**Weekly report**

날짜 : **2023-01-26**

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
    - Data Integration and Machine Learning: A Natural Synergy 읽고 공부, Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks을 인용한 논문 중 연구와 관련된 것 읽어보기, Valentine: Evaluating Matching Techniques for Dataset Discovery 읽고 공부
  + **GAN 논문 세미나 :**
    - **Towards the Automatic Anime Characters Creation with Generative Adversarial Networks** 논문 읽고 세미나 준비
* **수행결과**
* **관계형 테이블 임베딩** : (임베딩 / 융합 가능성 파트 분류해서 작성)
  + Data Integration and Machine Learning: A Natural Synergy 논문 읽고 공부
    - Data Integration(DI) 과 Machine Learning(ML) 기술이 합쳐졌을 때 어떤 시너지가 나는지 간략하고 다양하게 설명하는 논문.
    - 현재 ML이 Entity Resolution / Data Fustion / Data Extraction / Schema Alignment등의 DI 분야에서 좋은 활약을 하고 있고 DI도 큰 스케일의 데이터셋 생성과 Data Cleaning 분야에서 활약해 ML 모델 성능을 높이는데 기여하고 있는 것을 간략하게 설명함.
    - 추가적으로 앞으로 DI와 ML이 추가적으로 Multi-modal DI와 Data Augmentation 에서도 시너지가 날 것임을 시사함.
    - 다만 관계형 테이블 임베딩 연구와 관련해 직접적으로 도움이 되는 부분은 없었음
* **GAN 논문 세미나 :**
  + **Towards the Automatic Anime Characters Creation with Generative Adversarial Networks 논문 읽고 발표 자료 작성**
* **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, Valentine: Evaluating Matching Techniques for Dataset Discovery 읽고 공부
  + **GAN 논문 세미나 :**
    - **Pix2Pix | Image-to-Image Translation with Conditional 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