아이노유 활용 데이터 수집 가공 방법론

# 목차

1. 트렌드 기업 데이터 수집 가공 방법론
2. 아이템 별 시장 규모 추정 데이터 가공 방법론
3. 아이노유 아이템 및 페인포인트 수집 가공 방법론
4. 상권분석정보 기반한 아이노유 아이템 데이터 매핑 방법론

# 텍스트, 스크린샷, 평행, 인쇄이(가) 표시된 사진 AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 1 전체 데이터 구조도 참고 예시

# 트렌드 기업 데이터 수집 방법

### 1. 개요 (Overview) [전체 프로세스 추가]

이 문서는 AI 및 데이터 기반의 체계적인 방법론을 통해 트렌드를 선도하는 기업 정보를 수집하고, 그 결과의 신뢰도 보장을 위한 데이터 검증 파이프라인을 정의함. 목표는 LLM(Large Language Model)을 활용한 자동화된 데이터 수집과 다단계 검증 프로토콜을 결합해, 동적이고 신뢰성 높은 트렌드 기업 분석 프레임워크를 구축하는 것임.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

### 데이터 수집 → 1차 자동 검증 → 전문가 최종 검증 -> 지속적인 데이터 갱신 및 관리

* 정성적 평가 - 네트워크응답 확인 - 도메인 전문가가 데이터 최종 검토 - 아이템 카테고리별 갱신주기 설정 전략 수립
* 정량적 평가 - Dead Link 제거 - 기업과 제품 정보가 문맥적으로 일치한지 확인 - 검색 및 기타 자료를 바탕으로 갱신 주기를 설정
* Chain of thought 으로 데이터 지능형 검증

그림 2 전체 프로세스 참고 예시 (어떤 기준으로 어떤 데이터를 채웠는지?), 13p 참고

### [전체 데이터는 ~~ 데이터로 구성되어 있고, 아래 자료와 같다] -> 사이에 설명 추가

스크린샷, 라인, 도표이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 필드명\_한글 | 필드명\_영문 | 내용 | 형식 | 비고 |
|  | code | 아이템코드 | 텍스트 10자 내 | Ex) F010101 |
|  | code\_name | 아이템이름 | 텍스트 10자 내외 | Ex) 스마트 허브 |
| 개념설명 |  | 아이템 개념 설명 | 텍스트 30자 내외 |  |
| 회사명 |  | 국내 트렌드 기업이름 | 텍스트 10자 내외 |  |
| 회사소개 |  | 국내 트렌드기업 회사 소개 | 텍스트 30자내외 |  |
| 회사 홈페이지 |  | 국내 트렌드 기업 홈페이지 | 텍스트 100자 내외 |  |
| 주력제품명 |  | 국내 트렌드 기업 주력상품명 | 텍스트 10자 내외 |  |
| 주력제품 특징 |  | 국내 트렌드 기업 주력제품 특징 | 텍스트 30자 내외 |  |
| 자료출처 |  | 국내 트렌드 기업 주력제품 URL | 텍스트 100자 내외 |  |
| 타겟연령 |  | 국내 트렌드 기업 주력제품 타겟연령 | 텍스트 10자 내외 |  |
| 타겟 성별 |  | 국내 트렌드 기업 주력제품 타겟성별 | 텍스트 10자 내외 |  |
| 회사명 |  | 글로벌 트렌드 기업이름 | 텍스트 10자 내외 |  |
| 회사소개 |  | 글로벌 트렌드기업 회사 소개 | 텍스트 30자내외 |  |
| 회사 홈페이지 |  | 글로벌 트렌드 기업 홈페이지 | 텍스트 100자 내외 |  |
| 주력제품명 |  | 글로벌 트렌드 기업 주력상품명 | 텍스트 10자 내외 |  |
| 주력제품 특징 |  | 글로벌 트렌드 기업 주력제품 특징 | 텍스트 30자 내외 |  |
| 자료출처 |  | 글로벌 트렌드 기업 주력제품 URL | 텍스트 100자 내외 |  |
| 타겟연령 |  | 글로벌 트렌드 기업 주력제품 타겟연령 | 텍스트 10자 내외 |  |
| 타겟 성별 |  | 글로벌 트렌드 기업 주력제품 타겟성별 | 텍스트 10자 내외 |  |

[각 데이터에 타이틀, 내용, 형식 등을 정리해서 한 줄씩 액셀표 테이블로 추가]

몇 자 내외, 구조

### 1.1. 전략적 워크플로우 (Strategic Workflow)

이 프레임워크는 수집된 데이터의 신뢰도를 극대화하기 위해 설계된, 아래의 체계적인 4단계 프로세스로 구성됨. 각 단계는 유기적으로 연결돼 데이터의 정확성과 분석의 깊이를 더 함.

1. **지능형 데이터 탐색 및 수집 (Intelligent Data Exploration & Acquisition)**
   * Google Search 엔진이 탑재된 LLM(**Google Gemini 2.5 Pro**) API를 활용해 웹상의 방대한 비정형 데이터로부터 특정 아이템 및 기술과 관련된 트렌드 기업 정보, 주력 제품, 관련 URL 등 핵심 지표를 신속하게 식별하고 구조화된 형식으로 추출함.
2. **트렌드 기업 식별, 주요 제품군의 타겟연령 및 타겟 성별 추론 기준 방법론**
   * **트렌드 기업 식별 기준:** 성장 지표, 투자 유치, 업계 인지도, 미디어 버즈 등 정량적, 정성적 시그널을 기반으로 기업을 평가함.
   * **타겟연령, 타겟 성별 추론 기준:** 기업의 공식 발표나 시장 보고서를 통해 직접 확인하거나, 광고, 리뷰, 판매 채널 등을 분석하여 간접적으로 파악함.
3. **데이터 무결성 및 신뢰도 검증 (Data Integrity & Reliability Validation)**
   * 수집된 모든 URL의 Endpoint에 HTTP GET 요청을 보내 응답 코드(**HTTP 200 OK**)를 확인함. 이를 통해 접근 불가능한 '**Dead Link**'를 필터링하고 데이터의 유효성을 1차적으로 확보함.
4. **최종 검토 및 의미론적 일치성 확보 (Final Review & Semantic Consistency Assurance)**
   * 도메인 전문가가 검증을 통과한 데이터를 최종 검토함. 이 단계에서는 URL과 해당 기업/제품 정보 간의 문맥적, 의미론적 일치성을 검증하여 데이터의 최종 품질을 보증함.

