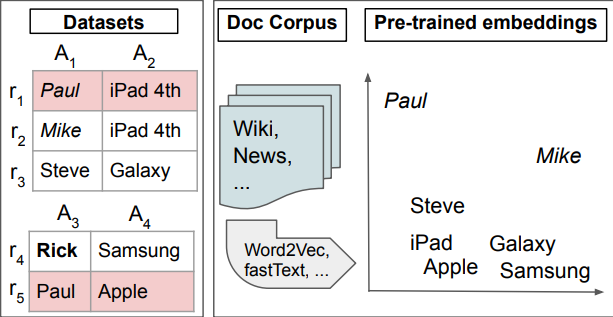
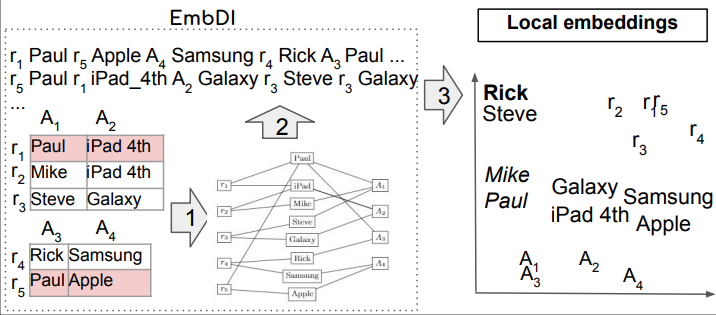
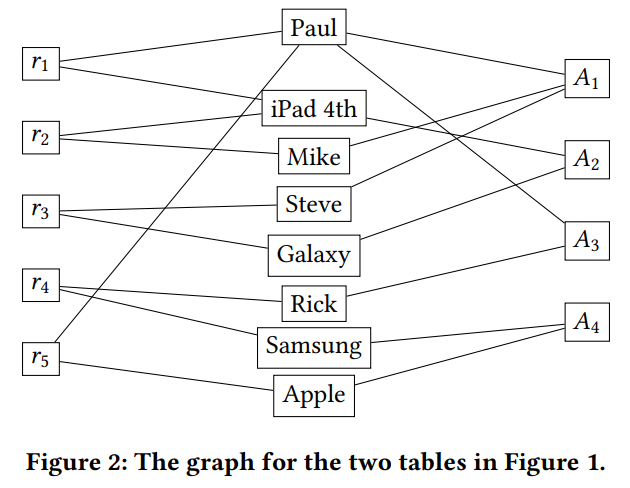
**Weekly report**

날짜 : **2023-01-20**

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 , 임베딩 방식의 문제 : 이전 논문에서 읽었던 Field-wise Network 방식을 적용
    - 연구재단 보고서 작성(1/18까지 교수님께 메일로) 최대한 빨리
    - 데이터셋 융합 관련 논문 2편 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)
  + **[임베딩] :** Creating Embeddings of Heterogeneous Relational Datasets for Data Integration Tasks 논문 공부
    - Pre-trained embedding을 사용하면 주어진 데이터셋에 맞는 의미적 정보를 얻을 수 없음
    - 논문에서 제시하는 Local embedding을 사용하면 주어진 데이터셋에 맞는 의미적 정보를 학습할 수 있음  
      
    - 먼저 주어진 데이터셋들에 Row와 Column에 인덱스를 붙여 RID와 CID를 생성함
    - RID와 CID 그리고 테이블 셀 안의 값들(Token)을 중복을 제하고 전부 노드의 값으로 하여 노드들을 생성
    - Token이 속해있는 RID, CID에 엣지들을 생성함
    - 한 노드에서 출발해 정해진 수만큼 Random Walk를 진행하여 만나는 노드에 있는 값들을 전부 적어 하나의 문장으로 만듦
    - 출발지 노드를 모든 노드로 해서 노드의 수만큼 문장을 얻음
    - 얻은 문장들을 입력 데이터로 활용해 Word2Vec이나 FastText같은 기존의 임베딩 모델을 사용해 Local embedding을 진행
    - 이를 통해 특정 데이터셋에 specific한 임베딩을 얻을 수 있음
  + 연구재단에 제출할 관계형 테이블 융합 보고서 초안 작성 및 교수님께 전달
* **GAN 논문 세미나 :**
  + **Towards the Automatic Anime Characters Creation with Generative Adversarial Networks 논문 공부**
* **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 , 임베딩 방식의 문제 : 이전 논문에서 읽었던 Field-wise Network 방식을 적용
    - 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 (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.
* **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