**Weekly report**

날짜 : **2023-05-25**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - EmbDI의 임베딩을 개선하여 더 좋게 만드는 방법에 대해 연구
      * W2V 대신에 Elmo / Bert 등의 임베딩을 사용하는 방안
    - Search and Join Algorithms for Tables in Data Lakes 논문 읽고 정리
    - 학술대회에 제출할 아이디어 정리 (한국전자거래학회)
  + **객체지향프로그래밍 TA**
    - 객체지향프로그래밍 과제3 채점
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성).
  + Search and Join Algorithms for Tables in Data Lakes 논문 읽고 정리 / 내 연구에 사용할 아이디어 추출
    - 3개의 논문이 합쳐진 형태로 구성되어 있음  
      1. LSH Ensemble : Approximate Joinable Table Search  
      2. JOSIE : Exact Top-K Joinable Table Search  
      3. Auto-Join : Smart Join with Transformations
    - 1. LSH Ensemble : Approximate Joinable Table Search
      * **두 column value간의 similarity 계산기준으로 Jaccard similarity 대신 containment를 제시함**.
      * Containment는 식은 다음과 같음  
        반대의 경우도 구해서 둘을 (가중)평균 내는 방법도 있음 ->>이렇게 해야되는 이유는 조인은 교환법칙이 성립  
         Q : query column , X : 임의의 column
      * Containment가 Jaccard Similarity 보다 합리적인 이유
        + Column Q = {A, B} , Column X = {A, B, C,…,Z}
        + Containment는 2/2 = 1로 높은 값이 나오지만  
          Jaccard similarity는 2/26 = 0.08로 낮은 값이 나옴
        + Jaccard similarity는 두 컬럼에 포함된 value 개수가 크게 차이나면 성능이 좋지 않음
      * Query column과 조인이 가능한 column들을 찾기 위해 Query column과 다른 모든 컬럼들의 containment 계산을 하여 threshold 값보다 높으면 조인 가능하다고 판단함.
      * 이를 빠르게 하기 위해 Locality Sensitive Hashing(LSH) 기법을 사용함
    - 2. JOSIE : Exact Top-K Joinable Table Search
      * 앞선 논문과 마찬가지로 containment를 기준으로 한 column과 조인 가능성이 가장 높은 k개의 column을 찾음. 논문의 주된 내용은 그 시간을 줄이는 데에 초점을 둠
  + 조인 가능한 테이블을 찾는 아이디어 구상중
    - 지난주 수업시간에 발표한 Dataset\_Discovery 논문의 5가지 거리계산을 통해 두 column이 조인 가능한지 판단하는 것을 기본 틀로 사용하고 나만의 아이디어를 추가하는 방식으로 생각 중
    - 나만의 아이디어는 아직 구상중
* **객체지향프로그래밍 TA**
  + 과제3 채점완료 및 엑셀에 점수 기입완료
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - EmbDI의 임베딩을 개선하여 더 좋게 만드는 방법에 대해 연구
      * W2V 대신에 Elmo / Bert 등의 임베딩을 사용하는 방안
    - 학술대회에 제출할 아이디어 정리 및 수업시간에 발표할 PPT 작성
      * 1단계 : 임베딩에 관한 고민
        + 데이터 융합차원에서(조인) 범주형 + 조인만 가정하고 더 잘 찾아내는 방법
        + 수치형 + 조인만 가정하고 더 잘 찾아내는 방법
        + Ex. 주소를 가지고 조인할 때 ) 값들의 포맷이 천차만별일텐데 이를 해결하는 임베딩 방법
        + Partially하게 임베딩하는 방법 -> 풀 네임을 임베딩하기 보다 부분만을 임베딩 하는

* **기타사항**
  + **리포트 작성 시, 논문의 내용정리 같이 남이 한 부분에 대한 설명은 자세하게 하지 말고 그 중에서 내가 얻은 아이디어나 내 연구에 관한 부분을 자세하게 설명할 것**
  + **테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각**
  + **데이터 융합 프로세스**





혼합형 데이터셋 예시