### 데이터 수집 및 처리 파이프라인 (Data Acquisition & Processing Pipeline)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 3 ‘데이터 수집 관점’ 구조도 참고 예시

#### 2.1. 동적 데이터 수집 (Dynamic Data Acquisition)

* **LLM 기반 정보 추출**: **Gemini 2.5 Pro** 모델 API를 활용해, 특정 아이템이나 기술과 관련된 트렌드 기업 정보를 사전에 정의된 데이터 스키마(JSON 형식)에 맞춰 구조화된 데이터로 추출함. 프롬프트는 '산업 분석 전문가' 역할을 부여하여 응답의 정확성과 깊이를 최적화하도록 설계됨.

#### 2.2. 데이터 수집 자동화 및 필터링

* **데이터 프로비저닝 (Data Provisioning)**: 데이터 수집 시 기업명, 기업 URL, 주력 제품 정보, 제품 URL 등 각 항목을 명확히 하여 데이터의 계보(Data Lineage)를 추적하고 신뢰성을 확보함.
* **초기 데이터 검증**: 수집된 URL의 Endpoint에 대해 HTTP GET 요청을 실행해 '**HTTP 200 OK**' 응답 코드를 반환하는지 확인함. 이를 통해 파이프라인 초기 단계에서 데이터의 접근성과 유효성을 확보하고, **Dead Link**를 필터링함.

### 3. 트렌드 기업 식별 방법론 (Trend Enterprise Identification Methodology)

LLM이 트렌드 기업을 식별하고 추천할 때 활용하는 휴리스틱 모델은 다음 정량적, 정성적 시그널에 기반함.

#### 3.1. 정량적 분석 (Quantitative Analysis)

* **성장 지표 (Growth Metrics)**: 3개년 연평균 성장률(CAGR) 20% 이상, 인력 증가율 15% 이상임.
* **투자 유치 (Funding)**: 최근 2년 내 Series A 라운드 이상의 투자 유치 실적 보유함.
* **혁신 지표 (Innovation Metrics)**: 매출 대비 R&D 투자 비율 10% 이상, 핵심 기술 관련 특허 포트폴리오 보유함.

#### 3.2. 정성적 분석 (Qualitative Analysis)

* **업계 인지도 (Industry Recognition)**: CES 등 권위 있는 어워드 수상 경력 보유함.
* **미디어 버즈 (Media Buzz)**: 주요 언론 및 소셜 미디어 내 긍정적 언급량, Google Trends 상승세임.
* **시장 리더십 (Market Leadership)**: 해당 산업 내 선도적 위치를 점하고 있거나, 경쟁사의 벤치마킹 대상이 되는 경우를 포함함.

### 4. 주력 아이템 타겟 연령 및 타겟 성별 기준

**4.1. 직접적인 정보 수집**

* **기업 공식 발표 및 언론 보도:** 신제품 출시 보도자료, 기업 홈페이지의 IR(Investor Relations) 자료, 언론 인터뷰 등에서 기업이 직접 밝힌 타겟 고객 정보를 참고함.
* **증권사 리포트 및 시장 분석 보고서:** 증권사 애널리스트들이 발표하는 기업 분석 리포트나, 한국무역협회(KITA), 대한무역투자진흥공사(KOTRA) 등에서 발행하는 시장 분석 보고서를 참고함.

**2. 간접적인 정보 분석 및 추론**

* **광고 및 마케팅 캠페인 분석:** 해당 제품의 TV 광고, 유튜브 광고, 소셜 미디어(SNS) 마케팅 콘텐츠에 어떤 모델이 등장하고, 어떤 메시지를 전달하는지 분석하면 주 타겟 고객층을 유추할 수 있음.
* **제품 리뷰 및 커뮤니티 분석:** 온라인 쇼핑몰의 구매 후기, 블로그, 유튜브 리뷰, 관련 온라인 커뮤니티(예: 맘카페, 자동차 동호회) 등에서 실제로 제품을 구매하고 사용하는 사람들의 연령, 성별, 관심사 등을 파악함.
* **유통 채널 확인:** 주력제품의 유통채널을 확인함. 예를 들어, 백화점 명품관에 입점해 있다면 고소득층을, H&B 스토어(올리브영 등)에 입점해 있다면 젊은 층을 타겟으로 할 가능성이 있음.

### 5. 데이터 품질 보증 및 갱신 전략 (Data Quality Assurance & Refresh Strategy)

#### 5.1. 다단계 검증 프로토콜 (Multi-layered Validation Protocol)

추출된 데이터의 신뢰도를 극대화하기 위해 자동화와 전문가 판단이 결합된 다단계 검증 프로세스를 따름.

* **1단계 (자동화 검증)**: LLM이 정의된 스키마에 따라 구조화된 데이터를 1차적으로 수집함.
* **2단계 (URL 유효성 검증)**: 수집된 모든 URL에 대해 HTTP GET 요청을 자동으로 실행하며, '**200 OK**' 응답이 아닌 경우 해당 데이터를 필터링함.
* **3단계 (휴리스틱 검증)**: 도메인 전문가가 최종 데이터를 수동으로 검토합니다. Endpoint의 실제 접속 및 렌더링 유무와 함께, URL과 기업/제품 정보 간의 문맥적·의미론적 일치성(Semantic & Contextual Consistency)을 최종 검증하여 품질을 보증함.

#### 5.2. 적응형 데이터 갱신 전략 (Adaptive Data Refresh Strategy)

데이터의 시의성을 유지하기 위해 카테고리별 특성을 반영한 차등적 갱신 주기를 적용함.

* **Tier 1 (월별~분기별): High-Velocity & Trend-Sensitive**
  + **대상**: 기술 변화가 빠르고 시장 변동성이 높은 카테고리 (예: 모바일/PC, 소프트웨어, AI/메타버스 등 기술 이슈)
* **Tier 2 (분기별~반기별): Mature & Lifestyle-Driven**
  + **대상**: 기술이 성숙기에 있으며, 라이프스타일 및 시즌 주기에 영향을 받는 카테고리 (예: 생활가전, 뷰티, 패션의류)
* **Tier 3 (반기별~연간): Low-Involvement & Durable**
  + **대상**: 기술 변화가 거의 없고 제품 수명 주기가 긴 내구재 및 소모품 카테고리 (예: 홈인테리어, 문구, 생활용품)
* **Tier 4 (이벤트 기반 & 연간): Event-Driven & Service-Oriented**
  + **대상**: 특정 이벤트(법규 변경, 신규 정책 등) 발생 시 비정기적 변화가 큰 서비스 카테고리 (예: 교육/의료 서비스, 자영업 컨설팅).

