**팀 회의 #1**

| 일시: 2023.11.15(Wed) AM 11:00 |
| --- |
| 장소: 명지대학교 MCC 학식당 |
| 참석자: 허재석(60171791), 정성훈(60191686), 신지훈(60201687), 신지영(60211674) |

[프로젝트 주제]

신지훈

* AIHUB 여행 로그 데이터 기반 장소 추천 → 정량적으로 분석하기에 유리할 것 같음

신지영

* 증상별 어떤 질병 예측하는 것 → 증상별 유무 데이터
* 범죄 데이터(Toronto) → 범죄 예측

정성훈

* arxiv 논문 수집 → 분석 → 원하는 논문 검색 & 찾기

허재석

* SNS

[팀 프로젝트 - 계획서]

문제 정의

* 원하는 주제의 논문을 찾기 어려움
  + 보통 원하는 주제를 Keyword로 논문 검색 → 동음이의어가 존재하므로 원하는 분야의 논문이 아닌 경우 있음
* 현재 보고 있는 논문과 비슷한 주제의 논문 찾기 어려움
* 검색 Keyword를 포함하고 있지만 전체 내용은 다른 내용인 경우 존재

프로젝트 주제:

* 유사 연구 주제 논문 프로그램?
* 찾았다! 내 Reference! (참고 논문)

프로젝트 목표:

* keyword 혹은 논문의 일부 내용을 입력 시 원하는 주제와 가까운 논문 리스트 제공

실행 계획

* 데이터 수집 방법   
  [https://arxiv.org](https://arxiv.org/) 사이트에서 논문 데이터 수집
* 데이터 분석 방법

1. 데이터 수집 단계: arXiv 사이트에서 크롤링하여 논문 데이터 수집.

2. 데이터 전처리: 수집된 데이터 중 필요한 정보 추출 및 정제.

3. 모델 개발 단계: 자연어 처리 기법을 활용하여 입력된 Keyword 또는 논문 일부 내용과 유사한 주제를 가진 논문을추출하는 모델 개발.

4. 프로그램 구현: 개발한 모델을 기반으로 사용자가 입력한 정보를 받아 유사한 주제의 논문 리스트를 반환하는 프로그램 구현.

[작년 프로젝트 예시]

1. OTTO 추천 시스템

* 데이터 개수: 약 2억개
* 원핫 인코딩의 통합
* 특정 아이템(item N)에 대한 Action(1, 2, 3)
* 기존 one hot encoding으로 아이템을 표현한 뒤 1이 아니라 Action의 값을 바로 할당

⇒ 교수님 피드백: 프로젝트 완성이 중요함 & 최종 진행도까지의 결과 & 한계점

1. 중증 환자를 위한 음압 병동 설치 입지 분석

* 전체 시스템 아키텍처

1. 데이터 수집 = {블로그 본문 데이터, 지자체별 유동인구, 병원 수, 병상 수 데이터, 고령인구 비율 데이터}
2. 데이터 분석 = {텍스트 전처리, 데이터 수치화, 감정지수 분석}
3. Output:

3. CRAFT: text detection

⇒ 이걸 왜 Hadoop에서 해야 하는지!