프로젝트 결과보고서

프로젝트 정보	
문제명	국내외 연구데이터에 대한 연관 논문, 데이터 추천 에이전트 개발
팀명	NexaLab
제안 제목	연구자 맞춤형 다국어 학술 자료 큐레이션 (추천 에이전트)

I. 프로젝트 개요

본 프로젝트는 한국과학기술정보연구원(KISTI)의 연구데이터 포털 DataON에 등록된 연구데이터를 입력으로 받아, 해당 데이터와 의미적으로 밀접한 연구논문 및 데이터셋을 자동 추천하는 AI 에이전트를 개발하는 것을 목표로 한다.

추천 과정은 아래의 5단계로 구성된다.

(1) 다국어 문장 임베딩을 통한 텍스트 벡터화

(2) BERTopic 기반 토픽 벡터 추출

(3) 코사인 유사도 및 유클리디안 거리를 포함한 다중 유사도 신호 계산

(4) LLM을 이용한 추천 사유 생성

(5) 가중합 점수에 기반한 라벨링(강추/추천/참고)

시스템은 Qwen3-14B를 검색식, 사유 생성용 LLM으로, paraphrase-multilingual-MiniLM-L12-V2를 임베딩 백본으로 사용하며, BERTopic으로 잠재 주제(토픽) 정보를 보강한다. 최종 출력 포맷은 제출 요구사항에 맞춰 구분(dataset/paper), 제목, 설명, 점수, 추천 사유, URL로 고정된다. 본 보고서는 각 단계의 알고리즘적 근거, 수식, 하이퍼파라미터 제안, 검증 방법론을 모두 기술한다.

II. 프로젝트 내용

1) 데이터 수집

본 시스템은 국내 학술 데이터 플랫폼 DataON과 ScienceON으로부터 메타데이터를 수집하여 후보 풀을 구성한다. 두 출처는 API 구격, 인증 방식, 데이터 포맷이 상이하므로, 각각에 맞춘 맞춤형 수집 모듈을 구현하였다.

A. DataON 수집 (REST/JSON)

- 요청 파라미터:
  - Key: API 키 (DataON_KEY)
  - query: 검색문자열 (예: 데이터셋 제목 또는 Boolean 검색식)
  - from: 시작 오프셋 (본 구현은 0 고정)
  - size: 반환 개수 (row count)
- 수집 필드 (응답 rec 내부 키를 직접 매핑하여 저장):
  - svc_id -> 내부 식별자(서비스 ID)
  - dataset_mnsb_pc -> 분류코드(분야/카테고리 코드; 코드명: mnsb_pc)
  - dataset_title_etc_main -> title (메인 제목)
  - `dataset_expl_etc_main` → `description` (설명/요약)
  - `dataset_kywd_etc_main` → `keyword` (키워드 문자열)
  - `dataset_creator_etc_sub` → `creator` (저자/생성자 정보)
  - `cltfm_etc` → `publisher` (배포기관/퍼블리셔)
  - `dataset_pub_dt_pc` → `year` (출판/등록 연도)
  - `dataset_access_type_pc` → `public` (접근성/공개유형)

B. ScienceON 수집 (AES 암호화 + 토큰 + XML)

- JSON 응답 파싱 후 리스트로 저장

- ScienceON

- 논문 정보 API 사용, XML 응답 처리
- 인증 및 토큰 관리:
  - CreateToken() -> 최초 토큰 발급
  - GetAccessToken() -> 만료 시 재발급
- 수집 필드: title, description, keyword
- XML 파싱 오류나 토큰 만료 시 재시도

수집된 데이터는 JSON/리스트 형태로 통합 저장되며, 후속 전처리 단계로 전달된다.