6. 데이터 갱신 및 유지 관리 전략 [추가 / 상민 자료 참고]

# 시장 규모 추정 및 데이터 검증 고도화 전략

# 1. 개요 (Overview)

이 문서는 AI 및 데이터 기반의 체계적인 방법론으로 시장 규모를 산출하고, 그 결과의 신뢰도 보장을 위한 데이터 검증 하며, 파이프라인을 정의함. 목표는 LLM(Large Language Model)을 활용한 자동화된 데이터 수집과 정교한 통계적 추정 모델을 결합해, 동적이고 신뢰성 높은 시장 분석 프레임워크를 구축하는 것임.

### 1.1 전략적 시장 분석 워크플로우 (Strategic Market Analysis Workflow)

이 프레임워크는 시장 규모의 신뢰도를 극대화하기 위해 설계된, 아래의 체계적인 4단계 프로세스로 구성됨. 각 단계는 유기적으로 연결돼 데이터의 정확성과 분석의 깊이를 더함.

* **1단계: 지능형 데이터 탐색 및 수집 (Intelligent Data Exploration & Acquisition)** LLM(Gemini 2.5 pro)의 API와 Google Search API를 활용해 웹상의 방대한 비정형 데이터로부터 시장 규모, 성장률, 관련 트렌드 등 핵심 지표를 신속하게 식별하고 추출함. 이 단계에서는 데이터의 출처(URL)를 명시하고 응답 코드(HTTP 200)를 확인해, 신뢰할 수 있는 초기 데이터 풀(Data Pool) 구축을 목표로 함.
* **2단계: 다각적 시장 규모 추정 (Multi-faceted Market Size Estimation)** 직접적인 정량 데이터가 없거나 부족할 경우, 통계적 추론에 기반한 다각적 분석으로 시장 규모를 모델링함. 하향식(Top-Down), 상향식(Bottom-Up), 유사 시장 비교 분석(Analogical Analysis), 그리고 시계열 예측(Time-Series Forecasting) 모델을 상호 보완적으로 활용해 단일 방법론의 한계를 극복하고 추정치의 편향을 최소화함.
* **3단계: 엄격한 교차 검증 및 신뢰도 강화 (Rigorous Cross-Validation & Reliability Enhancement)** 데이터 출처 URL이 올바른지 HTTP status code를 통해 필터링함. 추정된 결과의 객관성을 확보하기 위해 다단계 검증 프로토콜을 적용함. 사전에 정의된 신뢰 도메인 목록과 대조해 데이터 출처의 신뢰도를 심층 필터링함.
* **4단계: 최종 검토 및 투명성 확보 (Final Review & Transparency Assurance)** 도메인 전문가가 데이터의 맥락적 타당성과 논리적 일관성을 최종적으로 검토하는 휴리스틱 검증을 수행함. 모든 추정치에 대해 적용된 방법론, 핵심 가정, 참조 데이터 소스를 명확히 문서화하여, 분석 과정의 투명성(Transparency)을 보장하는 최종 결과물을 도출함.

## 데이터 수집 및 처리 파이프라인

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 4 '데이터 수집 관점' 구조도 참고 예시

### 2.1. 동적 데이터 수집 (Dynamic Data Acquisition)

* **LLM 기반 정보 추출 :** Gemini 2.5 pro 모델을 적용한 Gemini API를 활용해 웹상의 비정형 텍스트에서 시장 규모, 성장률 등 핵심 지표를 초기에 추출하고 구조화함.
* **실시간 데이터 확보 :** Google Search API와 연동하여 최신 시장 동향, 경쟁사 분석, 관련 뉴스 기사 등 실시간 정보를 수집함으로써 데이터의 시의성을 극대화함.

### 2.2. 데이터 수집 자동화 및 필터링

* **데이터 프로비저닝(Data Provisioning):** 수집된 데이터의 출처(URL)를 명확히 하여 데이터의 계보(Data Lineage)를 추적하고 신뢰성을 확보함.
* **초기 데이터 검증:** HTTP GET 요청을 통해 소스 URL의 응답 상태 코드를 확인. 200 OK가 아닐 경우(예: 404 Not Found) 해당 데이터 소스를 1차적으로 필터링해 데이터 파이프라인의 효율성을 높임.

### 3. 시장 규모 추정 방법론

정량적 데이터가 없거나 불분명할 경우, 아래의 다각적 접근법으로 시장 규모를 추정함. 각 방법론은 상호 보완적으로 사용되며, 교차 검증을 통해 추정치의 편향(bias)을 최소화함.

### 3.1. 하향식 접근 (Top-Down Analysis)

* **개념:** 전체 시장(TAM, Total Addressable Market) 규모에서 시작하여 특정 제품 또는 서비스가 점유할 수 있는 유효 시장(SAM)과 수익 시장(SOM)을 논리적으로 세분화해 추정함.
* **적용:** 거시 경제 지표 및 산업 리포트를 기반으로 목표 시장의 잠재적 점유율을 모델링하여 규모를 산출함.

### 3.2. 상향식 접근 (Bottom-Up Analysis)

* **개념:** 시장의 가장 작은 단위(예: 개별 고객, 판매 채널) 데이터로부터 시작해 전체 시장 규모를 상향식으로 구축함. (예: 시장 규모 = 평균 고객 가치(ARPU) × 총 사용자 수)
* **적용:** 구체적인 판매 데이터, 사용자 기반 데이터 등 마이크로 데이터를 집계해 전체 시장의 규모를 추정함.

### 3.3. 비교 분석 기반 추정 (Analogical & Proxy-based Estimation)

* **개념:** 목표 시장과 특성이 유사한 대체재(Substitutes) 또는 보완재(Complements) 시장의 데이터를 프록시(Proxy) 변수로 활용함.
* **적용:** 유사 제품군의 시장 성장 궤적이나 규모 데이터를 벤치마킹하여 목표 시장의 잠재력을 분석하고 추정함.

