**Weekly report**

날짜 : **2024-03-29**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 시간 단축 문제는 가장 나중으로 진행
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(use case diagram)
      * PPT로 작성
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - 테이블 임베딩을 위한 Feature selection 방법 탐색
      * 테이블에서 임베딩 해야 하는 중요 컬럼을 자동 / 휴리스틱을 이용해 알아내는 방법 탐색
      * 아래의 논문 읽고 아이디어 얻기
        + <https://arxiv.org/pdf/2310.07736.pdf>
        + <https://arxiv.org/pdf/2207.02773.pdf>
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가
      * 추가 학습한 모델과 기존의 pre-trained 모델로 n-join-pair 성능평가를 진행해 더 좋은 containment 그래프를 얻은 모델의 성능이 더 좋다고 판단
    - 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
    - 실험에 사용할 테이블 데이터 추가로 수집
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 실험에 사용할 테이블 데이터 추가 수집
      * 김승준 연구원의 크롤링 코드를 활용해 Kaggle에서 700개의 테이블 데이터를 수집
      * 700개의 테이블 데이터를 입력으로 CNE-Join 코드를 돌려보았지만, 코드가 돌아가는 데 걸리는 예상 시간이 너무 길어 램을 추가로 장착하고 용량이 적은 테이블 데이터 300개를 활용하여 코드를 돌리는 중

텍스트, 스크린샷, 번호, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* + - BERT 최신 데이터로 추가 학습한 모델의 성능평가
      * 앞서 300개의 테이블 데이터를 활용해 돌린 CNE-Join 코드의 결과가 나오면 이를 n-join-pair의 입력으로 넣고, n-join-pair에서 사용하는 임베딩을 1. 기존의 Word2Vec / 2. Pre-trained된 BERT / 3. 최신 데이터로 추가 학습한 BERT 로 바꿔가며 Containment를 활용해 각각 임베딩에 대한 성능평가를 진행
    - Observatory: Characterizing Embeddings of Relational Tables 논문 읽는 중
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 시간 단축 문제는 가장 나중으로 진행
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(use case diagram)
      * PPT로 작성
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - 테이블 임베딩을 위한 Feature selection 방법 탐색
      * 테이블에서 임베딩 해야 하는 중요 컬럼을 자동 / 휴리스틱을 이용해 알아내는 방법 탐색
      * 아래의 논문 읽고 아이디어 얻기
        + <https://arxiv.org/pdf/2310.07736.pdf>
        + <https://arxiv.org/pdf/2207.02773.pdf>
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가
    - 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
* **기타사항**