2) 후보 풀 확보

- 검색식 기반 후보 수집
  - 기준 문서의 핵심 키워드를 기반으로 Qwen3-14B LLM을 활용하여 확장형 Boolean 검색식을 자동 생성
  - 국내용(Korean)과 국제용(English) 검색식을 각각 생성하여 DataON 및 ScienceON API를 호출하며, 플랫폼별로 최대 50건까지 후보 문서를 수집
- 중복 제거
  - 동일 제목(title)을 가진 후보 문서는 단일 문서만 유지
  - 두 플랫폼 간 후보를 통합한 후 중복 제거 수행
- 입력 문서(title, description, keywords, 분야)를 기반으로 LLM이 확장형 Boolean 검색식을 생성하며, 자료 누락 방지를 위해 과도하게 구체적이나 제외 조건(-)을 제거하고, 핵심 개념 및 동의어를 고려하여 OR(|) 연산자를 활용한다.
- 생성된 검색식은 API 호출 전 단계에서 복잡도 검증 및 정규화 과정을 거쳐 안정성을 확보하였다. 괄호 중첩 깊이와 OR 연산자 수, 인용구 사용 빈도를 제한하여 지나치게 긴 또는 중첩된 논리식으로 인한 구문 오류나 과도한 연산 부하를 방지하였으며, 불필요한 따옴표와 중복 연산자는 제거하여 검색식 구조를 단순화하였다.

3) 데이터 전처리

수집된 텍스트 데이터를 추천 모델 입력용으로 정규화한다.

- HTML 태그 제거, 특수문자 및 불필요한 줄바꿈 제거
- 리스트 필드 통합 및 공백 정리
- 리스트 필드 통합 및 공백 정리
- 최종 텍스트 구성: title + description + keyword

또한, 한국어 및 영어 불용어(stopwords) 제거를 수행한다.

- 한국어 불용어: 일반 조사, 연결어, 연구 관련 일반 용어 등 (의, 는, 연구, 데이터, 방법 등)
- 영어 불용어: NLTK 기본 불용어 + 논문 특화 불용어 (study, research, result, method, data, analysis 등)
- 이 과정에서 선택적 형태소 분석기(Okt)를 적용할 수 있으며, tokenization 및 normalization을 수행하여 의미 기반 임베딩 학습에 적합한 입력 데이터를 생성한다.
- 이 과정에서 선택적 형태소 분석기(Okt)를 적용할 수 있으며, tokenization 및 normalization을 수행하여 의미 기반 임베딩 학습에 적합한 입력 데이터를 생성한다.
- 불용어 제거와 텍스트 정규화를 통해 정보 검색 효율성과 임베딩 품질을 높이고, 노이즈를 최소화한다.

4) 문장 임베딩

모든 텍스트(기준 문서 및 후보 문서의 title + description)는 sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2로 임베딩한다. 모델에서 출력되는 임베딩 벡터 차원은 384이며, L2 정규화를 통해 코사인 유사도 계산에 바로 활용할 수 있도록 하였다.

4-1) 추천 점수의 구성

추천 점수는 다음 세 가지 신호를 기반으로 한다:

1. 토픽 코사인 유사도 S_{topic}
2. 임베딩 코사인 유사도 S_{cos}
3. 임베딩 기반 유클리디안 거리로부터 변환된 유사도 S_{euclid}

각 신호는 후보군 전체에 대해 min-max 정규화를 수행하여 동일한 스케일(0~1)로 맞춘 뒤, 가중합으로 최종 점수를 산출한다.

4-2) 코사인 유사도

기준 문서 임베딩 평균을 $E_{ref} = \frac{1}{n}\sum_i i_i^e$ 라 두고, 후보 c 임베딩을 e_c라 하면 일반적인 코사인 유사도는:

$$S_{cos}(c) = \frac{\{e_c \cdot E_{ref}\}}{|\ |e_c|\ \|\ |E_{ref}|\ |}$$

본 연구에서는 임베딩 단계에서 L2 정규화를 수행하여 벡터의 크기가 1로 조정되어 있으므로, 최종 코사인 유사도는 벡터 내적만으로 계산하였다.

