**Weekly report**

날짜 : **2023-08-24**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * DIALITE: Discover, Align and Integrate Open Data Tables가 참고한 논문들을 읽어보면서 새로운 아이디어 캐치
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
      * A. Top-N개의 조인 가능한 테이블을 반환하는 방식의 문제점(조인 가능성 값은 높은데 실제로는 조인이 불가능한 케이스가 존재)을 보완하는 코드 작성
      * B. 질의 방식(Query Table을 입력으로 주는 방식)의 CNE-join 만들기

* **수행결과**
  + **관계형 테이블 임베딩 :**
    - 박찬훈 연구원과 DataMap 개발
      * A. Top-N개의 조인 가능한 테이블을 반환하는 방식의 문제점(조인 가능성 값은 높은데 실제로는 조인이 불가능한 케이스가 존재)을 보완하는 코드 작성
        + 기존에 컬럼쌍의 조인 가능성을 계산할 때 사용한 식은 아래와 같음
        + 컬럼명 임베딩 벡터간의 코사인 유사도 값은 -1에서 1사이의 값을 가지고, Containment값은 0에서 1사이의 값을 가짐
        + 기존의 식을 사용할 때 컬럼쌍이 적은 종류의 값을 가지면 컬럼명이 비슷하지 않아도 조인 가능성이 높게 나오는 문제가 발생함

예를 들어, 컬럼쌍의 두 컬럼이 Yes, No 두개의 값만 가지는 경우, Containment 값이 1이 되어 컬럼명 임베딩 벡터간의 코사인 유사도 값이 낮아도 높은 조인 가능성 값을 가지게 됨.

* + - * + 이를 해결하기 위해 두 컬럼쌍의 조인 가능성을 계산할 때 컬럼쌍에 포함된 값의 종류가 5개 미만이라면 컬럼명 임베딩 벡터간의 코사인 유사도 값에 높은 비중을 두는 아래의 식을 사용함

텍스트, 폰트, 번호, 스크린샷이(가) 표시된 사진

자동 생성된 설명

Figure 1. 기존의 식을 사용하면 조인 가능성이 높아도 실제로는 조인이 불가능한 케이스가 존재

텍스트, 폰트, 스크린샷, 번호이(가) 표시된 사진

자동 생성된 설명

Figure 2. 실제로 조인 가능하지 않은 컬럼쌍에 대해 낮은 조인 가능성 값을 가지게 코드를 수정

* + - * + Fig 1은 조인 가능성을 계산할 때 아래의 기존 식을 사용함
        + Fig 2는 조인 가능성을 계산할 때 아래의 새로운 식을 사용함
        + Fig 1에서 조인이 불가능하지만 0.75의 높은 조인 가능성 값을 가진 컬럼쌍이 Fig 2에서는 0.61로 훨씬 낮아진 조인 가능성 값을 가지는 것을 확인
      * B. 질의 방식(Query Table을 입력으로 주는 방식)의 CNE-join 코드 작성

텍스트, 폰트, 화이트, 흑백이(가) 표시된 사진

자동 생성된 설명

Figure 3. 2019\_word\_happiness.csv를 질의 테이블로 주었을 때, 내가 가지고 있는 50개의 테이블셋 중 Top-N개의 조인 가능한 테이블셋을 반환함

* + - * + 특정 테이블(2019\_word\_happiness.csv)을 질의 테이블로 주었을 때,  
          **[질의 테이블, 조인 가능한 테이블, 질의 테이블의 조인 컬럼, 조인 가능한 테이블의 조인 컬럼, 조인 가능성 값]** 의 형식으로 Top-N개의 조인 가능한 테이블셋을 반환하도록 코드를 작성함

* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * DIALITE: Discover, Align and Integrate Open Data Tables가 참고한 논문들을 읽어보면서 새로운 아이디어 캐치
    - 박찬훈 연구원과 DataMap 개발(Type A, B)
* **기타사항**