**팀 회의 #3**

| 일시: 2023.11.22(Wed) PM 20:30 |
| --- |
| 장소: 명지대학교 도서관 스터디룸 11 |
| 참석자: 허재석(60171791), 정성훈(60191686), 신지훈(60201687), 신지영(60211674) |

[프로젝트 주제: 찾았다! 내 Reference! (참고 논문)]

[데이터 수집]

데이터 크롤링()

* Input: arXiv 홈페이지 관련 크롤링 정보
* Output:
  + Title, Authors, Abstract, Subjects, Date 로 구성된 csv 파일

csv 파일 저장하는 디렉토리 구조

Dataset

[데이터 전처리]

Input: 크롤링된 논문 csv 파일

* 불용어 처리(해야 한다면)
* 모델 / 알고리즘 요구사항에 맞게 진행

[모델 / 알고리즘 학습]

* Model / Algorithm

Text → vector : 모델 / 알고리즘

* Doc2Vec
* DocBERT

vector 간 유사도 검사: 모델 / 알고리즘

* 코사인 유사도 등 벡터 끼리 유사도 검증하는 기본적인 방법 ex 자카드, 코사인, n-gram

1. TFW2V

https://github.com/ruathudo/tfw2v

https://arxiv.org/pdf/2112.12489v1.pdf

2. Doc2Vec

https://arxiv.org/pdf/1405.4053.pdf

https://radimrehurek.com/gensim/models/doc2vec.html

3. DocBERT

https://arxiv.org/pdf/1904.08398.pdf

https://paperswithcode.com/paper/docbert-bert-for-document-classification

https://github.com/castorini/hedwig

Output: 유사한 논문 top n개

[결과값 분석]