RapidFuzz로 계산한 유사도가 높을 경우(70점 이상) 후보로 채택합니다.-
- 토큰 단위 비교: 단어를 조각내어 토큰 단위로 유사도를 평가(75점 이상)하여 정확도를 높입니다.

최종 후보 선택

위 과정을 통해 발굴된 후보군 중에서 점수가 가장 높은 항목을 우선 선택합니다.

만약 점수가 같다면, 유사도 점수가 더 높은 항목을 최종 선택합니다.

QC_Score 매칭 - 메뉴(1)

1. 초기 유사도 기반 매칭 - Category, Name, Theme_topic

name	category	theme_topic	matched_theme_name	matched_theme_score	matched_similarity	match_log
느삼청마당점	아이스크림	['소프트아이스크림', '베스킨라빈스', '와플']	None	NaN	NaN	모든 매칭 실패
기킹 도우	양식	['햄버거', '수제버거', '화덕피자', '피클', '초리조']	코스 양식 요리	4.0	100.0	category '양식' c '코스 양식 요리' → score_mapping 부분 매칭 (점수: 4)
면 안국점	국수	['매운맛', '비빔칼국수', '감자전', '연못', '비빔밥']	None	NaN	NaN	모든 매칭 실패
레스토랑	양식	['프랑스음식', '레스토랑', '디저트', '스타', '코스요리']	코스 양식 요리	4.0	100.0	category '양식' c '코스 양식 요리' → score_mapping 부분 매칭 (점수: 4)
회화나무	카페	['디저트', '쿠키', '케이크', '홍차']	None	NaN	NaN	❌ 카페/디저트/베이커리 category → 매칭 제외
•피방앗간	카페	NaN	None	NaN	NaN	❌ 카페/디저트/베이커리 category → 매칭 제외
차차티클럽	카페, 디저트	['홍차', '찻집', '시그니처', '녹차', '디저트']	None	NaN	NaN	❌ 카페/디저트/베이커리 category → 매칭 제외
로우루프	카페, 디저트	['파운드케이크', '스콘', '수정과', '한옥카페', '타르트']	None	NaN	NaN	❌ 카페/디저트/베이커리 category → 매칭 제외
#스테이블	스파게티, 파스타전문	['고르곤졸라', '피자', '루꼴라피자', '크림파스타', '리조또']	시카고피자	5.0	100.0	[직접 포함] ✓ '피자' c '시카고피자' → 직접 포함
베테카텐	이탈리아음식	['카르보나라', '바질', '콜키지', '아기의자', '토마토파스타']	None	NaN	NaN	모든 매칭 실패
#미술관서울점	한정식	['코스요리']	None	NaN	NaN	모든 매칭 실패
계동피자	피자	['카르보나라', '화덕피자', '화덕피자', '스타']	시카고피자	5.0	100.0	category '피자' c '시카고피자' → score_mapping 부분 매칭 (점수: 5)
조선김밥	김밥	['미술관', '어묵', '오뎅', '뚝배기', '김밥집']	해물탕 / 해물뚝배기	4.0	100.0	[직접 포함] ✓ '뚝배기' c '해물탕 / 해물뚝배기' → 직접 포함
카츠 안국점	돈가스	['모밀정식', '나베', '치즈카츠', '우동', '모짜렐라치즈']	일본 나베 요리	5.0	100.0	[직접 포함] ✓ '나베' c '일본 나베 요리' → 직접 포함
동순두부	한식	['순두부찌개', '순두부', '파전', '순두부찌개', '돌솥비빔밥']	한상차림식 한식백반	2.0	100.0	category '한식' c '한상차림식 한식백반' → score_mapping 부분 매칭 (점수: 2)
한경 그로서란트	한식	['전통주', '밀키트', '잡채', '꼬막비빔밥', '육개장']	한상차림식 한식백반	2.0	100.0	category '한식' c '한상차림식 한식백반' → score_mapping 부분 매칭 (점수: 2)
#띠끄경성	스테이크, 립	NaN	스테이크/와규 덮밥	5.0	100.0	category '스테이크' c '스테이크/와규 덮밥' → score_mapping 부분 매칭 (점수: 5)
고가	HV(BAB)	[[아조], [고리스마사], [코인], [나그나친], [와스카]]	None	NaN	NaN	모든 매칭 실패

1. "부분 포함" 매칭의 오류 가능성

- a. 비슷한 문자열 포함 여부로 매칭하기 때문에 정확도 부분에서 제대로 매칭이 어려울 수 있음.
- b. 단순 키워드 포함 여부가 아닌 해당 키워드가 얼마나 중요하고 빈도 및 의미 파악이 필요함.

