**Weekly report**

날짜 : **2023-04-27**

연구원 : **최윤석**

* **이전 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
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성).
  + Efficient Joinable Table Discovery in Data Lakes: A High-Dimensional Similarity-Based Approach 읽고 정리
    - 논문은 테이블간 Join하는 방법에 초점을 두기보단, 테이블간 Join작업을 어떻게 하면 효율적으로 할 수 있을지에 초점을 두고 있음.
      * 나의 연구에서는 Join하는 방법에 관심이 있고, Join 작업의 효율성에는 크게 관심이 없기 때문에, 논문에서 제시한 Join하는 방법에 관해서만 소개함.

도표이(가) 표시된 사진

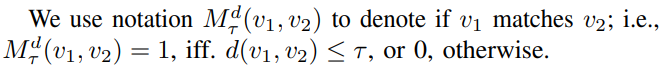
자동 생성된 설명

Figure . 논문에서 제시하는 Joinable table search 프로세스

* + - Join에 사용할 key column을 정하는 방법은
      * 1. 사용자가 직접 key column 지정
      * 2. 가장 unique한 값이 많은 컬럼을 key column으로 사용
      * 3. 모든 컬럼을 한 번씩 전부 key column으로 사용
    - Key column이 주어지면 column에 포함된 값에 대해 전부 pre-trained language model(ex : fasttext)을 사용하여 벡터로 바꿔줌
    - 두 컬럼이 Join이 가능한지는 아래 식을 활용해서 알아냄

텍스트이(가) 표시된 사진

자동 생성된 설명



* + - Jn(Q,S) 가 미리 지정한 threshold 값보다 크면 두 컬럼 Q와 S가 Join 가능하다고 판단하고 threshold 값보다 작으면 두 컬럼 Q와 S가 Join 불가능하다고 판단함.

* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - 테이블 융합 관련 논문 추가 검색
    - EmbDI의 임베딩을 개선하여 더 좋게 만드는 방법에 대해 연구
      * W2V 대신에 Elmo / Bert 등의 임베딩을 사용하는 방안
    - 수업시간에 발표할 PPT 작성
* **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