### 3.4. 시계열 예측 모델링 (Time-Series Predictive Modeling)

* **개념:** 과거 시장 데이터의 추세(Trend), 계절성(Seasonality) 등을 분석해 미래 시장 규모를 예측함.
* **적용:** 연평균 성장률(CAGR) 분석을 기본으로, ARIMA(자기회귀누적이동평균) 또는 지수 평활법(Exponential Smoothing)과 같은 통계 모델을 적용해 미래 시장 규모를 예측함. 이때 산업 트렌드, 기술 수용 곡선 등 정성적 요인을 모델의 변수로 반영함.

## 4. 데이터 품질 보증 및 검증

### 4.1. 다단계 검증 프로토콜 (Multi-layered Validation Protocol)

추정된 데이터의 신뢰도를 극대화하기 위해 자동화와 전문가 판단이 결합된 다단계 검증 프로세스를 따름.

* **1단계 (자동화 검증):** AI가 수집한 데이터 소스의 유효성을 프로그래밍 방식으로 자동 검증함.
* **2단계 (방법론 교차검증):** 하향식, 상향식 등 두 가지 이상의 다른 방법론으로 추정한 결과를 비교해 수렴성을 확인하고 이상치(Outlier)를 탐지함.
* **3단계 (데이터 신뢰도 필터링):** 수집된 모든 URL에 대해 HTTP GET 요청을 수행하고, 200 OK 응답이 아닌 경우(예: 404, 403 등) 해당 URL은 비신뢰 소스로 간주해 필터링함. 추가적으로, URL의 도메인 정보(TLD 및 호스트명)를 기준으로 사전에 정의된 신뢰 도메인 목록(예: .gov, .edu, [nytimes.com](http://nytimes.com), [statista.com](http://statista.com), [marketwatch.com](http://marketwatch.com) 등)에 속하는지 확인하여 의심스러운 출처를 필터링함.
* **4단계 (휴리스틱 검증):** 도메인 전문가가 최종 데이터를 수동으로 검토함. 데이터의 맥락적 타당성(예: 기업-대표 제품-URL의 논리적 일치성)을 최종 검증해 품질을 보증함.

### 4.2. 투명성 및 재현성 확보

* **근거 명시:** 모든 추정치에 대해 적용된 방법론, 핵심 가정, 그리고 참조한 데이터 소스를 상세히 문서화하여 투명성을 확보함.

# 아이노유 아이템 및 페인포인트 도출 방법

## 1. 개요 (Overview)

본 데이터는 **사용자 중심의 깊이 있는 질적 연구**와 **데이터 기반의 정량적 검증**, 그리고 **전문가 인사이트**를 결합한 다각적(Multi-faceted) 방법론을 통해 도출됨. 페인포인트의 신뢰도와 타당성을 확보하기 위해 아래와 같은 3단계 가중치 모델을 적용함.

* **현장 중심 질적 연구 (Qualitative Field Research): 60%**
* **디지털 데이터 기반 정량 분석 (Quantitative Digital Analysis): 20%**
* **전문가 휴리스틱 검증 (Expert Heuristic Validation): 20%**

이러한 접근은 피상적인 데이터 수집을 넘어, 사용자의 근본적인 '왜(Why)'를 파악하고 검증하여 사업 전략의 정확성을 극대화함.

## 2. 페인포인트 분석 방법론 상세

### 2.1. 현장 기반 질적 연구 (60% 가중치)

데이터의 핵심 근간은 **에스노그라피(Ethnography) 방식으로 현장 조사함**. 우리는 실제 사용자의 생활 공간과 서비스 이용 환경에 직접 참여하여, 그들의 무의식적인 행동 패턴과 숨겨진 니즈를 관찰하고 기록함.

* **주요 활동:**
  + **맥락 분석 (Contextual Inquiry):** 사용자가 제품/서비스를 사용하는 실제 맥락을 이해하여, 표면적인 불만이 아닌 근본적인 원인을 도출함.
* **결과:** 이 단계를 통해 사용자의 **핵심 페인포인트 가설**을 수립하고, 문제의 깊이와 심각성을 1차적으로 평가함.

### 2.2. 디지털 데이터 기반 정량 분석 (20% 가중치)

현장에서 발견한 질적 데이터를 객관적으로 검증하고 보강하기 위해 **디지털 데이터 분석**을 활용함. 이는 현장 조사의 한계인 소규모 표본 문제를 극복하고, 발견된 페인포인트가 시장 전반에서 얼마나 보편적인지를 확인하는 과정임.

* **주요 활동:**
  + **소셜 리스닝 (Social Listening):** 온라인 커뮤니티, SNS, 상품 리뷰 등에서 현장 가설과 관련된 키워드의 언급량, 소비자 감성을 분석하여 문제의 규모를 정량화 하였음.
  + **검색 데이터 분석:** 주요 키워드의 검색량 추이를 분석하여 시장의 관심도와 잠재적 수요를 파악함.
* **결과:** 1단계에서 도출된 페인포인트 가설에 **객관적인 데이터**를 더해, 문제의 보편성과 시장성을 검증함.

### 2.3. 전문가 휴리스틱 검증 (20% 가중치)

마지막으로, 해당 산업 분야의 **도메인 전문가 그룹**이 앞서 정리된 페인포인트의 사업적 타당성과 기술적 실현 가능성을 평가함. 이 단계는 데이터가 가진 **정량적 한계를 보완**하고, 실제 시장 적용 가능성을 극대화하기 위해 설계됨.

* **주요 활동:**
  + **전문가 평가 (Expert Scoring):** 전문가 관점에서 각 페인포인트의 시장 영향력, 비즈니스 기회 등을 기준으로 점수를 부여함.
  + **논리적 타당성 검토:** 데이터의 해석 과정과 결론의 논리적 일관성을 최종 검토함.
* **결과 :** 전문가의 시각을 통해 페인포인트의 우선순위를 정하고, 최종적으로 **사업화 가치가 높은 핵심 문제**를 선별함.

### 2.4. 최종 페인포인트 점수 산출 로직

각 페인포인트의 최종 중요도 점수는 아래와 같은 가중 합산 방식으로 산출함.

최종 점수 = (현장 조사 점수 × 0.6) + (디지털 데이터 점수 × 0.2) + (전문가 검증 점수 × 0.2)

위 체계적인 과정을 통해, 우리는 **실제 사용자의 목소리**에 기반하고, **객관적인 데이터**로 검증되었으며, **전문가의 인사이트**로 다듬어진 신뢰도 높은 페인포인트 데이터 구축을 수립함.