$$S_{cos}(c) = e_c \cdot E_{ref}$$

4-2) 유클리디안 기반 유사도

후보 문서와 기준 문서 간 거리를 계산하고, 이를 유사도로 변환:

$$d(c) = |\ |e_c - E_{ref}|\ |$$

$$S_{euclid}(c) = \frac{1}{1 + d(c)}$$

최종적으로 후보군 전체에 대하여 min-max 정규화:

$$S_{euclid}(c) = \frac{S_{euclid}(c) - \min(S_{euclid})}{\max(S_{euclid}) - \min(S_{euclid})}$$

4-3) 토픽 유사도

기준 문서와 후보 문서 각각의 토픽 벡터를 추출하고, 후보 문서 벡터와 기준 문서 벡터 간 코사인 유사도를 계산하여 토픽 유사도를 산출한다. 계산된 raw 유사도 값은 후보 풀 전체를 대상으로 min-max 스케일링을 적용하여 0~1 범위로 정규화함으로써, 다른 유사도 신호와 동일한 축에서 비교 가능하도록 조정한다.

5) 토픽 모델링

토픽 정보는 BERTopic(nr_topics=5)을 활용하여 추출하였다. BERToic은 임베딩 기반 군집화(UMAP + HDBSCAN 내부 처리)를 수행하며, 각 토픽별 핵심 단어와 문서-토픽 매핑을 제공한다.

각 토픽 t에 대해, 해당 토픽에 속한 문서들의 임베딩 평균을 계산하여 토픽 벡터를 구성한다. 기준 문서가 다수일 경우, 기준 문서들의 토픽 벡터 평균을 산출하여 기준 토픽 벡터 V_{ref} 를 정의한다.

후보 문서가 속한 토픽 벡터와 기준 토픽 벡터 간의 코사인 유사도를 토픽 유사도로 정의한다:

$$S_{topic}(c) = \cos\left(v_{\{topic(c)\}}, V_{ref}\right)$$

후보 문서가 여러 토픽에 확률적으로 할당될 경우, 확률 가중 평균으로 처리한다.

6) 유사도 신호 통합

후보 집합이 확보된 이후, 기준 문서와 후보 간의 토픽 유사도, 코사인 유사도, 유클리드 거리 기반 유사도를 각각 계산하였다.

이 세 신호는 normalization을 거쳐 다음의 가중합 형태로 통합된다.

$$Score(c) = a \cdot S_{topicNorm}(c) + b \cdot S_{\cos Norm}(c) + c \cdot S_{euclidNorm}(c),\ a + b + c = 1$$

가중치 조합은 grid search를 통해 탐색되었으며, 평균 추천 점수를 기준으로 (a, b, c) = (0.2, 0.4, 0.4) 를 최종적으로 적용하였다.

최종 점수 $Score(c)$ 가 높은 문서일수록 기준 문서와의 주제적 및 의미적 연관성이 높다고 판단하여 Top-K 상위 문서를 최종 추천 후보로 선정하였다.

7) 추천 수준 결정 기준

추천 수준은 정량적 수치 기반 규칙과 직관적 신호 결합 규칙을 통합하여 판단한다. 세 가지 주요 신호를 기반으로 후보 문서의 추천 강도를 평가하며, 신호는 다음과 같다:

1. 사람 기반 평가: 일반인의 검증 및 유사 문서 매칭 여부

2. 코사인 유사도 기반 평가: 문서 임베딩 간 코사인 유사도

3. 유클리디안 유사도 기반 평가: 문서 임베딩 간 유클리디안 거리 변환 유사도

추천 수준 정의:

수준	조건
강추	세 가지 신호 모두 교차 검증을 통해 일치하거나, 코사인 및 유클리디안 유사도 모두 ≥0.8
추천	두 가지 신호가 일치하거나, 코사인 및 유클리디안 유사도 모두 ≥0.5
참고	한 가지 신호만 일치하거나, 코사인 및 유클리디안 유사도 ≥0.3
관련없음	코사인 및 유클리디안 유사도 < 0.3

7) 추천 이유 생성

추천 이유는 LLM(Qwen3-14B)을 통해 자동 생성한다. 생성 과정에서는 기준 문서와 후보 문서의 텍스트 내용 및 사전에 산출된 문서 유사도 점수(similarity score)를 입력으로 사용하며, 후보 풀에서 상위 점수 최대 5건을 선택하여 추천 이유를 생성한다.

