**Weekly report**

날짜 : **2023-10-04**

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
      * 시간 단축 문제는 가장 나중으로 진행
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
    - 한국인공지능 학술대회 논문 발표

* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 한국인공지능 학술대회 논문 발표완료
    - 박찬훈 연구원과 DataMap Type B를 어떻게 개발할지에 대해 토론
      * Type B : Query 형식의 CNE-join 코드와 사용방법을 박찬훈 연구원에게 전달
    - CNE-join의 성능을 개선할 방법 제안
      * 새로운 아이디어 : CNE-join(Query version) + 테이블 클러스터링을 통해 사용자가 CNE-join의 결과로 출력된 테이블 중 조인하고 싶은 n개의 테이블을 수월하게 선택할 수 있게 도와주는 모델

텍스트, 스크린샷, 도표, 디자인이(가) 표시된 사진

자동 생성된 설명

* + - * 위의 그림은 아이디어를 구체화한 것
      * **1**. 사용자는 하나의 테이블과 그 테이블에서 조인 기준이 될 컬럼을 지정함. 이후에 이 테이블을 조인 후보 테이블들이 존재하는 테이블 저장소와 같이 CNE-join에 입력함  
          
        **2**. CNE-join은 입력에 대한 결과로 입력된 테이블에 대해 조인 가능한 테이블들과 조인 기준이 되는 컬럼을 제시해줌  
        (1, 2번까지의 과정은 기존의 CNE-join(Datamap B타입)과 유사함)  
          
        **3**. CNE-join의 결과로 나온 테이블들을 클러스터링 기준에 맞춰 클러스터링  
        (클러스터링 기준 : 각 테이블에 포함된 컬럼명들을 전부 임베딩하고 임베딩한 벡터들을 더해 하나의 벡터로 만들고 벡터들의 거리를 기준으로 가까운 것들끼리 클러스터링)  
          
        4. 각 클러스터에 포함된 테이블의 모든 컬럼명을 활용해 WordCloud를 만들어 사용자가 각 클러스터에 포함된 테이블이 어떤 내용의 테이블인지 한눈에 알 수 있게 하여 사용자가 조인할 테이블을 선택하는 것을 도움.  
          
        아래의 그림은 이에 대한 예시임.

텍스트, 스크린샷, 폰트, 소프트웨어이(가) 표시된 사진

자동 생성된 설명

* 입력 테이블은 국가명, 코로나 확진자 수, 사망자 수, 완치자 수 컬럼으로 구성된 테이블이고 조인 컬럼은 국가명으로 지정되었음.
* 다양한 조인 후보 테이블들이 있는 테이블 저장소와 함께 입력 테이블을 CNE-join에 입력함.
* CNE-join 출력으로 입력된 테이블과 조인 가능한 테이블들과 조인 기준이 되는 컬럼이 나옴.
* 이를 클러스터링 기준에 따라 클러스터링하고 각 클러스터에 포함된 테이블의 모든 컬럼명을 활용해 WordCloud를 만듦.
* Cluster 1은 코로나와 국가별 백신에 관련된 테이블들이 있고, Cluster 2는 코로나와 국가별 GDP에 관련된 테이블들이 있음. Cluster3은 코로나와 국가별 인구 밀집도에 관련된 테이블들이 있음. 각 클러스터에 포함된 테이블 정보를 WordCloud를 보고 빠르게 파악할 수 있음
* 사용자는 클러스터를 보고 조인하여 유용할 것 같은 2, 3, 4, … n개의 테이블을 직접 고름.
* (추가사항)골라진 테이블들과 입력 테이블을 자동조인하여 최종적으로 사용자에게 반환함
* **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
        + 이를 위해 2개의 sequence들이 있을 때 이를 n개의 sequence로 확장시키는 알고리즘에 대해 공부
      * 시간 단축 문제는 가장 나중으로 진행
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
* **기타사항**