2. 기존 유사도 알고리즘의 한계

- a. 단어 순서가 바뀌거나, 분리가 안된 경우, 여러 정보 조합할 때 유사도 판단 정확도 떨어짐.

2. 초기 유사도 기반 매칭 - Category, Name, Theme_topic (Tanimoto 계수 활용)

- 언급 신호 점수 (Mention Score)

- 매칭의 중요 단서에 가중치를 부여하여 합산하며, 합이 2점 이상이어야 초기 후보로 고려됩니다.
 - 테마/메뉴 일치: 가중치 2
 - 매장 이름 유사도: 가중치 1 (유사도 70점 이상일 때)
 - 카테고리 포함: 가중치 0.5

- 종합 유사도 점수 (Combined Similarity)

- 5가지 다른 유사도 지표를 가중 합산하여 계산합니다.
 - cat_sim (카테고리 유사도): 25%
 - theme_sim (테마 유사도): 25%
 - name_sim (이름 유사도): 20%
 - mix_sim (전체 조합 유사도): 20%
 - tanimoto_sim (토큰 기반 유사도): 10%

- 최종 점수 (Final Score)

- 기본 점수 \times (종합 유사도 점수 / 100) 공식을 통해, 유사도가 높을수록 기본 점수에 가깝게 조정됩니다.

최종 통과 조건

- mention_count ≥ 2
- combined_sim $\geq \text{sim_threshold}$ (기본 60)
- base_score > 0

QC_Score 매칭 - 메뉴(2)

2. 초기 유사도 기반 매칭 - Category, Name, Theme_topic (Tanimoto 계수 활용)

theme_topic	matched_theme_name	matched_theme_score	matched_similarity	matched_final_score	mention_count	match_log
아이스크림', '베스킨라빈스', '와플']	None	NaN	NaN	NaN	0.0	대상: '수제 햄버거', base:5, mention:2.0, sim:43.9, final:2.2, 임계치 미달
'화면피자', '피클', '초리조']	None	NaN	NaN	NaN	2.0	대상: '수제 햄버거', base:5, mention:0.0, sim:38.9, final:1.9, 임계치 미달
', '김자전', '연못', '비빔밥']	None	NaN	NaN	NaN	0.0	대상: '파스타 레스토랑', base:4, mention:5.0, sim:61.4, final:2.5, 통과 [{}]
'디저트', '파스타', '코스요리']	파스타 레스토랑	4.0	61.423576	2.456943	5.0	대상: '파스타 레스토랑', base:4, mention:2.0, sim:55.7, final:2.2, 임계치 미달
'루꼴라피자', '크림파스타', '리조또']	None	NaN	NaN	NaN	2.0	대상: '프랑스 코스 요리', base:5, mention:0.0, sim:30.5, final:1.5, 임계치 미달
콜카지', '야기의자', '토마토파스타']	None	NaN	NaN	NaN	0.0	대상: '프랑스 코스 요리', base:4, mention:2.0, sim:34.4, final:1.4, 임계치 미달
['코스요리']	None	NaN	NaN	NaN	2.0	대상: '화면 피자 전문점', base:4, mention:4.5, sim:66.3, final:2.7, 통과
역피자', '화면피자', '파스타']	화면 피자 전문점	4.0	66.289683	2.651587	4.5	대상: '해물탕, 해물뚝배기', base:4, mention:2.0, sim:31.7, final:1.3, 임계치 미달
'오뎅', '뚝배기', '김밥집']	None	NaN	NaN	NaN	2.0	대상: '유카쵸', base:5, mention:2.0, sim:46.7, final:2.3, 임계치 미달
'치즈카츠', '우동', '모짜렐라치즈']	None	NaN	NaN	NaN	2.0	대상: '부대찌개', base:3, mention:4.0, sim:44.4, final:1.3, 임계치 미달
한진', '순두부찌개', '돌솥비빔밥']	None	NaN	NaN	NaN	0.0	대상: '초밥 무한 리필', base:5, mention:0.0, sim:34.6, final:1.7, 임계치 미달
트', '잡채', '고막비빔밥', '육개장']	None	NaN	NaN	NaN	0.0	대상: '스테이크 전문점', base:5, mention:0.5, sim:42.9, final:2.1, 임계치 미달
과일', '시그니처', '위스키']	None	NaN	NaN	NaN	0.0	대상: '일본 나베 요리', base:5, mention:0.0, sim:24.2, final:1.2, 임계치 미달
NaN	None	NaN	NaN	NaN	0.0	대상: '프랑스 코스 요리', base:4, mention:0.0, sim:29.4, final:1.2, 임계치 미달
'초콜릿', '드림커피', '피스티치오']	None	NaN	NaN	NaN	2.0	대상: '프랑스 코스 요리', base:4, mention:2.0, sim:38.1, final:1.5, 임계치 미달
'와인', '레터링', '콜카지']	None	NaN	NaN	NaN	2.0	대상: '스테이크 전문점', base:5, mention:2.0, sim:48.3, final:2.4, 임계치 미달
NaN	None	NaN	NaN	NaN	0.0	대상: '스페인 레스토랑', base:5, mention:0.0, sim:32.5, final:1.6, 임계치 미달
-통지', '스테이크술밥', '덮밥']	None	NaN	NaN	NaN	4.0	대상: '스테이크, 와규 덮밥', base:5, mention:4.0, sim:44.1, final:2.2, 임계치 미달
, '테라스', '오리', '까르보나라']	None	NaN	NaN	NaN	2.0	대상: '오리고기 구이', base:3, mention:2.0, sim:37.2, final:1.1, 임계치 미달
NaN	None	NaN	NaN	NaN	0.0	대상: '연어 요리', base:5, mention:0.0, sim:33.1, final:1.7, 임계치 미달
'와인', '소세지', '오픈샌드위치']	None	NaN	NaN	NaN	0.0	대상: '참치 무한 리필', base:5, mention:0.0, sim:29.5, final:1.5, 임계치 미달
파스타', '닭다리살', '레드와인']	None	NaN	NaN	NaN	2.0	대상: '파스타 레스토랑', base:4, mention:2.0, sim:36.2, final:1.4, 임계치 미달
'한복', '명란떡볶이', '테라스']	None	NaN	NaN	NaN	2.0	대상: '즉석떡볶이', base:3, mention:2.0, sim:30.8, final:0.9, 임계치 미달
삼계탕집', '돌깨삼계탕', '인삼튀김']	삼계탕	3.0	91.250000	2.737500	5.5	대상: '삼계탕', base:3, mention:5.5, sim:91.2, final:2.7, 통과
정식', '바게트', '에그베네딕트']	None	NaN	NaN	NaN	2.0	대상: '일본 가정식', base:5, mention:2.0, sim:56.3, final:2.8, 임계치 미달
, '부추전', '쌈채소', '문어보쌈']	None	NaN	NaN	NaN	0.5	대상: '족발', base:3, mention:0.5, sim:46.1, final:1.4, 임계치 미달
NaN	None	NaN	NaN	NaN	0.0	대상: '프랑스 코스 요리', base:4, mention:0.0, sim:31.1, final:1.2, 임계치 미달

