**Weekly report**

날짜 : **2023-11-08**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - 박찬훈 연구원과 분담하여 빅데이터자료구조 중간고사 채점
    - 조인 테이블 개수를 3,4개 이하로 테이블을 임베딩한 벡터들을 2차원 공간에 시각화하기
    - 고급정보검색론 수업시간에 발표할 논문 읽고 발표 PPT 작성

* **수행결과**
  + **관계형 테이블 임베딩 :**
  + On the resemblance and containment of documents 읽고 정리
    - 두 개의 문서 A, B 가 주어졌을 때 두 문서의 유사도와 포함도를 계산하는 식을 제안하는 논문

수식 . 유사도 식

폰트, 텍스트, 화이트, 친필이(가) 표시된 사진

자동 생성된 설명

수식 . 포함도 식

폰트, 텍스트, 라인, 화이트이(가) 표시된 사진

자동 생성된 설명

* n-join-pair에 관한 아이디어는 얻지 못함.
* 조인 테이블 개수를 3,4개 이하로 테이블을 임베딩한 벡터들을 2차원 공간에 시각화하기
  + 윤종찬 연구원의 제안모델인 TabBERT를 일부분 수정하고 사용해, 하나의 테이블을 하나의 벡터로 임베딩을 진행하고자 함

텍스트, 포스트잇 노트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Figure . 윤종찬 연구원의 TabBERT 제안모델

* + 윤종찬 연구원의 코드를 수정하고, 이를 이용해 CNE-join에 사용한 50여개의 테이블들을 각각 하나의 벡터로 매핑할 생각임
  + 이후 매핑한 테이블 벡터를 2차원 공간에 시각화 할 예정
    - * 현재 코드를 수정하는 중
* 고급정보검색론 수업시간에 발표할 논문 읽는 중
  + Text classification using embeddings: a survey
* **빅데이터 자료구조 중간고사 채점 :**
  + 박찬훈 연구원과 분담하여 중간고사 채점완료
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - 조인 테이블 개수를 3,4개 이하로 테이블을 임베딩한 벡터들을 2차원 공간에 시각화하기
    - 고급정보검색론 수업시간에 발표할 논문 PPT 작성

* **기타사항**