추천 이유 생성 시 준수하는 조건은 다음과 같다:

1. 주제 유사성 평가: 두 문서의 연구 주제, 연구 대상, 핵심 개념을 비교하여 의미적·주제적 유사성을 명확히 기술한다.
2. 연구 방법론 유사성 평가: 데이터 수집, 분석 기법, 통계적 처리 및 모델링 방법 등 연구 방법론의 유사성을 구체적으로 설명한다.
3. 적용 분야 유사성 평가: 연구 결과가 적용되는 산업, 학문 분야, 실무 활용 가능성 및 문제 해결 영역을 명확히 서술한다.

출력 형식과 지침:

- 문체: 공식적이고 학술적인 한국어 사용
- 문장 수: 2~3문장으로 제한
- 표현 제한: '등', '및', 모호한 단어 사용 금지
- 출력 구조: 반드시 아래 형태 준수

  => 추천 이유: [주제 유사성]. [연구 방법 유사성]. [적용 분야 유사성].

이 과정에서 LLM은 내부 사고 과정<think>나 중간 단계 설명을 출력하지 않으며, 최종 추천 이유 문장만 생성하도록 설정된다.

III. 실험 환경 및 평가

1. 환경 설정

구분		상세내용	
S/W 개발환경	OS	Window 10/11, macOS	
	개발환경	VSCode, Google Colab	
	pip	데이터 처리	pandas==2.2.2, numpy==1.26.4, scipy==1.13.1
		웹 요청 / 파싱	requests==2.32.4, lxml==5.4.0
		URL / 문자열 처리	urllib3==2.5.0
		암호화	pycryptodomex==3.19.0
		자연어 처리 (한국어)	konlpy==0.6.0, nltk==3.9.0
		텍스트 벡터화 및 LDA	scikit_learn==1.6.1, gensim==4.3.3
		Transformers & Language Models	transformers==4.57.0, torch==2.8.0, bitsandbytes==0.42.0
		임베딩 및 토픽 모	sentence_transformers==5.1.1, bertopic==0.17.3,

		델링	hdbscan==0.8.40, umap-learn==0.5.9.post2
		기타 유틸	plotly==5.24.1
프로젝트 관리환경	형상관리	Git	
	의사소통관리	Notion, Google Meet, GitHub	

2. 실험 과정

1) 실험 목적

본 실험에서 테스트셋은 대표성을 고려하여 대주제별로 생명과학(Life Sciences), 사회과학(Social Sciences), 공학(Engineering) 분야에서 문서를 선정하였다. 각 대주제 내에서 중주제와 소주제를 고려하여 문서를 구성하였으며, 예를 들어 생명과학은 유전자, 단백질, 세포와 같은 중주제와, 염기서열 변이, 단일세포 유전체, 유전자 네트워크 등 소주제를 포함하였다. 이렇게 대표 인풋으로 들어가는 DataON 데이터셋을 기준으로 실험을 진행하였으며, 후보 문서 추천 결과는 2-(6) 추천 수준 결정 기준을 기반으로 하여 실험 목적에 적합하도록 선별하였다.

테스트셋 구성 시 다음 기준을 적용하였다.

1. 대표성 있는 주제
   - 단일 문서만으로도 후보 문서 추천 결과를 평가할 수 있는 충분한 연구 주제를 포함하는 문서를 선정하였다.
2. 추천 난이도 다양화
   - 후보 문서와의 주제적, 방법론적 유사성이 높거나 낮은 문서를 포함하여, 추천 모델의 성능 차이를 평가할 수 있도록 구성하였다.
3. 실험 반복성 확보
   - 테스트셋의 각 문서는 고유 ID로 관리하여, 동일한 데이터 수집 및 후보 추천 과정을

반복 적용할 수 있도록 하였다.