1. 의미는 일치할 수 있지만 키워드가 반복되지 않는 경우의 한계

- 수집하는 데이터에 따라 키워드가 반복되지 않는다면 의미론적 해석을 하기 위한 빈도수 의미 파악이 어려워짐. (베트남 음식이 확실하지만 쌀국수, 분짜, 반미 등 다양한 메뉴명으로 언급될 때 확인 어려움.)

2. Tanimoto 유사도의 낮은 영향력

- 앞서 생긴 문제를 위해 토큰으로 나눠서 사용했지만 정확도를 높이기 위한 임계점 파악도 어려웠으며 효과가 좋지 못함.

3. 하이브리드 매칭 알고리즘 - Category, Name, Theme_topic + Menu_list, Review_info 정보 추가

알고리즘 실행 순서 (Waterfall 방식)

알고리즘은 아래 3단계를 순서대로 진행하며, 중간에 하나라도 성공하면 즉시 결과를 반환하고 종료함.

- **Step 1: 룰 기반 포함 매칭 (Rule-based Matching)**
 - 핵심 키워드(라벨, 카테고리 등)가 매장 정보 텍스트에 문자 그대로 포함되면 즉시 매칭
- **Step 2: 유사도 기반 매칭 (TF-IDF & Fuzzy Matching)**
 - 1단계에서 실패 시, Fuzzy Matching으로 부분 문자열의 유사도를 평가, TF-IDF와 코사인 유사도를 통해 단어의 중요도까지 고려한 유사도를 계산함. 정해진 임계값을 넘으면 매칭에 성공
- **Step 3: 임베딩 기반 의미 유사도 매칭 (Semantic Matching)**
 - 1, 2단계에서 모두 실패 시, SentenceTransformer와 같은 언어 모델을 사용하여 매장 정보 전체를 의미를 담은 벡터로 변환. 벡터 간의 코사인 유사도를 계산하여 의미적으로 가장 가까운 항목을 찾기
- **최종 실패:** 위 3단계에서 모두 매칭에 실패하면 최종적으로 0점을 부여하고 실패 사유를 기록

