**팀 회의 #4**

| 일시: 2023.12.04(Mon) PM 20:00 |
| --- |
| 장소: 학생회관 5층 명지대기독교연합실 |
| 참석자: 허재석(60171791), 정성훈(60191686), 신지훈(60201687), 신지영(60211674) |

[프로젝트 주제: 찾았다! 내 Reference! (참고 논문)]

[Goal]

* Abstract → 비슷한 Abstract 내용의 논문 리스팅

[빅데이터 프레임워크 사용 이유]

ex) 뉴스 데이터(100년 동안 수집) → WordCount → 년도별 가장 많이 나온 단어 수 → 그 해는 이런 키워드가 이슈였구나

논문(100년 동안 수집한) →

단순히 우리가 하려고 했던 것

Abstract 입력 넣었을 시

1년치 Abstract 데이터를 다 가져옴

→ BERT로 유사도 값 계산 → 정렬 후 가장 Top 논문

[데이터 전처리] → Task 1

* Title, Authors, Subject, Abstract, Year, Month
* 논문 데이터셋이 없어서 Local에서 만듬

[논문 데이터 분석]

* 1. 하둡 기반 불용어 처리한 후 새로운 csv파일 생성하는 코드 → Task 2
* 연도별 논문 트렌드 분석 - Task 3

1. 연별, 월별 Abstract → WordCount → Top K + (threshold)
2. 연별, 월별 Title → WordCount → Top K
3. 연도별, 월별 논문 게시 수 → 시각화(그래프)  
   ex) 2023년 1~12월 각 별 논문 게시 수

* 키워드 별 논문 그룹화 - Task 4  
  → Top K keyword에 대한 Title or Abstract → Title
* tf-idf, word2vec spark rdd api 사용 가능 → 어떻게 활용할지는 생각  
  https://spark.apache.org/docs/latest/mllib-feature-extraction.html

Task 1 : 데이터 전처리 → 지훈

Task 2 : 하둡 기반 불용어 처리 후 새로운 csv 파일 생성하는 코드 → 지영

Task 3 : 연도별 트렌드 분석 → 성훈

Task 4 : 키워드별 논문 그룹화 → 재석