아하모멘트

현업에서 겪었던 첫 데이터 분석 경험

강사 : 김진용

입사 전 경험

비전공자에서 AI, 빅데이터에 입문하다

대학 4년, 기계 공학, 4대 역학 ~> 이걸 배워서 어디에 써먹지? ~> 대학원?! ㅠ.ㅠ

3학년 때, 학부 연구생 ~> 드론 자율 주행 연구 ~> 프로그래밍에 대한 굉장히 재밌었던 경험(간단한 코드 한줄로 드론 조정)

네이버 부스트캠프 AI Tech 2기 시작!

입사 전 경험

네이버 부스트캠프 AI Tech 2기

- 약 5개월 간의 고군분투
- 공부 시간 : 10 to 19 -> 8 to 22
- AI 경진 대회를 통한 데이터 EDA, 시각화, 모델링
- 각 대회에서 필요한 논문 스터디

Naver Boost Camp AI TECH 2기

인공지능, 딥러닝 활용 방법과 AI Production의 End-to-End 교육 프로그램

Image Classification Competition 21.08.23~21.09.02	• 이미지 분류 모델을 활용한 마스크 착용 / 성별 / 나이에 대해 총 18개의 클래스로 나눠 판별하는 시스템을 구현하는 대회 • 이미지에서 사람 얼굴을 탐지 및 크랍하여 모델의 성능 개선
Object-Detection Competition 21.09.27~21.10.15	• Object detection 모델을 활용해 쓰레기를 탐지하고 10종류의 쓰레기 class를 detection하는 시스템을 구현하는 대회 • Heavy Augmentation, 앙상블 기법을 적용하여 모델의 성능 개선
Semantic Segmentation Competition 21.10.18~21.11.04	• Segmentation 모델을 활용해 쓰레기를 탐지하고 10종류의 쓰레기 class를 segment 단위로분류하는 시스템을 구현하는 대회 • 재활용 쓰레기를 copy paste하여 모델의 성능 개선
글자 검출 대회 21.11.08~21.11.18	• 글자 검출 모델을 활용해 글자 검출 시스템을 구현하는 대회 • 오라벨링된 데이터 검수 후 삭제하여 모델의 성능 개선
최적화 대회 21.11.22~21.12.02	• AutoML, Quantization 등을 활용한 모델 경량화 및 최적화 대회 • Optuna를 활용해 100번의 모델 실험하여 모델의 성능 개선
모델 서빙 21.11.08~21.12.27	• "간편한 식단 관리 Food Log 어플리케이션" 개발 • 데이터 전처리, Heavy Augmentation, fast API로 GCP에 모델 배포

입사 전 경험

현 직장 취업 성공!

- 기업 네트워킹데이
- 총 3번의 면접
- 부스트캠프 종료 후 한달 만에 취업

목차

1. 입사 후 겪었던 난항

1. 입사 후 겪었던 난항

1.1 나 홀로 팀

- 국방부 사업 단독 개발자
- 업무 경험이 처음인 나
- 빅데이터 경험이 전무했던 국방부 PM님과 나
- 입사해보니 중간보고 일정이 한달 지나있는 상태
- 개발 현황
  - 지난번 사업 때 전임자가 사용했던 일부 코드 존재
  - 분석해야할 이미지, 3D 데이터만 적재되어 있는 상황
- 개발 목표
  - 2달 안에 AI로 분석한 빅데이터 결과물 산출하기

1. 입사 후 겪었던 난항

1.1 나 홀로 팀

AI Training & Inference

- 데이터에 대한 이해(라벨링 방식, 학습 방식, 추론 결과 등)
- AI가 학습할 수 있도록 데이터 가공
- 만족할만한 정확도가 나올 때까지 AI 모델 학습
- 수집된 데이터에 대해 AI 분석 후 원하는 형태로 가공
- 매달마다 추가적으로 수집되는 데이터에 대해 학습