QC_Score 매칭 - 메뉴(3)

3. 하이브리드 매칭 알고리즘 - Category, Name, Theme_topic + Menu_list, Review_info 정보 추가

theme_keywords	crawling_date	QC점수_매핑_label	QC점수_ncategory	QC점수_매핑_keywords	QC점수	매칭_사유	대분류	중분류	소분류
	2025-05-27	코스 양식 요리	[양식]	[코스]	4	룰 기반(포함) 매칭	양식		양식
	2025-05-27	고막비빔밥	[한식, '한정식']	[고막]	4	룰 기반(포함) 매칭	한식	국수	국수
	2025-05-27	파인다이닝 레스토랑	[이탈리아, '일식당', '양식']	[파인다이닝]	5	룰 기반(포함) 매칭	양식		양식
	2025-05-27	화덕 피자 전문점	[양식]	[화덕피자]	4	룰 기반(포함) 매칭	양식	이탈리아음식	스피게티, 파스타진
		파스타 레스토랑	[양식]	[스파게티, 파스타]	4	TF-IDF:0.59, Fuzzy:50.0	양식	이탈리아음식	
	2025-05-27	고막비빔밥	[한식, '한정식']	[고막]	4	TF-IDF:0.71, Fuzzy:100.0	한식	한정식	한정식
	2025-05-27	분식	[분식]	[순대, '튀김', '김밥', '오뎅']	1	TF-IDF:0.71, Fuzzy:100.0	분식		김밥
	2025-05-27	일식돈카츠	[돈카츠]	[카츠]	4	TF-IDF:0.51, Fuzzy:100.0	일식	돈가스	돈가스
	2025-05-27	고막비빔밥	[한식, '한정식']	[고막]	4	룰 기반(포함) 매칭	한식		한식
	2025-05-27	고막비빔밥	[한식, '한정식']	[고막]	4	룰 기반(포함) 매칭	한식		한식
	2025-05-27	파스타 레스토랑	[양식]	[스파게티, 파스타]	4	TF-IDF:0.47, Fuzzy:62.5	양식	스테이크, 립	스테이크, 립
	2025-05-27	코스 양식 요리	[양식]	[코스]	4	룰 기반(포함) 매칭	양식		양식
	2025-05-27	샐러드, 포케 전문점	[한식, '초밥, 류']	[샐러드, 포케]	4	룰 기반(포함) 매칭	한식		한식
	2025-05-27	갈비 무한 리필, 고기 무한 리필	[고기뷔페, '육류, 고기요리']	[무한리필]	5	TF-IDF:0.66, Fuzzy:100.0	한식	육류, 고기요리	육류, 고기요리
	2025-05-27	미국 기정식	[보란치]	[사미바나]	4	TF-IDF:0.46, Fuzzy:100.0	양식	보란치	보란치
	2025-05-27	곱도리탕	[요리주점, '이자카야', '닭볶음탕']	[곱도리탕]	3	TF-IDF:0.73, Fuzzy:100.0	술집	요리주점	요리주점
	2025-05-27	한상차림식 한식백반	[찌개, 전골, '한식', '생선구이', '기사식당']	[백반, '정식']	2	룰 기반(포함) 매칭	한식		한식
	2025-05-27	코스 양식 요리	[양식]	[코스]	4	룰 기반(포함) 매칭	양식		양식
	2025-05-27	파인다이닝 레스토랑	[이탈리아, '일식당', '양식']	[파인다이닝]	5	룰 기반(포함) 매칭	양식		양식
	2025-05-27	닭볶음탕	[한식, '닭요리', '백숙, 삼계탕']	[닭볶음탕, '닭도리탕']	3	룰 기반(포함) 매칭	한식	육류, 고기요리	백숙, 삼계탕
	2025-05-27	코스 양식 요리	[양식]	[코스]	4	룰 기반(포함) 매칭	양식		양식

==== 하이브리드 매칭 결과 분석 ===

전체 매장 수: 956개

매칭 성공: 622개

매칭 실패: 334개

성공률: 65.1%

==== 점수별 분포 ===

0점: 334개 (34.9%)

1점: 22개 (2.3%)

2점: 58개 (6.1%)

3점: 149개 (15.6%)

