**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
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - 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
    - DeepJoin: Joinable Table Discovery with Pre-trained Language Models 읽고 정리
    - 테이블 융합 관련 논문 추가 검색
    - 논문 관련 아이디어 정리
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + ITRC 포럼에 제출할 연구내용, 그림, 제목 작성하여 정주은 연구원한테 전달
  + DeepJoin: Joinable Table Discovery with Pre-trained Language Models 읽고 핵심 아이디어 및 내 연구에 도움이 될 부분 정리
    - 연구의 목적
      * Pre-trained language model(PLM)을 사용하여 컬럼(컬럼명, 컬럼에 포함된 값들, …)을 벡터로 임베딩하고, 임베딩한 벡터간의 거리에 기반해 조인이 가능한 테이블을 찾는 것.
    - 논문에서 제시한 모델의 구조

도표이(가) 표시된 사진

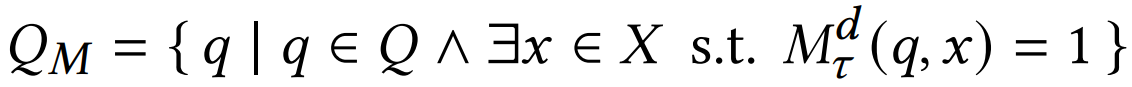
자동 생성된 설명

* + - * Training을 통해 PLM을 fine tuning하고, fine tuning한 PLM을 이용해 조인 가능한 테이블을 찾음.
      * Column-to-text transformation은 PLM이 인식할 수 있게 컬럼명과 컬럼에 포함된 값을 텍스트 형태로 나열하는 것
      * 텍스트이(가) 표시된 사진

        자동 생성된 설명학습에 사용하는 데이터셋을 만드는 방법
        + 텍스트이(가) 표시된 사진

          자동 생성된 설명테이블을 셀프 조인하여 조인이 가능한 컬럼 쌍(X,Y)을 찾음. 이때 조인이 가능한지 여부는 jn(X,Y) 함수가 threshold를 넘는지로 판단함.

Equivalent – join 일 때

  
  
Semantic – join 일 때

* + - * + 조인이 가능한 컬럼 쌍(X,Y)를 제외한 컬럼 쌍 (X, C)들은 전부 조인이 불가능한 데이터셋으로 사용됨
      * 학습이 완료된 PLM을 통해 text를 벡터로 임베딩하고, Approximate Nearest Neighbor Search(ANNS)를 이용해 조인이 가능한 테이블을 찾음
      * ANNS는 고차원 벡터 공간에서 가장 가까운 이웃을 탐색하는 알고리즘으로, 대규모 데이터 집합에서 가장 가까운 이웃을 찾는 데 사용됨.
    - 내 연구에 사용할 부분
      * 나의 연구에서는 컬럼명, 컬럼값들을 임베딩할 때, EmbDI를 사용함.
      * 이를 사용하는 가장 큰 이유는 각 테이블에 맞는 의미적 정보를 고려하기 때문임.
      * 다만 EmbDI는 시간이 매우 오래걸리는 단점이 있음.
      * 하지만 PLM을 fine tuning하여 사용하면, 각 테이블의 의미적 정보도 어느정도 고려하면서 임베딩 시간을 단축할 수 있을 것 같음.
      * 실험할 때 EmbDI와 PLM 두 가지를 사용하여 성능을 비교해볼 예정.
  + 학술대회 논문의 문제와 아이디어 정의
    - 문제 : 어떤 테이블과 조인이 가능한 테이블들을 찾아내는 것을 자동으로 하고 싶음.
    - 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.
  + 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
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - 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
* **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**
  + **데이터 융합 프로세스**



* **융합 가능도 척도**
  + 임베딩 거리 기준 값을 정하고 기준값보다 임베딩 거리가 가까우면 융합 가능한 데이터셋이라 판단함
  + 융합 가능한 데이터셋을 사람이 하나씩 확인하여 유의미한 융합이 되는 데이터셋을 찾음
  + 을 융합 가능도 척도라 정의함



혼합형 데이터셋 예시