**Weekly report**

날짜 : **2023-04-20**

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 마무리 및 실험   
      추가할 부분 : 성능평가 Baseline을 정해서 다른 기법과 비교
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성).
  + 학술대회 논문의 아이디어를 기반으로 코드 작성 완료
    - 코드 작성현황
      * (지난 주 작성한 것)두 테이블의 컬럼명을 임베딩하여 나온 벡터간의 코사인 유사도를 계산하는 부분 작성완료
      * (지난 주 작성한 것)두 테이블의 컬럼에 속한 값을 임베딩하여 나온 벡터간의 코사인 유사도를 계산하는 부분 작성완료
        + (이번 주 작성한 것)두 테이블의 컬럼에 속한 값에 대해 코사인 유사도를 계산하는 것을 기존의 모든 값에 대해 전부 코사인 유사도를 계산하고 더하는 방식이 아닌, 모든 값에 대해 전부 코사인 유사도를 계산하고 그 중 가장 큰 값을 선택하는 방식으로 변경
      * (이번 주 작성한 것)각 테이블간 조인 가능성을 계산하는 부분 작성완료
    - 코드 실행결과

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure . EmbDI에서 예시로 사용한 데이터셋

* + - * EmbDI에서 예시로 사용한 amazon에서 파는 물품과 google에서 파는 물품에 대한 테이블
      * Fig 1의 테이블을 보면 컬럼명에 0\_ 과 1\_로 시작하는 것을 볼 수 있는데, 이는 EmbDI에서 Schema Matching을 진행할 때 권장하는 구조임

도표이(가) 표시된 사진

자동 생성된 설명

Figure . Schema Matching시 EmbDI가 권장하는 구조, 왼쪽 두 테이블을 합친 오른쪽 테이블을 EmbDI 입력으로 넣어줘야 함

* + - * Fig 1의 테이블을 내가 만든 테이블의 입력으로 넣으면 최종적으로 각 컬럼에 대한 조인 가능성을 내림차순으로 제시함

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure . 내 코드의 최종 산출물, 각 컬럼끼리의 조인 가능성을 내림차순으로 제시함

* + - * 원하는 결과는 0\_maufacturer 과 1\_maufacturer, 0\_price와 1\_price 그리고 0\_title과 1\_title이 높은 조인 가능성을 보이는 것이었는데 그렇지 못함
      * 그 이유는 Fig 3는 두 테이블의 컬럼에 속한 값에 대해 코사인 유사도를 계산하는 것을 기존의 모든 값에 대해 전부 코사인 유사도를 계산하고 더하는 방식으로 나와서 그런 것 같음
      * 모든 값에 대해 전부 코사인 유사도를 계산하고 그 중 가장 큰 값을 선택하는 방식으로 변경하여 다시 코드를 돌려보는 중임
  + EmbDI 논문 내용으로 수업시간에 발표할 PPT 작성 완료
  + Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽는 중

* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - 테이블 융합 관련 논문 추가 검색
    - EmbDI의 임베딩을 개선하여 더 좋게 만드는 방법에 대해 연구
* **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