**Weekly report**

날짜 : **2023-03-16**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기 (Embedding + Semantic Overlap(외부 knowledge base)을 섞는 방식 관련해서)   
      -> SANTOS: Relationship-based Semantic Table Union Search (외부 Knowledge Base를 사용하여 Table Unionability를 알아내는 알고리즘 제시)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Contextual word embeddings for tabular data search and integration 논문 읽고 정리
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + Contextual word embeddings for tabular data search and integration 논문 읽고 핵심 아이디어와 나의 연구에 적용할 부분 정리
    - 논문의 목적
      * word embedding을 이용해 테이블 데이터의 검색과 융합 작업을 하는 것

도표이(가) 표시된 사진

자동 생성된 설명

* + - 위의 그림은 논문에서 제시한 모델의 구조
    - Input tabular data를 embedding하여 벡터로 만들고, Tabular data repository의 테이블 벡터들과의 similarity를 기준으로 Union / Join이 가능한 테이블들을 구함. 이후 Input tabular data를 Union / Join이 가능한 테이블과 융합하여 새로운 데이터를 만들 수 있음.
    - 각 테이블을 embedding하는 것은 테이블에 있는 모든 컬럼을 embedding하는 것으로 진행됨.
      * 테이블의 각 컬럼당 2개의 벡터를 만듦.
        + 1. 컬럼명을 임베딩한 벡터
        + 2. 컬럼에 속한 값들을 임베딩한 벡터
      * 두 컬럼 벡터간의 Similarity 계산은 다음과 같이 진행됨.

텍스트, 손목시계, 게이지이(가) 표시된 사진

자동 생성된 설명

* + - * + 은 컬럼명 벡터들끼리의 코사인 유사도 값이고, 는 컬럼에 속한 값의 벡터들끼리의 코사인 유사도 값, 는 1에서 0사이의 실수로 각 식의 가중치를 정해줌
    - 논문의 핵심 아이디어
      * Union이 가능한 테이블을 찾는 기준 : 두 테이블이 많은 수의 높은 similarity 값을 가지는 컬럼쌍을 가지고 있는가?  
        텍스트, 안테나이(가) 표시된 사진

        자동 생성된 설명
        + 는 각 테이블에 속한 모든 컬럼의 집합
        + 는 1번 테이블에 속한 i번째 컬럼, 는 2번 테이블에 속한 j번째 컬럼
      * Join이 가능한 테이블을 찾는 기준 : 두 테이블의 컬럼쌍 similarity 중 가장 큰 값이 Threshold 값보다 높은가?  
        

* + - 내 연구에 적용할 부분
      * 나의 연구에서는 융합가능한 테이블을 찾기 위해 두 테이블의 컬럼, 레코드, 태그를 전부 embedding 하여 하나의 벡터로 합치고, 그 벡터들의 코사인 유사도를 기준으로 융합 가능성을 평가했었음.
      * 융합 가능한 테이블을 찾기 위해 논문에서 제시한 Join이 가능한 테이블을 찾는 기준을 활용하여 융합 가능성을 평가하면 좋은 성능을 기대할 수 있을 것 같음.
  + SANTOS: Relationship-based Semantic Table Union Search 논문 읽는 중
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각
    - SemProp(Attribute overlap + Value overlap + Embedding) 문제점을 찾고 문제점을 타개할 해결방안 찾기, 큰 부분은 모방해도 되지만 독창성 있는 아이디어로 나의 연구를 할 것
    - Contextual word embeddings for tabular data search and integration 저자가 올린 소스코드 설치 및 실행
    - SANTOS: Relationship-based Semantic Table Union Search 읽고 정리
* **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**
  + **데이터 융합 프로세스**



* **융합 가능도 척도**
  + 임베딩 거리 기준 값을 정하고 기준값보다 임베딩 거리가 가까우면 융합 가능한 데이터셋이라 판단함
  + 융합 가능한 데이터셋을 사람이 하나씩 확인하여 유의미한 융합이 되는 데이터셋을 찾음
  + 을 융합 가능도 척도라 정의함



혼합형 데이터셋 예시

등급의 semantic

융합 가능도 딥러닝 아키텍처 : 임베딩 벡터 + 메타데이터(융합 가치를 정량화할 수 있는 무언가)   
조인할 수 있는 컬럼이 필요함(A : 주민번호, B : 주민등록 번호로 조인이 가능한 테이블이면?) 융합을 하는 건 가치가 있기 때문, 가치가 있다? -> 임베딩이 가깝게 되어있다.

조인할 수 있다, 없다 -> 사람이 판단

이걸 조인해서 가치있는 정보가 나오나? -> 모델이 판단

Id를 가지는 컬럼이 의미는 같지만 겉모양이 다를때

그래프 임베딩

정형테이블을 그래프로 바꾸는 논문읽기

그래프 정보를 임베딩한 정보와 기존 윤종찬 연구원과의 차이점 발견

나만의 새로운 패러다임 찾기

데이터셋은 영문 데이터만 사용

융합 가능성 평가는 임베딩 성능을 확보한 후에

테이블 융합의 연산은 Union / Join이 있음 -> Union / Join 을 구분해줄 수 있는 모델이 있으면 좋겠음.

Ex. Output 값이 1이면 Union 0이면 Join

조인키만 있어도 조인이 되는게 아닐 수 있음. 나중에 코드를 짤 때 조인(유니온)이 가능한 예제를 sql로 조인해보고 조인이 가능한 예제를 모델에 넣어서 조인이 가능하다고 판독이 되는 것을 확인해보기