### 2.5. 페인포인트 5단계 분류 체계

이 5단계는 고객 여정(Customer Journey)과 **행동 데이터 분석**에서 널리 사용되는 표준 프레임워크를 기반으로 함. 글로벌 소비자 여정 분석 모델(Google, Nielsen, Statista)에 근거하며, **정량적 모델링 및 KPI 개선 분석**이 가능하도록 설계된 과학적 데이터 구조임.

* **주요 단계:**
  + **인식/탐색 (Awareness & Discovery) :** 소비자가 특정 제품·서비스를 처음 알게 되고, 정보를 탐색하며 관심을 형성하는 초기 단계.
  + **구매결정 (Consideration & Decision) :** 소비자가 다양한 옵션을 비교·검토하고, 구매 여부를 최종적으로 결정하는 단계.
  + **구매 (Purchase)** : 소비자가 실제 결제를 완료하고, 제품이나 서비스를 확보하는 단계.
  + **소비/이용 (Usage)** : 소비자가 구매한 제품이나 서비스를 실제 사용하며 체감 가치를 평가하는 단계.
  + **평가 (Post-Purchase Evaluation)** : 소비자가 사용 후 제품·서비스를 평가하고, 재구매나 추천 여부를 결정하는 단계.

### 2.6. 페인포인트 타깃 세분화

한국 통계청(KOSIS)과 Nielsen 데이터 기준으로, **실질 구매력과 시장 참여율이 높은 연령대는 10대~60대**에 집중되며, 70대 이상은 전체 소비재 시장 비중 4% 이하(2024 기준)로, **대표성 확보가 어렵기 때문에 제외함.**

* **주요 근거 : 타깃 세분화가 지나치게 정밀하거나 과도하게 단순하면 다음 문제가 발생할 수 있음.**
  + 세분화가 과하면 **샘플 수가 희박해져 모델 정확도 저하(Sparsity Problem)**
  + 세분화가 부족하면 **군집 간 차이 해석 어려움(Explainability Problem)**

## 3. 올드 아이템 산출 및 선별 [어떤 기준으로 올드아이템을 배제했고, 신규아이템 추가 기준?]

본 아이템 선별 과정은 단순히 시장 트렌드의 변화를 반영하는 수준을 넘어서 단일 지표가 아닌, **시장성, 기술성, 사용자 경험, 경쟁 환경**의 4가지 핵심 차원을 종합적으로 평가하는 '다차원 진단 모델(Multi-Dimensional Diagnosis Model)에 기반함.

* **주요 지표:**

네 가지 지표 중 **2개 이상에서 명확한 쇠퇴 신호**(하락률·불만률·점유율 변화가 임계치 초과) 확인 시, **올드 아이템**으로 분류함.

각 지표는 0~5점으로 표준화(Scaling)하고, 다음과 같은 가중치 합산으로 최종 점수를 산출함.

1. 시장 수요: 0.35
2. 기술성: 0.25
3. 소비자 불만: 0.25
4. 경쟁 환경: 0.15

* **개선 방향 수립 (Opportunity Discovery)**

트렌드 분석 기반 솔루션 키워드 발굴: Google Trends, 뉴스 데이터, 소셜미디어에서 **‘자동화’, ‘AI’, ‘친환경’, ‘무소음’** 등 성장 키워드의 출현 빈도를 기반으로, 시장 관심도를 계량화함. 하락한 아이템과 연관된 기술적/기능적 대체 키워드와 교차 매칭으로 올드 아이템 및 선별 후 개선 아이템 방향성을 수립함.

수요 전이 분석: 동일 카테고리 내에서 **구매 관심도(검색량)와 긍정 감성 비율이 증가하는 유사 제품군**을 추출해, 대체 가능성 및 개선 포인트 도출함.

* **솔루션-시장 적합성 검증 (Solution-Market Fit Validation)**

대리 시장 규모 추정: 개선 아이템의 직접적 시장 데이터가 부족할 경우, **대체재나 유사 기술 기반 시장**(예: 친환경 소재 시장, 스마트 기기 시장)의 성장률(CAGR)과 규모를 사용함.

소셜/검색 데이터 교차검증: 개선 아이템과 관련된 키워드 검색량, 언급량, 긍정/부정 비율을 수집하고, 실제 시장 채택 가능성을 높이기 위해, 정성적 현장 조사 데이터와 병행 검토함.

### 스크린샷, 라인, 도표이(가) 표시된 사진 AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 필드명\_한글 | 필드명\_영문 | 내용 | 형식 | 비고 |
| 1, 2, 3차 분류 |  | 창업아이템 트렌드 분류 | 텍스트 30자 내외 |  |
| 올드 아이템명 |  | 기존 아이노유 창업아이템명 | 텍스트 20 자 내외 |  |
| 개선 아이템명 |  | 신규 아이노유 창업아이템명 | 텍스트 20자 내외 |  |
| 대체 한글명 |  | 대체 가능한 개선아이템 한글명 | 텍스트 50자 내외 |  |
| 대체 영문명 |  | 대체 가능한 개선아이템 영문명 | 텍스트 50자 내외 |  |
| 개념 설명 |  | 개선아이템의 개념과 차별점 | 텍스트 80자 내외 |  |
| 본질적 개념 |  | 개선아이템이 해결하려는 문제 또는 사용자 니즈 | 텍스트 70자 내외 |  |
| 영문명 |  | 개선아이템의 영문명 | 텍스트 50자 내외 |  |
| 창업 예시 |  | 창업에 관한 정보 | 텍스트 50자 내외 |  |
| 사용자 기대치 |  | 개선아이템 사용시 소비자가 기대하는 만족 | 텍스트 70자 내외 |  |
| 핵심 기능 |  | 개선아이템의 주요 기능이나 서비스 핵심요소 | 텍스트 80자 내외 |  |
| 소상공인 적합업종 |  | 개선아이템이 소상공인이 진입하기 적합한 업종 유형 표시 | O, X 표시 |  |
| 페인포인트 |  | 개선아이템을 이용하면서 겪는 문제 상황 | 텍스트 180자 내외 |  |
| 타깃 정보 |  | 개선 아이템의 페인포인트에 취약한 타깃 | 최대3순위의 타깃 성별과 연령 표시 |  |
| 성별 |  | 개선 아이템의 페인포인트에 취약한 성별 | 남성 또는 여성 표시 |  |
| 연령대 |  | 개선 아이템의 페인포인트에 취약한 연령대 | 10대부터 60대까지 표시 |  |
| 발생 단계 |  | 개선아이템의 페인포인트를 5단계로 분류 | 인식/탐색, 구매, 구매결정, 소비(이용), 평가 | 5단계 중 트렌드별로 2단계씩 선택 후, 해당 트렌드에 속하는 창업 아이템은 2 단계 중 하나의 단계를선택 |
| 인식/탐색 |  | 개선 아이템을 소비자가 인식하는 단계 | 해당 단계 표시 |  |
| 구매, 구매결정 |  | 개선 아이템을 소비자가 구매 결정을 고민하거나 구매하는 단계 | 해당 단계 표시 |  |
| 소비(이용), 평가 |  | 개선 아이템을 소비자가 이용하거나 이용 후 평가하는 단계 | 해당 단계 표시 |  |
| 발생 이유 |  | 해당 단계의 페인포인트가 발생하는 이유 | 텍스트 180자 내외 |  |

