**Weekly report**

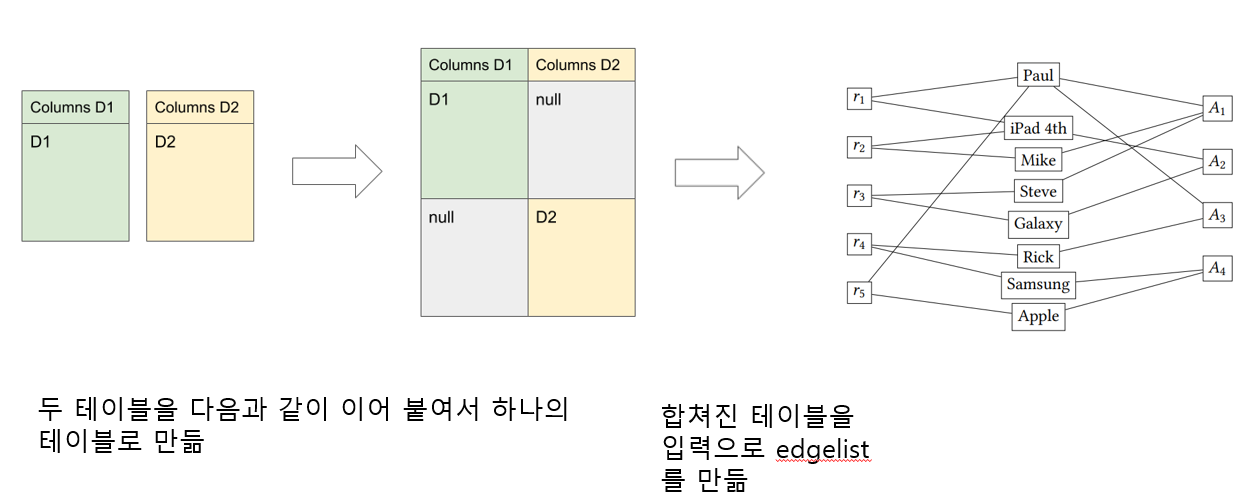
날짜 : **2023-02-09**

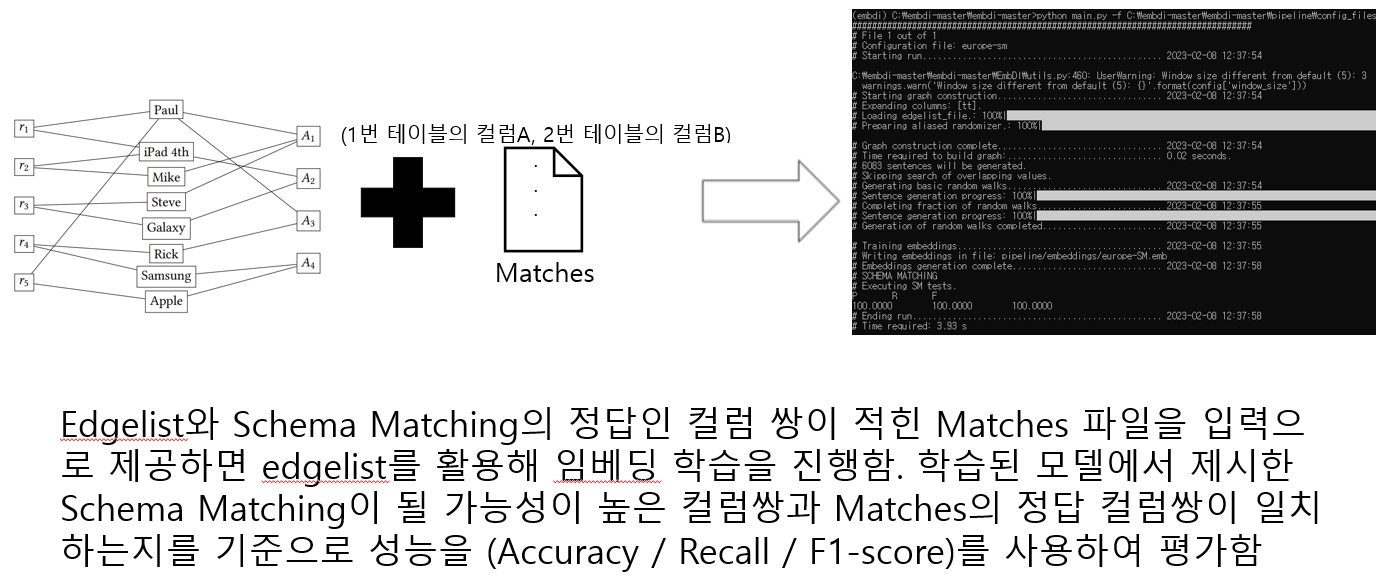
연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - 수치형 데이터를 이산화한 것을 범주형 데이터로 바꾸는 임베딩 방식을 보완하는 방법을 찾고 실험하여, 기존의 윤종찬 연구원의 연구와 **융합 가능도 척도**를 사용하여 성능 비교
    - 각 컬럼마다의 중요도를 고려해서 임베딩하는 방식과 수치형 컬럼에서 숫자의 크기가 크고 작음에 부여된 의미를 고려하는 임베딩 방식 연구
    - 나만의 임베딩 아키텍처 / 융합 가능성 아키텍처(임베딩 아키텍처에 dependant한) 그려보기, 기존의 것과 나만의 것을 분리하여 정리 (PPT로)
    - 모델의 성능이 안 좋은 이유 찾고 해결. (데이터의 문제 / 임베딩 방식의 문제) – 데이터의 문제 : 융합 가능한 라벨들을 찾아서 추가
    - 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)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각해보기
    - EmbDI 코드에 커스텀 데이터를 입력하여 임베딩 및 성능 확인
  + **GAN 논문 세미나 :**