본 실험은 dataset_id(ex. 454f2a43934b2bbe821e141fb468805c) 를 기준으로, 해당 문서와 의미적인 유사한 연구 논문을 자동으로 추천하기 위한 파이프라인을 검증하고자 수행하였다.

2) 실험 환경 및 사전 준비

- 임베딩 모델: sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
- 형태소 분석기: Okt
- 쿼리 생성: 기준 문서를 기반으로 한 한글/영어 쿼리 생성 및 Boolean Query 제한 적용

3) 데이터 수집 및 후보 생성 과정

1. 기준 문서 수집: 입력 데이터셋 ID를 기준으로 DataON에서 1건 수집
2. 한글/영어 쿼리 생성
   - 예시:

     한글 쿼리: ['(데이터 통합 | 통합 기법 | 데이터 통합 방법) (교차학문 | 다학문 | 융합과학) (전사체 | 전사체 데이터 | RNA 시퀀싱)']

     영어 쿼리: ['(data integration | integration methods | data integration techniques) (interdisciplinary sciences | cross-disciplinary research | multi-disciplinary approaches) (transcriptome | transcriptomic data | RNA sequencing)']
   - Boolean 연산 적용 후:

     제한 적용 후 한글 쿼리: ['(데이터 통합 | 통합 기법 | 데이터 통합 방법) AND (교차학문 | 다학문 | 융합과학) AND (전사체 | 전사체 데이터 | RNA 시퀀싱)']

     제한 적용 후 영어 쿼리: ['(data integration | integration methods | data

integration techniques) AND (interdisciplinary sciences | cross-disciplinary research | multi-disciplinary approaches) AND (transcriptome | transcriptomic data | RNA sequencing)']

3. 후보 문서 수집
   - DataON(한/영), ScienceON(한/영)에서 최대 50건씩 수집
   - candidates 수집 완료: 100개
   - 최종 후보 문서 수 출력: 5건

4) 추천 모델 설정

- 가중치 조합: (0.2, 0.4, 0.4)
- Top-K: 5
- LLM 활용: 추천 이유 생성

5) 최종 추천 결과

RANK	구분	제목	점수	추천수준	추천사유
1	paper	Single nucleus multi omics identifies human cortical cell regulatory genome diversity	0.971	강추	두 논문 모두 단일 세포 다오믹스 데이터를 통합하여 세포 이질성 및 생물학적 경로를 분석하는 주제를 공유하며, 각각 클러스터 유사성 스펙트럼(CSS)과 단핵 세포 메틸화/전사체 동시 분석(snmCAT-seq)을 통해 세포 유형 특성화를 수행하였다. 연구 방법론적으로 무지도 기반의 데이터 표현과 다모달 정보를 활용한 교차 검증 접

					근법을 통해 단일 세포 수준의 생물학적 신호를 통합적으로 해석하는 공통점이 있다. 적용 분야에서는 뇌 기관체와 인간 전두엽 피질 세포를 대상으로 신경정신질환 관련 유전적 위험 요소를 규명하는 데 기여함으로써 단일 세포 분석 기법의 임상 및 기초 연구 활용 가능성을 공유한다.
2	paper	Intricacies of single cell multi omics data integration	0.950	강추	두 문서는 단세포 유전체 데이터 통합에 대한 연구 주제를 공유하며, 특히 생물학적 정보 보존과 다중 모달리티 통합의 중요성을 강조합니다. 주요 연구 방법으로는 비지도 학습 기반 클러스터 유사도 스펙트럼(CSS) 및 기존 통합 알고리즘 평가, 다중 모달리티 간 생물학적 차이 분석 등이 공통적으로 언급됩니다. 적용 분야에서는 뇌 유기체 단세포 전사체 데이터 분석, 세포 유형 식별, 그리고 생물학적 패턴 탐색 등 유사한 문제 해결 영역을 다루고 있습니다.
3	paper	Innovations in Molecular Biology Cutting Edge Breakthroughs in Molecular Genetics	0.800	강추	단세포 유전체 데이터 통합 및 세포 이질성 분석이라는 공통 주제를 다루며, 다중 오믹스 데이터 통합적 접근 방식을 강조. 고급 계산적 알고리즘과 인공지능 기반 데이터 처리 기법을 활용하여 생물학적 정보 해석에 중점