4점: 263개 (27.5%)

5점: 130개 (13.6%)

==== 매칭 방법별 분포 ===

룰 기반: 321개 (33.6%)

해당 카테고리 점수 없음: 307개 (32.1%)

TF-IDF: 299개 (31.3%)

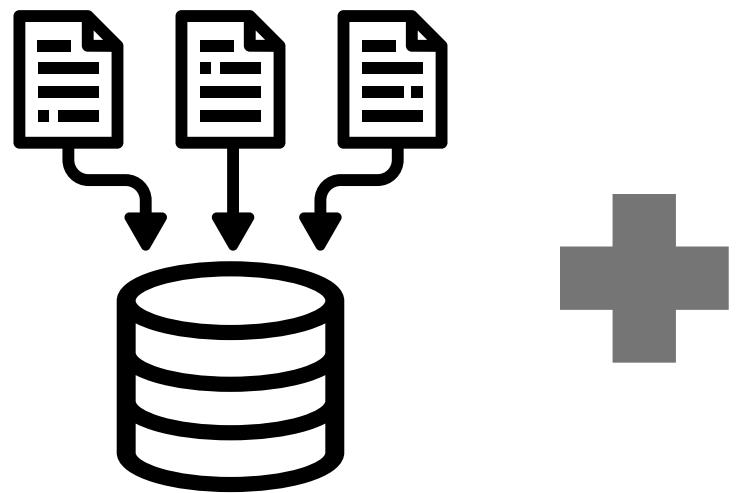
문제점: 키워드 기반 매칭의 명확한 한계

- 알고리즘을 통하여 매칭 시 복잡한 데이터의 특성을 제대로 고려하지 못하여 False Positive를 대거 발생시킴.

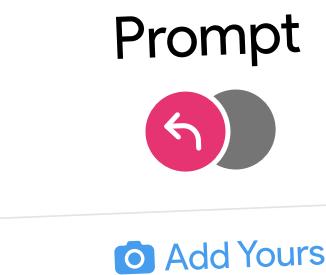
개선 방향: LLM을 통한 의미론적 분석 및 탐색 범위 제한

- LLM은 분산된 데이터로부터 복잡한 의미를 읽어내는 동시에, 명시적인 산출 근거를 함께 제공하게 하여 정해진 가이드라인에 따라 제시되는 이 근거는, '알 수 없는' AI의 판단을 '신뢰할 수 있는' 정보로 변환시키는 핵심적인 역할을 가능하게 해야함.

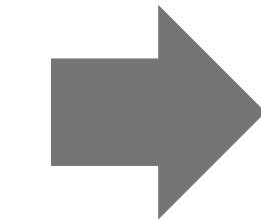
4. LLM을 통한 매칭 알고리즘(최종)



사전에 제공하는 데이터
(Polygon 좌표, 메뉴 정보)



프롬프트 가이드



```
[{"crawling_date": "2025-05-23", "theme_keywords": null, "대분류": "음식점", "중분류": "양식", "소분류": "이탈리아음식", "메뉴_라벨": "파스타 레스토랑", "메뉴_점수": 4.0, "메뉴_추론근거": "기존 카테고리가 '스파게티,파스타전문'이고, 메뉴 정보에 다양한 파스타(크림 파스타, 오일 파스타, 토마토 파스타)와 피자가 있는 점, 리뷰 요약에 '음식이 맛있어요'가 높은 비중을 차지하는 점을 고려하여 '파스타 레스토랑' 라는 추론 근거로 파악되었습니다."}]
```

프롬프트 가이드라인에 관한 공식문서를 참고하여 설계하였고, 모델의 고정된 출력과, 판단을 해야 하는 예시를 들었음.
(few shot learning의 방식을 사용함.)

QC_Score 매칭 - Total

■ Total 점수 기준표에 대해서

이때 TV 방영은 어떤 TV 프로그램을 맛집으로 할 것인지에 대한 정의가 없기 때문에 맛집 기준으로 하는 것이 아니라 TV 존재 유무로 0.3점을 부여하고 나중에 수동적으로 상), 놀거리(평균 3.0점 이상), 카페(평균 3.7점 이상), 술집(평균 3.4점 이상) 진행 (2025.04.14 개정) <확인이 가능함>

