**Weekly report**

날짜 : **2024-04-19**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - 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(그림으로) – 논문 서론에 사용할 용도로 간략화 추상화된 그림을 작성
      * 쿼리 바이 테이블도 설명은 하되, 나는 어디까지 구현을 했는지를 표현
      * 서비스 사용자 입장에서 어떻게 융합된 테이블을 얻게 되는지
      * 서비스 개발자 입장에서 무엇을 하는지
      * PPT로 작성 Type A B 구분하지 말고
      * 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
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가
      * 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 300개의 테이블 데이터를 입력으로 CNE-Join 코드 실행
      * CNE-Join 결과로 642개의 조인 가능한 2개의 테이블쌍을 찾아냄
      * 642개의 테이블쌍을 조합해, 조인 가능한 3개의 테이블쌍 3300개를 찾아냄
    - 3300개의 테이블 조합에 대해 조인을 진행
      * Memory / Key / 테이블 불러오기 에러 등을 제외하고 600개의 테이블 조합만 조인됨
    - 600개의 조인된 테이블들에 대해 N-Join-pair(Word2Vec) 코드를 실행한 결과
    - N-Join-pair(BERT) / N-Join-pair(BERT-FineTuning) 코드 작성 및 실행중
      * 코드 실행 중 오류가 발생하여 해결하는 중
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - 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(그림으로) – 논문 서론에 사용할 용도로 간략화 추상화된 그림을 작성
      * 쿼리 바이 테이블도 설명은 하되, 나는 어디까지 구현을 했는지를 표현
      * 서비스 사용자 입장에서 어떻게 융합된 테이블을 얻게 되는지
      * 서비스 개발자 입장에서 무엇을 하는지
      * PPT로 작성 Type A B 구분하지 말고
      * 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
      * 나만의 Feature selection 아이디어가 있으면 연구의 Sub결과로 활용 가능하니 아이디어의 변형이 있으면 고민해보기 - CNE-Join에서 Feature selection으로 시간 줄이기 가능 / N-Join에서 Feature selection은 성능 향상으로 가능
    - BERT 최신 데이터로 추가 학습한 모델의 성능평가
      * 추가 학습한 모델로 두 테이블의 유사도를 판단하는 방법 구현
    - N-Join-pair에서 Containment를 평가 척도로 쓰지 말고 조인된 테이블의 가치를 평가하는 새로운 평가 척도 만들기 : 조인된 테이블에 포함된 단어의 임팩트를 조합하여 평가하는 방법에 대해 고민 (TF-IDF와 같은 느낌의 새로운 척도)
      * https://www.enigma.com/resources/blog/tf-idf-for-tabular-data-featurization-and-classification
    - 5월 3일 랩미팅에서 발표할 중간심사 리허설 PPT 작성
* **기타사항**
  + **5월 3일 랩미팅에서 중간심사 리허설**
    - **20분 발표**