

**SK네트웍스 Family AI과정 3기  
 모델링 및 평가 수집된 데이터 및 전처리 문서**



**□ 개요**

* 산출물 단계 : 모델링 및 평가
* 평가 산출물 : 수집된 데이터 및 전처리 문서
* 제출 일자 : 2024.12.23 (월)
* 깃허브 경로 : <https://github.com/SKNETWORKS-FAMILY-AICAMP/SKN03-FINAL-1Team>
* 작성 팀원 : 서민정

1. **논문 요약 목적 데이터 전처리**

| **개요** | **논문 요약 기능 목적:** 사용자가 특정 논문을 선택했을 때, 해당 논문의 전반적인 목적과 내용을 쉽게 이해할 수 있도록 상세 요약을 제공한다. 이를 통해 사용자는 자신이 찾고자 하는 논문이 맞는지 판단할 수 있다.  **논문 요약 전처리 목적:** 논문 원문(.pdf)에서 요약에 필요한 핵심 내용을 추출하고, 요약에 적합한 형태로 정리하여 추후 입력 token 수 제한과 LLM의 End-of-Sequence Bias 현상을 고려한 입력값을 생성한다. |
| --- | --- |
| **데이터 자동화 및 검증** | AWS의 Lambda 및 Batch를 활용하여 자동화 프로세스를 구축하였다. 과정은 다음과 같다.  **I. ACLAnthology API**   1. 연도와 학회에 해당하는 논문 리스트를 조회한다. 2. 조회된 리스트에서 저작권 문제가 없고, ACL 마킹이 있는 논문을 선별한다. 3. 선별된 논문은 AWS의 S3에 저장되며, 유지보수의 용이성을 고려하여 저장 경로는 “{conference}/{year}/{doi.replace('/', '\_')}.pdf” 형식을 따른다. |
| **데이터 저장 및 관리** | AWS S3에 학회별, 연도별 논문 원문 데이터베이스를 구축하여 논문의 DOI를 알면 S3를 통해 사용자가 웹 인터페이스에서 시각화된 논문을 읽을 수 있도록 관리한다. |
| **데이터 전처리 과정** | [그림 1. 논문 요약 목적 데이터 전처리 과정] |
| **데이터 전처리 결과** | PDF Reader 외에도 추가 전처리 및 eng-eng 요약 단계를 추가하여 다음과 같이 토큰 수를 줄인 후 더 효과적으로 요약을 추출할 수 있었다.  **[2024.acl-long.834] 토큰 수 변화**  11702(PDF Reader 결과) → 6294(문장 결합 및 내용 추출) → 531(eng-eng 요약)  **[2022.emnlp-main.759] 토큰 수 변화**  10240(PDF Reader 결과) → 5989(문장 결합 및 내용 추출) → 496(eng-eng 요약) |

1. **논문 검색 목적 데이터 전처리**

| **개요** | **논문 검색 기능 목적 :**  특정 검색어 입력 시 유사도 검색을 통해 사용자가 원하는 논문을 효율적으로 탐색할 수 있도록 지원하는 것을 목표로 한다.  **논문 검색 전처리 목적 :**  사용자가 입력한 검색어(영어 키워드)와 관련된 논문을 효과적으로 추천할 수 있도록, 유의미한 주요 영어 키워드를 다양하게 추출한다. |
| --- | --- |
| **데이터 자동화 및 검증** | AWS의 Lambda 및 Batch를 활용하여 자동화 프로세스를 구축하였다. 과정은 다음과 같다.  **I. ACLAnthology API**   1. 연도와 학회에 해당하는 논문 리스트를 조회한다. 2. 조회된 리스트에서 저작권 문제가 없고, ACL 마킹이 있는 논문을 선별한다. 3. title, authors, venue, publicated\_year, publicated\_month에 해당하는 정보를 수집한다.  * title : 논문 제목, 영문 문자열 * authors : 저자(들), 영문 문자열 * eng\_abstract : 영문 초록, 영문 문자열 * venue : 학회, 영문 문자열 (acl, emnlp) * publicated\_year : accept 연도, 정수형 * publicated\_month : accept 월, 영문 문자열 * doi : 논문의 고유 doi, 영문 문자열   **II. SemanticScholar API**   1. ACLAnthology에서 수집한 논문 리스트를 기반으로, citation, references 정보를 수집한다.  * citation : 논문 인용수, 정수형 * reference\_papers : 참조 논문, Json형식 객체 리스트   + paper\_doi : 참조 논문의 고유 doi, 영문 문자열   **III. Google Translate**   1. ACLAnthology에서 수집한 논문 리스트를 기반으로, eng\_abstract를 kor\_abstract로 영한번역을 수행한다.  * kor\_abstract : 국문 초록, 국문 문자열   **IV. BART, OpenAI**   1. ACLAnthology에서 수집한 논문 리스트를 기반으로, BART를 사용하여 eng\_abstract를 요약한다. 2. OpenAI에 요약된 영문 문자열을 포함하여 프롬프트를 입력하고, 한국어로 적절히 추출된 논문의 상세 요약 내용을 제공받는다.  * generated\_summary : 생성된 정리 내용, Json형식 객체   + 논문의 키워드 : 논문의 키워드, 문자열   + 논문 요약 : 논문의 전반적 요약 내용, 문자열   + 논문의 핵심 방법론 : 핵심 방법론 리스트, 문자열 리스트   + 핵심 활용 기술 및 설명 : 핵심 활용 기술과 그에 대한 설명, 문자열   + 실험 내용 : 실험의 설계 및 평가 방법 내용, 문자열   + 실험 결과 : 실험의 결과 내용, 문자열   **V. SciBERT**   1. 데이터 베이스에 구축된 논문의 리스트를 기반으로, 아래의 데이터를 임베딩화한다.  * eng\_abtract 문장 단위 리스트 * title * 논문의 키워드 * 논문의 핵심 방법론 리스트  1. 해당 임베딩 결과를 FAISS 형식으로 AWS의 S3에 저장한다. |
| **데이터 저장 및 관리** | 논문 상세 데이터를 RDS(MySQL)의 paper 테이블에 저장하고, 논문의 내용이 포함된 벡터데이터를 S3에 저장함으로써 검색된 논문의 DOI를 기반으로 다양한 상세 정보를 웹 인터페이스에서 효율적으로 제공할 수 있다. |
| **데이터 전처리 과정** | [그림 2. 논문 검색 목적 데이터 전처리 과정] |
| **데이터 전처리 결과** | [그림 3. 논문 검색 목적 데이터 전처리 결과]  위와 같이 논문의 제목, 목적, 기여, 키워드, 핵심 방법론 등 주요 요소를 추출하여 검색 정확도를 향상시킬 수 있었다. |

1. **논문 요약 목적 데이터 전처리**

| **개요** |  |
| --- | --- |
| **데이터 자동화 및 검증** |  |
| **데이터 저장 및 관리** |  |
| **데이터 전처리 과정** |  |
| **데이터 전처리 결과** |  |