**Weekly report**

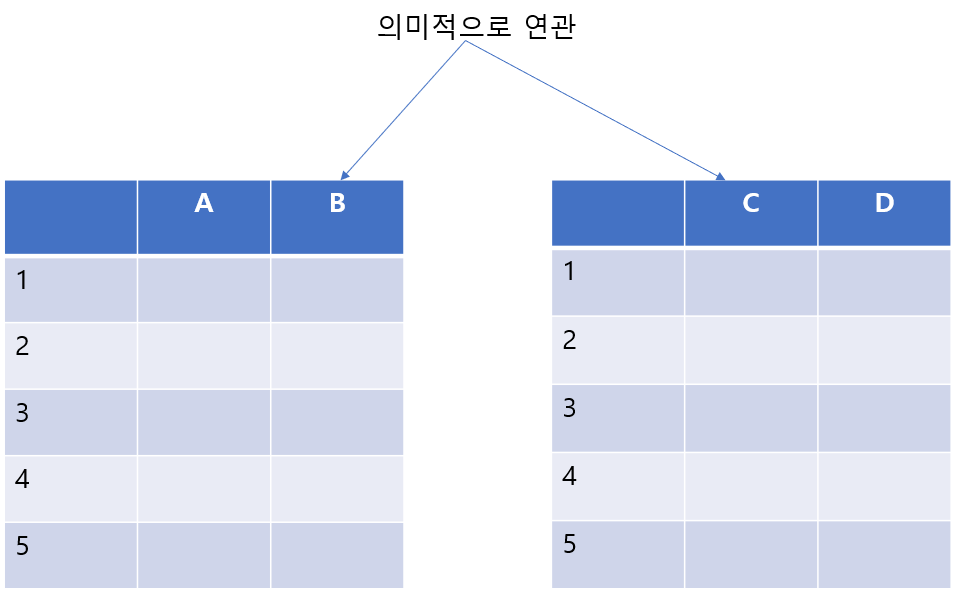
날짜 : **2023-02-16**

연구원 : **최윤석**

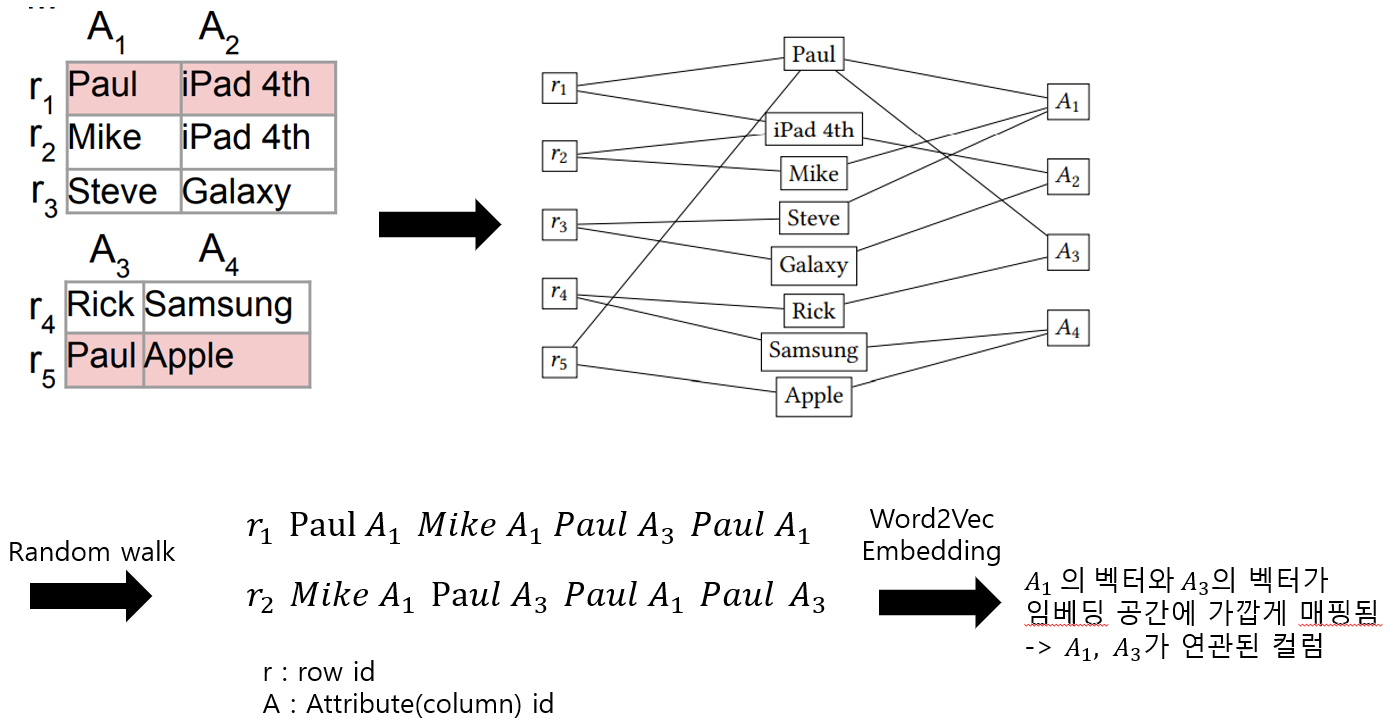
* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - Value overlap + Semantic overlap + Embedding(EmbDI) 을 사용하여 테이블 융합이 가능한지 확인하는 나만의 아키텍처 연구
    - Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기 (Embedding + Semantic Overlap(외부 knowledge base)을 섞는 방식 관련해서)   
      -> SANTOS: Relationship-based Semantic Table Union Search (외부 Knowledge Base를 사용하여 Table Unionability를 알아내는 알고리즘 제시)  
      -> 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

* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + EmbDI가 스키마 매칭이 가능한 이유
    - Schema matching은 두 테이블에서 의미적으로 연관된 컬럼 쌍을 찾아내는 것



* + - 기존의 Schema matching 기법은 컬럼명이 비슷하거나 각 컬럼에 속한 데이터 타입이나 분포가 비슷한지를 기준으로 의미적으로 연관된 컬럼 쌍을 찾아냄.
    - EmbDI는 두 컬럼이 의미적으로 연관되어 있는 것을 다음과 같이 알아냄.



* + - Random walk를 진행할 때 Attribute id도 문장에 나타나기에, 같은 값(Paul)을 공유하는 컬럼들(A1, A3)은 문장에서 자주 같이 나오게 됨.
    - 이러한 문장을 입력으로 임베딩을 진행하게 되면 A1, A3는 임베딩 공간에 가깝게 매핑되어, 의미적으로 연관된 컬럼이라고 판단할 수 있음
    - 컬럼명(Attribute id)을 임베딩하여 두 컬럼이 연관되었는지를 판단할 수 있기에 Schema matching이 가능함.
  + EmbDI 수치형 데이터가 완전히 겹치지 않는 경우에 정수로 바꿔서 추가 실험
    - 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.
    - 실제로 Pizza\_Restaurants\_and\_the\_Pizza\_They\_Sell 테이블의 위도 컬럼에 속한 값들 중 USA-lation의 위도값과 완전히 겹치는 값은 하나도 없었음. 경도값도 마찬가지로 완전히 겹치는 값이 하나도 없었음.
    - 위도와 경도 컬럼에 있는 값들을 정수로 바꾸었더니 95%이상의 값들이 다른 테이블의 위도, 경도 값 중 하나 이상과 서로 완전히 겹침

검은색, 기타이(가) 표시된 사진

자동 생성된 설명

* + - 실수를 정수로 바꿔 진행해 보았지만, Schema Matching으로 찾아내지 못함.
  + 깃허브에 관계형 테이블 임베딩 관련 코드, 데이터셋, 참고 논문 탑재 완료
  + SemProp 논문 읽는 중
    - 논문에서 이해가 안되는 부분이 있어서 마저 이해를 진행하고 다음주에 논문 내용을 정리할 예정.
* **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
    - SemProp(Attribute overlap + Value overlap + Embedding) 관련 논문 이해하고 정리
  + **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