데이터 수집 및 데이터베이스 설계 문서

SKN04-오정연

1. 데이터 수집

A. 데이터 정보

1) 논문 기본 정보 수집

- 데이터 출처: Arxiv 논문 웹사이트
- 수집 방법: Arxiv API를 활용하여 데이터 크롤링
  (참조: https://info.arxiv.org/help/api/basics.html)
- 수집 기간: 2016년 10월 1일 - 2024년 12월 24일
- 수집 논문 수: 70,000개
- 수집 대상 논문 분류: Computer Science > Computational and Language (cs.CL)

2) 수집된 논문 데이터 항목

- entry_id: 논문 고유 식별자
- updated: 논문 새로운 버전 업데이트 날짜
- published: 논문 게재 날짜
- title: 논문 제목
- authors: 저자 목록
- summary: 논문 요약 (Abstract와 동일)
- comment: 추가 정보
- journal_ref: 저널 정보
- doi: DOI 정보
- primary_category: 논문의 주요 분류
- categories: 논문의 전체 분류 (예: cs.LG, cs.AI, cs.CL)
- links: 논문 소개 페이지, 논문 원문 pdf 링크

3) 논문 원문 pdf 다운로드

- 각 논문의 고유 ID를 사용하여 원문 PDF 파일 다운로드
- 수집된 모든 논문의 PDF 파일을 별도 저장
- 논문 게재가 철회되었거나 PDF 파일에 접근할 수 없는 경우, 해당 데이터 삭제 처리

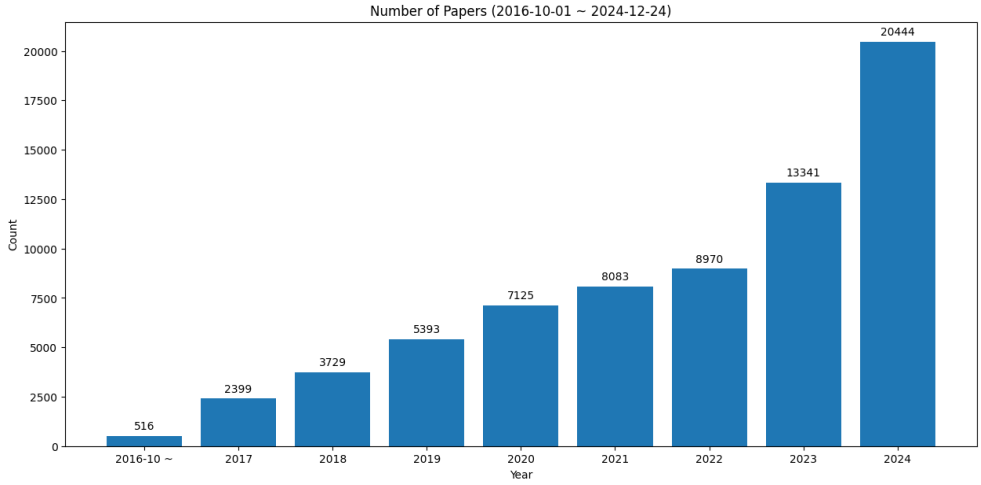

그림 1. 연도별 논문 수

B. 데이터 전처리

- 논문 PDF 파일에서 텍스트 추출하여 분석 가능한 형식으로 변환
- 각 논문에서의 Abstract, Related Works, Background 부분 추출 후 개별 저장
- 추출된 텍스트에서 불필요한 문구 제거

2. 데이터베이스 설계

1) Paper

- 각 논문 별 기본 정보와 abstract, background, related work에 대한 내용을 모아둔 테이블

2) Paper_Keywords

- 각 논문 별 추출된 키워드에 대한 정보 저장

3) Keywords

- 키워드에 대해서 저장해둔 테이블