**Weekly report**

날짜 : **2023-04-06**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - 테이블 융합 관련 논문 추가 검색
    - 논문 아이디어를 기반으로 코드 작성 및 실험
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성).
  + 학술대회 논문의 아이디어를 기반으로 코드 작성 및 실험
    - 아이디어 :
      * 모든 테이블에 속한 각 컬럼(컬럼명, 컬럼에 속한 값)들을 EmbDI를 이용해 임베딩.
      * 서로 다른 두 테이블의 임베딩된 컬럼끼리 벡터간의 코사인 유사도를 계산하고, 그 중 가장 큰 값을 두 테이블간의 조인 가능성으로 사용함.



* + - * 조인 가능성이 높은 순서대로 테이블들을 제시함.
    - 코드 작성현황
      * 두 테이블의 컬럼명을 임베딩하여 나온 벡터간의 코사인 유사도를 계산하는 부분 작성 완료
      * 두 테이블의 컬럼에 속한 값을 임베딩하여 나온 벡터간의 코사인 유사도를 계산하는 부분 작성완료
    - 코드 추가할 부분
      * 각 테이블간 조인 가능성을 계산하는 부분

* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - 테이블 융합 관련 논문 추가 검색
    - 논문 아이디어를 기반으로 코드 작성 마무리 및 실험   
      추가할 부분 : 성능평가 Baseline을 정해서 다른 기법과 비교
* **기타사항**
  + **리포트 작성 시, 논문의 내용정리 같이 남이 한 부분에 대한 설명은 자세하게 하지 말고 그 중에서 내가 얻은 아이디어나 내 연구에 관한 부분을 자세하게 설명할 것**
  + **테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각**
  + **데이터 융합 프로세스**





혼합형 데이터셋 예시