**팀 회의 #2**

| 일시: 2023.11.20(Mon) PM 21:00 |
| --- |
| 장소: 명지대학교 MCC 학식당 |
| 참석자: 허재석(60171791), 정성훈(60191686), 신지훈(60201687), 신지영(60211674) |

[프로젝트 주제: 찾았다! 내 Reference! (참고 논문)]

[데이터 수집]

Q. 어디에서 데이터 수집할 것인가?

* arXiv, google scholar

arXiv

* 장점: 논문이 분야별로 잘 카테고리화됨.
* 단점: keyword 검색 시 원하는 게 잘 안 나옴. ㅠ

google scholar

* 장점: keyword 검색 시 관련 논문을 꽤 잘 찾아줌
* 단점: 분야별 카테고리화가 잘 안되어 있음.

Computer Science에 대한 논문 수집

⇒ arXiv에서 크롤링 진행하기로 결정

[크롤링 방법]

* Computer Science 카테고리: Artificial Intelligence, Machine Learning, Computer Vision and Pattern Recognition
* 데이터가 더 필요할 경우

1. Computer Science 안의 다른 카테고리까지 포함하도록 범위 확장

2. 모든 카테고리가 포함될 경우 논문 기간 범위 확장

* Meta 데이터 범위
  + Title:
  + Authors:
  + Abstract:
  + Subjects:
  + Date(2순위):

[데이터 형태]

* Title:
* Authors:
* Abstract:
* Subjects:
* Date:

[팀 룰 - 일정 관리]

[데이터]

* 데이터 크롤러 제작
  + Date:
* 데이터 크롤링 & 전처리
  + Date:

[알고리즘]

* 모델 설정
* 모델 구현
  + Date:
* 모델 학습 & 검증
  + Date:

[최종 발표]

[깃 허브]

[전체 시스템 아키텍쳐]

Input:

* 내가 현재 보고 있는 논문의 Abstract

Output:

* 입력한 Abstract의 내용과 비슷한 주제의 논문 정보
* 논문 Title, 저자, Abstract 등

핵심 기능

* 입력 Abstract와 유사한 논문 찾기 : Input Abstract ←→ (다른 Abstract) ⇒ 얼마나 유사한가?

[모델 구현 방식]

1. 유사도에 대한 기준을 어떻게 정할 것인가? ⇒ 텍스트 간의 유사도 구하기 ⇒ Top N 추천

Abstract에 대한 Multi-label Classification 문제로 접근하는 방법

→ Text를 어떻게 의미를 포함하게 Vector화 할 것인가?

1. Input Abstract에 대한 Multi-label 출력

데이터셋의 여러 논문의 Abstract에 대한 Multi-label 출력 ⇒ 유사도? ⇒ Top N 추천

→ String(Abstract) → Multi Label(Categories)

[전체 일정]