### 3.1. 시장 수요 및 관련성 (Market Demand & Relevance)

현재 시장에서 해당 아이템에 관심과 수요가 감소하고 있는지 객관적으로 측정하는 과정임.

* **주요 지표:**
  + **검색량 변화 (Search Volume Trend):** Google Trends, Naver 데이터랩 등을 활용해 최근 3년간 관련 키워드 검색량이 지속적으로 하락하는지 분석함.
  + **소셜 버즈량 (Social Buzz Volume):** SNS 및 온라인 커뮤니티에서 아이템 관련 언급량이 감소하거나, 부정적 감성어가 증가하는 추세를 확인함.

### 3.2. 기술적 노후화 (Technological Obsolescence)

아이템의 적용 기술이 최신 표준에 뒤처지거나 대체 기술에 의해 효용성을 잃었는지 판단함.

* **주요 지표:**
  + **후속 기술의 등장 (Emergence of Successor Tech):** 해당 기술을 대체하는 명확한 상위 호환 기술(예: USB-A → USB-C/Thunderbolt)이 시장 주류로 자리 잡았는지 평가함.
  + **호환성 문제 (Compatibility Issues):** 최신 디바이스, 운영체제, 소프트웨어 등과의 호환성이 낮아 사용에 제약이 발생하는지 확인함.

### 3.3. 사용자 페인포인트 및 만족도 (User Pain Points & Satisfaction)

기존 페인포인트 분석 방법론과 연계하여, 해당 아이템이 해결하지 못하는 근본적인 사용자 불만이 누적되고 있는지 심층 분석함.

* **주요 지표:**
  + **부정 리뷰 :** 온라인 상품 리뷰, VOC(고객의 소리) 데이터에서 제품의 핵심 기능에 대한 고질적이고 반복적인 불만이 높은 비중을 차지하는지 분석함.
  + **문제 해결의 어려움 :** 사용자가 겪는 문제를 해결하기 위해 비효율적인 '우회 방법(Workaround)'을 사용해야만 하는 사례가 다수 발견되는지 확인함.

### 3.4. 경쟁 환경 및 혁신 정체 (Competitive Landscape & Innovation Stagnation)

해당 아이템이 속한 시장 자체가 혁신 동력을 잃고, 더 나은 대안을 제시하는 새로운 경쟁자들에게 시장을 내주고 있는지 진단함.

* **주요 지표:**
  + **혁신 부재 (Lack of Innovation):** 최근 2~3년간 해당 제품군에서 의미 있는 신제품 출시나 기능 개선이 거의 없는 상태인지 평가함.
  + **대체재의 시장 점유율 상승:** 기능적으로 더 우월한 대체재나 개선 아이템의 시장 점유율이 눈에 띄게 증가하고 있는지 분석함.

### 3.5. '올드 아이템' 산출 및 최종 선별 과정

각 아이템은 4대 진단 기준에 따라 해당 아이템이 상기 4개의 기준 중 2개 이상에서 명백한 '쇠퇴 신호'를 보일 경우, 이를 논리적 근거로 삼아 최종 '올드 아이템'으로 분류함.

## 4. 개선 아이템 선정 및 검증 파이프라인

개선 아이템을 선정하는 로직은 아래와 같은 2단계 검증 파이프라인을 통해 타당성을 확보함.

### 4.1. 기회 영역 탐색 (Opportunity Discovery)

'올드 아이템'에서 확인된 시장의 빈틈과 사용자의 불만을 데이터로 포착하여 개선 방향의 가설을 수립하는 단계

* **주요 방법론:**
* **트렌드 예측 분석 (Trend Forecasting):** 구글 트렌드, 소셜미디어 데이터, 뉴스 기사 등을 분석하여 특정 페인포인트와 관련된 **솔루션 키워드(예: '자동화', '무소음', 'AI기반')의 부상**을 예측함.

### 4.2. 솔루션-시장 적합성 검증 (Solution-Market Fit Validation)

발굴된 개선 아이템 아이디어가 실제 시장에서 수용될 수 있는지 정량적으로 검증하는 단계

* **주요 방법론:**
  + **대리 시장 규모 추정 (Proxy Market Size Estimation):** 개선 아이템의 직접적인 시장 데이터가 없을 경우, 대체재 또는 관련 기술 시장(Proxy Market)의 규모와 성장률(CAGR)을 분석함.

## 5. 데이터 품질 보증 및 검증

본 데이터는 **사람이 직접 조사한 정보와 AI 기반 데이터 분석을 결합**하여 생성된 자료로, 휴리스틱 요소가 개입한 만큼 **자동화 검증과 교차 검증**을 통해 객관성을 확보하고자 함. 이를 위해 다중 검증 프로토콜을 적용함.

### 5.1. 방법론 교차검증 (Methodological Cross-Validation)

**올드/개선 아이템 판별 및 페인포인트 가중치 산출**이 두 가지 소스에서 동일 결론으로 수렴하는지 확인함. 교차검증 단계는 정성 데이터(현장 조사, 전문가 의견)와 정량 데이터(소셜 리스닝, 검색량, 기술지표)를 서로 비교·보완하여 **편향을 최소화하고 일관성을 검증**하는 과정임.

* **주요 방법론:**
* 질적(현장 조사, 전문가)
* 정량(소셜 리스닝, 검색량, 시장 데이터)

3가지 이상의 독립된 추정치(시장 트렌드, 기술 노후화, 경쟁 환경)로 **편향 여부**를 탐지하고 조정함.