**Pix2Pix | Image-to-Image Translation with Conditional Adversarial Networks** 논문 읽고 세미나 준비

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + EmbDI 코드에 커스텀 데이터를 입력해 임베딩 및 성능 확인
    - EmbDI는 두 가지 작업을 수행함
      * Entity Resolution과 Schema Matching
      * 다만 데이터 융합이 목적이므로 Schema Matching 파트만 활용함
    - EmbDI 모델이 Schema Matching을 하는 과정





* + - Europe.csv : 유럽의 국가와 각 국가별 면적과 인구수 정보가 있는 테이블
      * Europe.csv 테이블을 두 번 사용하여 Schema Matching 진행

텍스트이(가) 표시된 사진

자동 생성된 설명

* + - TripAdvisor European restaurants.csv : 유럽 전역에 있는 레스토랑 정보가 있는 테이블
      * Europe.csv랑 Schema Matching 진행 (Country name 기준)
      * 파일 크기가 크면 작동이 안됨 입력 csv 파일 500메가 기준 -> 에러 발생
    - USA-lation : 미국의 zipcode와 각 zipcode별 위도, 경도 정보가 있는 테이블
    - Pizza\_Restaurants\_and\_the\_Pizza\_They\_Sell : 미국 전역 피자집 이름, 위도, 경도 등의 정보가 있는 테이블

텍스트이(가) 표시된 사진

자동 생성된 설명

* + - 위도 경도 컬럼을 Schema Matching으로 찾아내지 못함, 이는 Local Embedding의 특성상 두 테이블에서 완전히 겹치는 값이 없으면 두 컬럼의 관계성을 잘 나타내지 못해서 이런 현상이 나타난 것으로 보임.
    - 대부분의 수치형 데이터는 완전히 겹치기 어렵기에, 수치형 데이터를 정수의 형태로 바꿔서 한 번 더 실험을 진행할 예정
    - EmbDI의 임베딩 방식은 사람이 보고 융합 가능성을 평가하는 것과 닮아 있다고 생각함(융합 가능성 평가를 할 때도 각 테이블에서 의미가 비슷한 두 컬럼을 본 후 각 컬럼에 포함된 값들이 겹치는지를 확인함)
      * 기존의 아이디어인 임베딩 + 융합 가능성 아이디어는 기준이 사람마다 주관적이라고 생각되는 단점이 있음. 이를 우리의 임베딩 후 코사인 유사도 계산하여 일정 수준 이상이 나온 테이블들만 추출하고 EmbDI를 사용하는 방식을 생각함
      * 우리의 임베딩을 하고 EmbDI을 하는 이유는 EmbDI은 우리의 임베딩보다 매우 많은 시간이 걸리기 때문임.
* **GAN 논문 세미나 :**
  + **Pix2Pix | Image-to-Image Translation with Conditional Adversarial Networks** 논문 읽고 발표 준비
* **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)  
      -> Deep Entity Matching: Challenges and Opportunities (Pretrained된 language model과 딥러닝을 활용해 Entity Matching을 진행)
    - 테이블 간 조인 / 유니온 하기위한 컬럼을 찾는 문제에 관해 생각 (Elmo / Bert / Word2Vec) 등도 생각 / 내가 손으로 하는 것을 자동화하는 방법으로 생각해보기
    - EmbDI 수치형 데이터가 완전히 겹치지 않는 경우에 정수로 바꿔서 추가 실험
    - EmbDI는 테이블 수준의 임베딩인데, 스키마 매칭이 가능한 이유 공부하기
    - SemProp(Attribute overlap + Value overlap + Embedding) 관련 논문 읽고 공부
  + **GAN 논문 세미나 :**
    - **StyleGan | A Style-Based Generator Architecture for Generative Adversarial Networks** 논문 읽고 세미나 준비 (PPT 30분 이내 발표) , 각 논문의 핵심사항을 알아내는게 목적. 연구 동기, 배경, 관련연구 등은 생략하고 이해한 범위 내에서 핵심사항 중심으로 발표 진행. 성능 평가는 한 페이지(메인 그래프만)로 충분함. 그림과 키워드 위주로 발표. 예제 위주 이해 위주 친절한 자신만의 이해한 언어로 진행
* **기타사항**
  + **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