**Weekly report**

날짜 : **2022-09-29**

연구원 : **최윤석**

* **수행결과**
* **차량 이미지 분류 :** 
  + 텍스트이(가) 표시된 사진

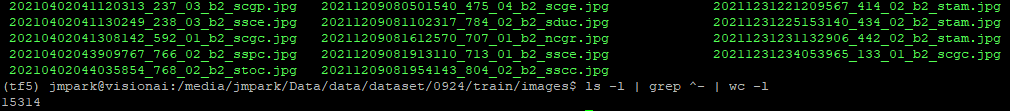
    자동 생성된 설명트라콤에서 새로 보내준 차량 이미지를 카테고리별로 1000개씩 추출하여 19000여개의 이미지를 얻고, 이를 Train : Test , 8 : 2의 비율로 나눔.  
    Train 15314개 / Test 3830개 이미지
  + 새로운 이미지로 SSD 학습에 필요한 파일 생성
    - 이미지 사이즈를 2048 \* 1536 에서 YOLO에서 학습할 때 사용하는 416 \* 416으로 리사이징
    - 이미지의 바운더리 박스정보가 적힌 txt 파일을 SSD 학습을 위해 xml 파일로 변환
  + SSD 모델 학습
    - 학습 중 오류가 발생하여 해결 중
* **관계형 테이블 임베딩** :



Figure 1. 혼합형 데이터셋 예시

* + 이산화한 것을 범주형 데이터로 (very high(5) / high(4) / medium(3) / low(2) / very low(1)) 변경하는 알고리즘에서 임베딩을 진행했을 때, 의미적으로 연관성이 없는 테이블들이 임베딩 공간에서 가깝게 위치함.

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 2. 이산화한 것을 범주형 데이터로 변경

* + - 실제로 Cars – Seattle – Food\_Production 테이블이 임베딩 공간에서 가까이 위치함
    - 하지만 실제 테이블 데이터를 보면 전혀 유사하지 않은 3개의 테이블임.
    - 테이블의 컬럼 대부분이 수치형 컬럼인 테이블들이 이산화되면 레코드내용이 이산화된 5개의 값으로 한정되기에 의미적으로는 전혀 달라도 비슷한 레코드 임베딩 벡터를 가짐.
    - 최종 임베딩 벡터는 전체의 60%를 레코드 임베딩 벡터에 의해 영향을 받기에, 크게 관련이 없는 테이블들이 벡터공간에 비슷하게 위치하게 되는 것으로 생각됨.

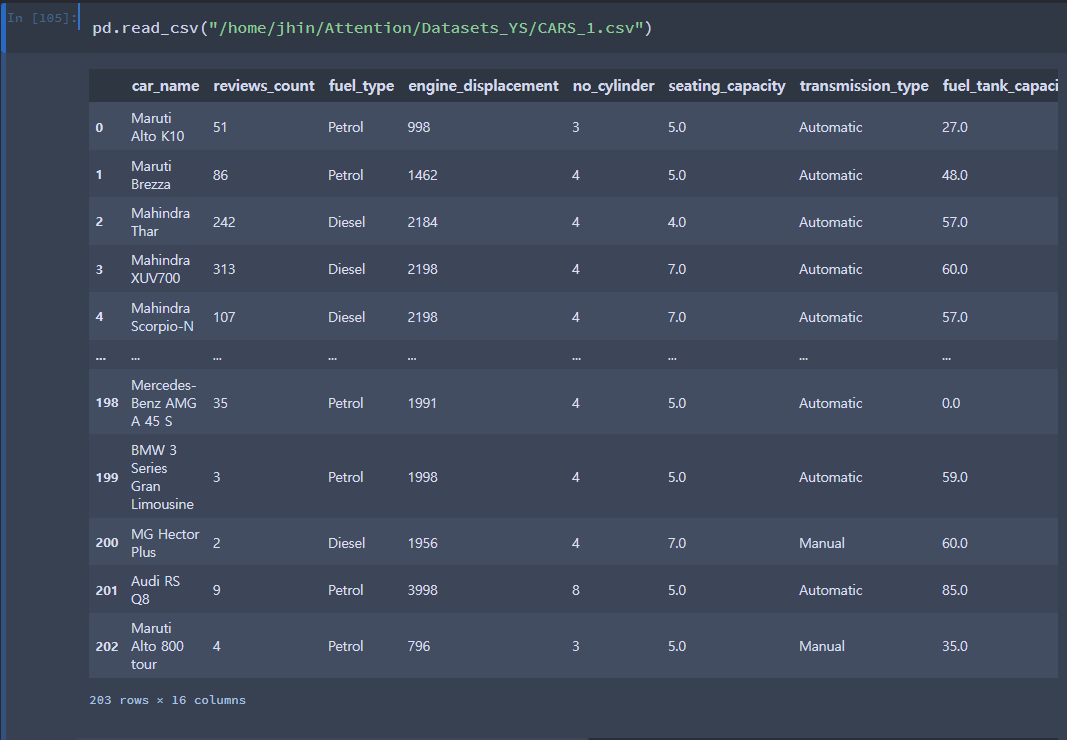


Figure 3. CARS\_1.csv 신차에 관한 평가 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트, 전자기기, 스크린샷, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

Figure 4. seattle\_01.csv 숙박 공유 사이트 Airbnb의 Seattle에서 일어난 숙박 공유 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 5. Food\_Production.csv 특정 음식 생산이 환경에 미치는 영향을 정리한 데이터

텍스트이(가) 표시된 사진

자동 생성된 설명

* + 이러한 문제를 개선하기 위한 방법으로 불필요한 수치형 컬럼을 지우는 작업을 진행.
    - 숫자의 대소가 아무 의미가 없는 ID 컬럼들을 지우는 코드를 추가
  + 텍스트이(가) 표시된 사진

    자동 생성된 설명텍스트이(가) 표시된 사진

    자동 생성된 설명코드를 추가한 이후 다시 임베딩한 결과 Cars – Seattle – Food\_Production 이 임베딩 공간에서 멀어지게 매핑됨

Figure 7. ID 관련 컬럼을 지운 후 임베딩한 결과

Figure 6. ID 관련 컬럼을 지우는 코드를 추가

* **수행계획**
* **차량 이미지 분류** : 학습 도중 발생하는 오류 해결 및 모델 학습
* **관계형 테이블 임베딩** :
  + 이산화한 것을 범주형 데이터로 바꾸는 임베딩을 보완하는 방법 찾고 실험
  + NDCG 계산에 필요한 Dataset\_rel.csv 파일 생성 및 NDCG로 임베딩 모델의 정량적 평가 진행
* **기타사항**
  + **데이터 융합 프로세스**