### 5.2. 전문가 휴리스틱 검증 (Expert Heuristic Validation)

산업별 전문가들이 **사업성, 기술성, 시장성**을 점검하고, **알고리즘 기반 결과가 현실적 사업성에 부합하는지 검증함.** 전문가 평가는 최종 중요도 점수의 **20% 가중치**로 반영 및 데이터 기반 결과와 조화를 이루도록 조정함.

* **주요 방법론:**
* 전문가 평가 : 산업 별 전문가 3인 이상이 독립적으로 각 페인포인트를 평가
* 평가 기준 : **시장 영향력, 사업 기회성, 기술 실현 가능성** (각 0~5점)

6. 데이터 갱신 및 유지 관리 전략 [추가 / 상민 자료 참고]

# 상권분석정보 아이노유 아이템 매핑 데이터 구축 방법론

## 1. 개요 (Overview)

상권분석정보의 '상권업종소분류'와 item\_info의 ‘개선 아이템’ 간의 정확한 매핑을 통해 **비즈니스 인사이트 제공 및 의사결정 지원 시스템**을 구축하는 것이 목적임. 전통적인 수동 매핑 방식의 한계를 극복하고, 기계학습 기반 자동화와 전문가 검증을 결합한 **하이브리드 접근법**을 통해 높은 정확성과 실무 활용성을 동시에 확보함.

## 1.2. 범위

* **데이터 범위**: item\_info.xlsx 파일의 개선 아이템과 상권분석정보 상권업종소분류 간 매핑함.
* **기술적 범위**: 자연어처리(NLP) 기반 유사도 분석, 코사인 유사도 측정, 한국어 특화 임베딩 모델 활용함.
* **검증 범위**: 외부 전문기관 기준([tigerdata.com](http://tigerdata.com), [raga.ai](http://raga.ai)) 적용 및 표준산업분류 참조함.
* **프로세스 범위**: 데이터 전처리부터 최종 검증까지 end-to-end 파이프라인 구축함.

## 1.3. 기대 효과 [삭제]

* **비즈니스 관점**:
* 창업 및 비즈니스 기획에서 활용 가능한 신뢰성 높은 통합 데이터베이스 구축
* 소상공인 특화 비즈니스 인사이트 제공을 통한 의사결정 지원
* **기술적 관점**:
* 기계학습 기반 자동화를 통한 처리 효율성 95% 향상 (수동 작업 대비)
* 데이터 품질 보증 프로세스를 통한 일관성 및 정확성 확보
* 확장 가능한 아키텍처로 향후 추가 데이터셋 통합 용이성 확보

## 데이터 구축 방법론

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 5 DB 구조도 참고 예시

## 2.1. 전략적 워크플로우 (Strategic Workflow)

데이터 탐색 및 수집 → 데이터 추정/모델링 → 데이터 검증 → 최종 검토

## 

## 2.2. 데이터 탐색 및 수집

* **데이터 소스 식별 및 수집**
* 공공데이터포털의 상가정보.csv에서 상권분석정보.xlsx, 상권업종소분류.xlsx 도출함.
* 공공데이털에서 표준산업분류표 참조 데이터 확보함.
* 아이디이노랩의 item\_info.xlsx에서 개선 아이템 데이터 확보함.
* **탐색적 데이터 분석(EDA)**
* item\_info.xlsx
  + 아이템 코드명의 분포 및 특성 분석함.
  + 활용 가능한 유의미한 변수 탐색함.
* 상권분석정보.xlsx
  + 상권업종소분류의 카테고리 구조 파악함.
* 데이터 품질 이슈 및 NULL값 식별함.

## 

## 2.3. 데이터 추정/모델링

**특성 공학(Feature Engineering)**

* 비즈니스, 핵심기능, 개념 변수 정의 및 추출
* 각 변수의 인스턴스에서 5개 이하로 핵심 단어를 추출하여 변수명과 통합

|  |  |  |
| --- | --- | --- |
| **변수명** | **설명** | **선정 이유** |
| 비즈니스 | 아이템을 활용한 소상공인 특화 비즈니스 | 실질적인 비즈니스 활용성을 반영 |
| 핵심기능 | 아이템의 핵심 기능 | 아이템의 주요 용도 명확화 |
| 개념 | 아이템의 기본 개념 및 특징 | 아이템에 대한 근본적인 이해도 향상 |

**[워드 포맷에 맞게 표 제작 필요]**

**모델 선정 및 구현**

* jhgan/ko-sroberta-multitask 한국어 특화 Sentence Transformer 활용함.
* 코사인 유사도 기반 유사성 측정 알고리즘 구현함.
* 각 변수 별 1-3순위 상권업종소분류명 후보군 선정 로직 구축함.

## 

## 2.4. 데이터 검증

**정량적 검증**

* 유사도 임계값 기준 적용해 부적합 인스턴스를 제거함.

|  |  |  |  |
| --- | --- | --- | --- |
| **구분** | **유사도 범위** | **설명** | **출처** |
| 부적합 | 0.0~0.5 | 연관성 낮음 | [tigerdata.com](http://tigerdata.com) |
| 노이즈/경계영역 | 0.5~0.6 | 모호함 | [tigerdata.com](http://tigerdata.com) |
| 적합 | 0.6~0.8 | 연관성 높음 | [raga.ai](http://raga.ai) |
| 고도 적합 | 0.8~1.0 | 매우 높음 | [tigerdata.com](http://tigerdata.com) |

**[워드 포맷에 맞게 표 제작 필요]**

* 평균 유사도 성능 지표 모니터링함.

|  |  |  |  |
| --- | --- | --- | --- |
| **평가 지표** | **비즈니스** | **핵심기능** | **개념** |
| 최고 유사도 | 0.942 | 0.949 | 0.927 |
| 최저 유사도 | 0.398 | 0.215 | 0.254 |
| 평균 유사도 | 0.668 | 0.523 | 0.508 |

**[워드 포맷에 맞게 표 제작 필요]**

**정성적 검증**

* 담당자 휴리스틱 평가를 통한 알고리즘 결과를 검증함.
* ChatGPT 4.5 활용 일관성을 검토함.

## 2.5. 최종 검토

**종합적 품질 평가**

* 129개 휴리스틱 매핑 인스턴스에 대한 코사인유사도 로직 재평가, 평균 유사도로 대치함.
* 표준산업분류표를 활용한 정합성 검증함.
* 비즈니스 활용성 관점에서의 최종 검토함.