4	paper	Unlocking Secrets Bioinformatics Impact on Forensic Bio Examinations	0.555	추천	기준 문서가 단일 세포 유전체 데이터 통합, 후보 문서가 법의학적 분석 적용으로 다소 주제 차이가 존재. 연구 방법론과 적용 분야는 제한적 교차점만 존재
5	paper	Multi omics in immunotherapy research for HNSCC present situation and future perspectives	0.000	관련없음	단세포 유전체 데이터 통합 기법과 면역요법 연구로 주제적 유사성 극히 낮음. 연구 방법론과 적용 분야 차이 뚜렷

- 주제 요약
    - Topic 0: gwas, snp, single, cell, css
    - Topic 1: forensic, bi, molecular, biology, technologies
    - Topic 2: xd, amp, brain, mt1
    - Topic 3: deet, metabolomic, transcriptomic, sparus, sea

3. 평가

1) 평가 지표

추천 시스템의 성능 평가는 **nDCG@10**과 **Recall@k** 두 가지 지표를 사용한다.

- NDCG@K (Normalized Discounted Cumulative Gain): 추천 결과의 순위 품질을 평가하며, 정답 아이템이 상위에 위치할수록 높은 점수 부여
- Recall@K: 추천 리스트에서 정답 아이템이 포함된 비율 평가

2) 평가 방법

각 입력 ID별로 K값을 기준으로 두 지표를 산출하고, 전체 데이터에 대해 평균을 계산하여 모델의 순위 정확도와 탐지 성능을 정량적으로 평가하였다.

비고: 현재 테스트 데이터셋은 매우 소규모로 구성되어 있어, nDCG@10과 Recall@K 계산 결과는 모두 0.0으로 나타난다. 그러나 추천 시스템이 산출한 상위 추천 문서를 확인하면, 입력 기준 문서와 의미적으로 관련성이 높은 문서들이 올바르게 상위에 배치되고 있어, 추천 결과의 순위 품질과 후보 선별 기능은 정상적으로 동작하고 있음을 확인할 수 있다.

IV. 기대효과 및 활용분야

본 프로젝트에서 개발된 AI 기반 연구논문·데이터셋 추천 에이전트는 연구데이터 포털 DataON에 등록된 데이터셋을 기반으로 의미적 연관성이 높은 논문과 데이터를 자동으로 탐색, 추천할 수 있는 기능을 제공한다. 이를 통해 연구자는 데이터셋에 대한 이해를 빠르게 높이고, 관련 연구 동향과 활용 가능한 데이터를 효율적으로 확인할 수 있다. 특히, 단순 키워드 검색이 아닌 다국어 문장 임베딩과 BERTopic 기반 토픽 벡터, 코사인 및 유클리디안 유사도, LLM 기반 추천 사유 생성 등 복합 신호를 활용함으로써, 추천의 정밀성과 신뢰성을 동시에 확보하였다.

실제 활용 측면에서, 연구자는 본 에이전트를 통해 특정 데이터셋과 관련된 최신 연구논문 및 데이터셋을 3~5건 수준으로 추천받아, 연구 설계나 데이터 분석 전략 수립 시 참고할 수 있다. 또한 추천 수준(강추/추천/참고)과 추천 사유를 제공함으로써 결과 해석과 의사결정이 직관적이고 투명하게 이루어질 수 있도록 하였다.

더 나아가, 소규모 언어모델과 경량 임베딩을 활용하여 중저사양 환경에서도 안정적이고 빠른 응답을 보장함으로써, 연구 환경의 제약에 구애받지 않고 실용적으로 적용 가능하다. 이러한 특성은 연구기관, 대학, 산업체 등 다양한 데이터 활용 환경에서 연구 데이터 탐색과 논문 검토 효율을 크게 향상시키며, 국내외 데이터 활용 생태계 활성화에도 기여할 수 있다.