Plus	매칭 데이터 속성	Minus	매칭 데이터 속성
TV 맛집 방영매장 (+0.5)	on_tv	공연/전시 할인율 고정 (-1.0)	X (크롤링이기 때문에)
TV 기타 방영매장 (+0.3)	on_tv	시그니처아닌 메뉴 픽스 (-0.5)	X (크롤링이기 때문에)
미쉐린 가이드 등재 (+0.5)	seoul_michelin	주말권 無 (-0.5)	X (크롤링이기 때문에)
인바운드 매장 (+0.3)	X (크롤링이기 때문에)	술집 자유이용권 酒불가 (-0.3)	X (크롤링이기 때문에)
소개 받은 매장 (+0.2)	X (크롤링이기 때문에)	요식업 불필요 예약필수 (-0.3)	X (크롤링이기 때문에)
일괄 계약 매장 (+0.2)	X (크롤링이기 때문에)	최소 3일전 예약 필수 (-0.3)	X (크롤링이기 때문에)
신규 계약 시 추가 현금 광고 (+0.3)	X (크롤링이기 때문에)	오피스텔 공방 (-0.3)	X (크롤링이기 때문에)
블로그 리뷰 300개 (+0.3)	blog_review_count	앱 지도에 없는 지역 (-0.5)	X (크롤링이기 때문에)
유명인/연예인 사장 (+0.3)	X (확인할 수 있는 지표가 없음)	이용권이 1개뿐 (-0.5)	X (크롤링이기 때문에)
재계약 매장 (+0.2)	X (크롤링이기 때문에)	네이버 플레이스 미등록 (-0.3)	X (크롤링이기 때문에)
개척 카테고리 (+0.2)	잔 카페 (-0.2)	X (크롤링이기 때문에)	
부가 놀거리 매장	?	주기적 상품 변동 (-0.3)	X (크롤링이기 때문에)
핫스팟 인접 매장 (+0.5)	gps_latitude, gps_longitude (확인 가능)	8만원 초과 상품 (공방) (-0.3)	X (크롤링이기 때문에)
후속 신규 계약 (+0.3)	X (크롤링이기 때문에)	샵인샵 매장 (-0.3)	X (크롤링이기 때문에)
자체 주차장 보유 (+0.2)	parking_available	학원, 교육원, 아카데미 (-0.3)	X (크롤링이기 때문에)

- **on_tv** 가 **True** 일 경우:
 - "방송 출연 (+0.3점)" 이 추가
 - 출력 시에는 **방송 출연 (0.3점)** 형식
- **seoul_michelin** 이 **True** 일 경우:
 - "서울 미쉐린 선정 (+0.5점)" 이 추가
 - 출력 시에는 **서울 미쉐린 선정 (0.5점)** 형식
- **blog_review_count** 가 300 이상일 경우 (정수 또는 실수 타입):
 - "블로그 리뷰 300개 이상 (N개) (+0.3점)" (**N**은 실제 **blog_review_count** 값)이 추가
 - 출력 시에는 **블로그 리뷰 300개 이상 (N개) (0.3점)** 형식
- **parking_available** 이 **True** 일 경우:
 - "주차 가능 (+0.2점)" 이 추가
 - 출력 시에는 **주차 가능 (0.2점)** 형식

QC_Score 매칭시 결과

```
[{"Crawl_Date": "2025-06-23", "kakao_score": 4.0, "kakao_review": 4.0, "kakao_taste": 2, "kakao_value": 3, "kakao_kindness": 2, "kakao_mood": 2, "kakao_parking": 3, "대분류": "음식점", "중분류": "한식", "소분류": "육류, 고기요리", "메뉴_라벨": "등갈비", "메뉴_점수": 4.0, "메뉴_추론근거": "매장의 주요 메뉴는 등갈비구이, 등갈비찜, 쭈꾸미볶음 등이며, '테마 토픽'에도 등갈비구이, 등갈비찜이 나타납니다. 리뷰 상세 정보에서 위치 점수는 5.0, 위치 산출근거는 핫플레이스, 위치 실패사유는 정상적으로 작동함.", "Total_점수": 4.7, "Total_산출근거": "메뉴 점수(4.0점) + 위치 점수(5.0점) \u00f7 2 = 4.5점; 추가 점수 항목: 주차 가능(+0.2점); 총 추가 점수: 0.2점"}
```

다음과 같이 메뉴, 위치, Total 점수와 함께 산출하고 산출근거를 명시하도록 설정함.

[카테고리 정보]

- 대분류: 음식점
- 중분류: 한식
- 소분류: 육류, 고기요리

[메뉴 평가]

- 메뉴 라벨: 등갈비
- 메뉴 점수: 4.0점
- 메뉴 추론 근거:
매장의 주요 메뉴는 등갈비구이, 등갈비찜, 쭈꾸미볶음 등이며, '테마 토픽'에도 등갈비구이, 등갈비찜이 나타남.
리뷰 상세 정보에서도 등갈비찜에 대한 언급이 있음.
따라서 '육류, 고기요리' 소분류가 적절하며
'등갈비' 라벨이 가장 적합하고 점수는 4점으로 평가됨.