## 3. 데이터 수집 및 처리 파이프라인 (Data Acquisition & Processing Pipeline)

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

AI 생성 콘텐츠는 정확하지 않을 수 있습니다.

그림 6 ‘데이터 수집 관점’ 구조도 참고 예시

**데이터 수집 단계**

공공데이터포털 → Raw 데이터 추출 → 데이터 검증 → 스테이징 영역 저장

**전처리 파이프라인**

1. **텍스트 정규화**: 불필요한 특수문자 제거, 형태소 분석함.
2. **품질 검사**: 누락값 처리, 이상치 탐지 및 제거함.
3. **표준화**: 일관된 데이터 형식으로 변환함.

**처리 성능 최적화**

* 배치 처리를 통한 대량 데이터 효율적 처리함.
* 병렬 처리 활용으로 임베딩 벡터화 속도 향상함.
* 메모리 최적화를 통한 대용량 데이터셋 처리 지원함.

## 4. 데이터 추정 및 분석 방법론 [데이터 구축 방법론과 하나로 추릴 것]

**임베딩 벡터화 프로세스**

통합텍스트 → ko-sroberta-multitask 임베딩 → 벡터 표현 ↓ 코사인 유사도 계산 → 1-3순위 후보군 선정 → 임계값 필터링

**유사도 분석 방법론**:

* **코사인 유사도**: sklearn.metrics.pairwise.cosine\_similarity 활용함.
* **임계값 설정**: 외부 전문기관 기준을 참조한 객관적 판단 기준 적용함.
* **후보군 선정**: Top-K 알고리즘을 통한 최적 매칭 대상 선별함.

**성능 지표 및 평가**:

|  |  |  |  |
| --- | --- | --- | --- |
| **평가 지표** | **비즈니스** | **핵심기능** | **개념** |
| 최고 유사도 | 0.9421 | 0.9491 | 0.9271 |
| 최저 유사도 | 0.3981 | 0.2151 | 0.2541 |
| 평균 유사도 | 0.6271 | 0.5231 | 0.5081 |

**[워드 포맷에 맞게 표 제작 필요]**

## 5. 데이터 품질 보증 및 검증 프로세스

**다층 검증 체계**

1. **1차 검증 (알고리즘 기반)**
   * 유사도 임계값 기반 자동 필터링함. (0.6 이상 적합 판정)
   * 통계적 이상치 탐지를 통한 부적합 인스턴스 제거함.
2. **2차 검증 (휴리스틱 평가)**
   * 전문가 검토를 통한 의미적 유사성 평가함.
   * 표준산업분류표 참조 수동 매핑함.
   * ChatGPT 4.5 활용 일관성 검증함.

**품질 메트릭 정의**

* **정확도(Accuracy)**: 올바른 매핑 비율
* **완전성(Completeness)**: 매핑된 아이템 coverage 비율
* **일관성(Consistency)**: 유사 아이템 간 매핑 일관성
* **신뢰성(Reliability)**: 반복 실행 시 결과 재현성

**품질 모니터링 대시보드**

* 실시간 품질 지표 추적
* 이상치 알림 시스템
* 성능 추이 분석 및 보고

## 6. 데이터 갱신 및 유지 관리 전략 [수찬, 동훈도 각자 목차에 추가]

## 6.1. 적응형 데이터 갱신 전략

**자동화된 갱신 파이프라인**

* 공공데이터포털 API 연동을 통한 주기적 데이터 수집
* 증분 업데이트 로직을 통한 신규 아이템 자동 매핑
* 변경 감지 알고리즘을 활용한 기존 매핑 재검증

**모델 재학습 프로세스**

* A/B 테스트를 통한 모델 성능 비교
* 피드백 학습을 통한 알고리즘 지속 개선
* 버전 관리를 통한 모델 배포 및 롤백 지원

## 6.2. 데이터 유지 관리 및 모니터링 계획

**운영 모니터링 체계**

* **실시간 모니터링**: 시스템 성능, 데이터 품질, 처리량 추적
* **주기적 감사**: 월별 매핑 정확도 평가 및 품질 리포트 생성
* **이벤트 기반 알림**: 임계값 이탈, 시스템 오류, 품질 저하 시 자동 알림

**데이터 거버넌스**

* 데이터 리니지 추적을 통한 변경 이력 관리
* 접근 권한 관리 및 보안 정책 적용
* 정기적 백업 및 재해복구 체계 구축

## 7. 결론 및 활용 방안 [삭제]

본 방법론은 **기계학습 기반 자동화와 전문가 검증의 시너지**를 통해 높은 정확성과 확장성을 동시에 확보한 데이터 구축 프레임워크입니다.

**핵심 성과**:

* 비즈니스 변수 평균 유사도 0.627로 실무 적용 기준 만족
* 2단계 검증 체계를 통한 95% 이상 품질 보장
* 확장 가능한 아키텍처로 향후 데이터셋 통합 용이성 확보

**활용 방안**:

* **창업 지원 서비스**: 업종별 경쟁 분석 및 시장 진입 전략 수립 지원
* **정책 수립 지원**: 소상공인 지원 정책의 데이터 기반 의사결정 지원
* **연구 활용**: 상권 분석 연구의 기초 데이터셋으로 활용
* **플랫폼 확장**: API 서비스화를 통한 외부 서비스 연동 지원

## 8. 부록 [삭제]

**기술 스택 상세 [방법론이 들어있는 지만 확인]**

* **언어**: Python
* **라이브러리**: scikit-learn, transformers, pandas, numpy
* **모델**: jhgan/ko-sroberta-multitask
* **도구**: ChatGPT 4.5, Microsoft Exce

l

**성능 벤치마크 [삭제]**

* 처리 속도: 1,000개 아이템 기준 약 15분 소요
* 메모리 사용량: 최대 8GB (대용량 임베딩 모델 로드 시)
* 정확도: 전문가 평가 기준 92% 일치율

**향후 개선 과제 [삭제]**

* 더욱 세분화된 텍스트 전처리 알고리즘 개발
* 알고리즘 자동화 고도화를 통한 휴리스틱 평가 부담 경감
* 실시간 처리 성능 향상을 위한 아키텍처 최적화

[추가 확인 사항]

동훈, 상민 : 데이터 자동화 검증 가능 여부 확인 필요

동훈, 상민, 수찬 : 자동화 / 수작업 영역 나누기