[위치 평가]

- 위치 점수: 5.0점
- 위치 산출 근거: 핫플레이스
- 위치 실패 사유: 정상적으로 작동함

[총점]

- Total 점수: 4.7점
- 산출 근거:
메뉴 점수(4.0) + 위치 점수(5.0) \u00f7 2 = 4.5점
추가 점수 항목: 주차 가능(+0.2점)
⇒ 최종 점수: 4.7점

Store Data Pipeline API 0.1.0 OAS 3.1

[/openapi.json](#)

default ^

POST [/pipeline/run](#) Start Pipeline Endpoint ▾

GET [/pipelines/status/{task_id}](#) Get Pipeline Status ▾

GET [/config](#) Get Config ▾

POST [/admin/consolidation](#) Trigger Consolidation Endpoint ▾

uvicorn src.api_server:app --reload 명령어를 통해서 사용하면

INFO: Uvicorn running on http://127.0.0.1:8000 (Press CTRL+C to quit) 다음과 같은 문구가 나타남.

이때 http://127.0.0.1:8000/docs 로 들어가서 사용 가능함.

API 사용법 Post(/pipeline/run)

POST /pipeline/run Start Pipeline Endpoint

파이프라인 실행을 요청하고 즉시 작업 ID를 반환합니다.

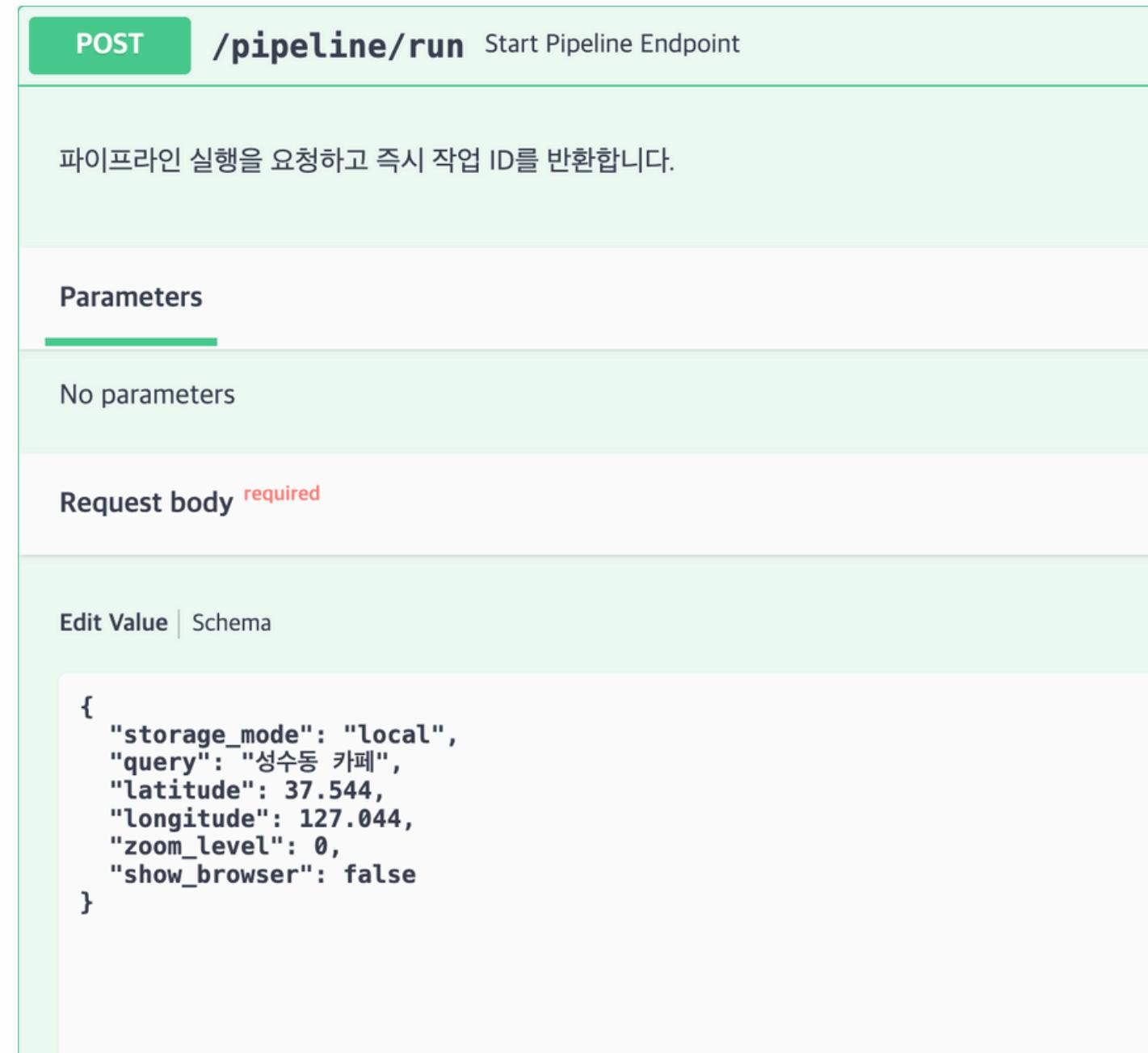
Parameters

No parameters

Request body required

Edit Value | Schema

```
{  
  "storage_mode": "local",  
  "query": "성수동 카페",  
  "latitude": 37.544,  
  "longitude": 127.044,  
  "zoom_level": 0,  
  "show_browser": false  
}
```



"storage_mode": local , s3
"query": 검색어 (필수)
"latitude": 위도
"longitude": 경도
"zoom_level": 영역 줌 확대 및 축소 값
"show_browser": false, true

자세한 내용은 config.yaml 참고

API 사용법

Get(/pipelines/status/{task_id})

Name	Description
task_id * required string (path)	cd77735c-03ba-4051-a0bf-f13d299a0e25

Execute

Responses

Curl

```
curl -X 'GET' \
  'http://127.0.0.1:8000/pipelines/status/cd77735c-03ba-4051-a0bf-f13d299a0e25' \
  -H 'accept: application/json'
```

Request URL

```
http://127.0.0.1:8000/pipelines/status/cd77735c-03ba-4051-a0bf-f13d299a0e25
```

Server response

Response body

```
{  
  "task_id": "cd77735c-03ba-4051-a0bf-f13d299a0e25",  
  "status": "processing",  
  "request_details": {  
    "storage_mode": "local",  
    "query": "성수동 카페",  
    "latitude": 37.544,  
    "longitude": 127.044,  
    "zoom_level": 0,  
    "show_browser": false  
  },  
  "progress": {  
    "네이버 크롤링": "running",  
    "카카오 크롤링": "pending",  
    "점수 산정": "pending",  
    "결과 저장": "pending"  
  },  
  "result_path": null,  
  "result_url": null,  
  "error": null  
}
```

Progress는 현재 어디 진행 중인지 확인이 가능하고, 아래 null은 result_path는 파이프라인이 성공 후 완료될 때 로컬 저장 경로를 의미하고, result_url은 s3 저장 모드일 때 다운로드 가능한 주소 반환을 한다.

error는 파이프라인 실행 중 오류 메세지를 준다.

API 사용법

Get(/config)

GET /config Get Config

서버에 로드된 전체 설정을 확인합니다.

Parameters

No parameters

Execute

Responses

Curl

```
curl -X 'GET' \
'http://127.0.0.1:8000/config' \
-H 'accept: application/json'
```

Request URL

http://127.0.0.1:8000/config

Server response

Code	Details
200	<p>Response body</p> <pre>{ "storage_mode": "local", "pipeline_stage": "full", "num_threads": 3, "headless_mode": true, "data_dir": "data", "output_format": "json", "local_config": { "output_dir": "results", "total_dir": "total", "master_file_prefix": "master_data" }, "s3_config": { "bucket_name": "your-s3-bucket-name", "output_results_prefix": "results/", "total_results_prefix": "total/", "master_file_prefix": "master_data" } }</pre>

config에서 설정한 파일의 기본 값을 확인 가능함.

API 사용법

Post(/admin/consolidation)

POST /admin/consolidation Trigger Consolidation Endpoint

데이터 통합 배치 작업을 수동으로 실행시킵니다. (관리자용)

Parameters

No parameters

Responses

Code	Description
202	Successful Response

Media type

application/json

Controls Accept header.

Example Value | Schema

```
{  
  "task_id": "string",  
  "message": "string"  
}
```

저장된 각 파일들을 병합하여 최신 파일로 만드는 API

이는 중복 수집을 막기 위해 최신 버전으로 병합하고 이를 읽게
하여 파이프라인을 실행시키기 위함