Database 관리

- 데이터 CRUD
  - AI 분석을 하기 위해 필요한 데이터 구축
  - AI로 분석된 데이터 관리
- API 서버 구축
  - AI 분석 서버와 DB간 필요한 데이터를 주고 받을 수 있는 서버 구축

BigData 분석

- 데이터가 기대되는 값으로 잘 도출이 되는지 확인
- 문제 발견 시 원인 파악
- 통제 가능한 문제일 시 해결방안 마련하여 문제 해결
- 데이터 분석하여 시사점 파악

1. 입사 후 겪었던 난항

1.2 AI 분석

AI Training & Inference

- 학습 코드 Readme 파일의 부재 -> 코드 한줄 한줄 뜯어보기
- 사내에 있는 AI 학습을 위한 데이터 가공 API 에러 -> 데이터에 대한 심도 있는 이해
- Multi GPU 처음 사용 -> 학습
- 분석 정확도가 떨어지는 케이스 발생 -> 원인 파악 및 재학습

1. 입사 후 겪었던 난항

1.2 DB 관리

Database 관리

- Database MySQL **처음 접함** -> DB Query/Procedure 개념 학습 및 Dev DB로 테스트
- Database의 각 테이블에 대한 **메타 데이터가 없어** 어떤 데이터인지 파악하기 어려웠음 -> 이전 사업 때 결과 산출물과 코드를 보며 파악, Database를 구성했던 개발자에게 문의
- **한글, 엑셀에 있는 표**를 csv 파일로 **파싱** -> 전임자 작업 코드 인계받아 처리
- 데이터의 품질이 낮은 경우 -> 알고리즘을 개발해 로직을 통한 품질 검사 실시
- **대량의 데이터**(10만 row) 추가나 업데이트 시 몇 시간 소요 -> 대량의 데이터를 한번에 처리할 수 있는 로직으로 개선
- fastAPI 처음 써보며 RESTful API 개념을 잘 몰랐음 -> 사내에 구축되어 있는 API 서버 참고하여 get, post 함수 구현

1. 입사 후 겪었던 난항

1.3 BigData 분석

BigData 분석

- 빅데이터 결과물을 내기까지가 **굉장히 리소스가 많이** 들어서 기한 내에 기대한 대로 잘 나오는지 확인을 못했음 -> 국방부 사업 PM님과 결과를 직접 확인해보며 계산식대로 나왔는지 검수
- 기대한 대로 나오지 않을 경우 원인이 코드 오류라면 수정하기 쉬웠지만 코드 원인이 아닐경우 **원인 파악에 애를 먹었음** -> 해당 결과로부터 거꾸로 파고 들어가 왜 이런 결과가 나왔는지 구체적으로 파악할 때까지 분석(데이터에 대한 심도 있는 이해하는 과정)
- 원인을 파악했음에도 **문제 해결이 쉽지 않은 경우가 있었음.** -> 새로운 알고리즘이 필요한 경우 리소스가 많이 들지 않는 한에서 개선의 가능성이 있다면 개발, 알고리즘의 한계를 극복하기 어려운 상황이라면 필터링 등을 활용해 보완

2. 실전 데이터 분석

2. 실전 데이터 분석

2.1 문제의 원인을 찾아라

5 whys?

- 제퍼슨 기념관의 대리석은 왜 빨리 부식할까?
  - 대리석을 비누로 자주 씻기 때문
- **왜** 대리석을 비누로 자주 씻을까?
  - 비둘기 배설물 때문
- **왜** 비둘기가 많을까?
  - 비둘기의 먹이인 거미가 많기 때문
- **왜** 거미가 많을까?
  - 거미의 먹이인 나방이 많기 때문
- **왜** 나방이 많을까?
  - 황혼 무렵에 점등되는 기념관 불빛 때문

-> 황혼 무렵에 점등되는 기념관 불빛을 2시간 늦게 점등하여 제퍼슨 기념관 대리석 부식을 막을 수 있었습니다.

2. 실전 데이터 분석

2.1 문제의 원인을 찾아라

5 whys?

- 5번을 질문하는 것이 중요한게 아닙니다.
- **어떤 현상에 대한 근본적인 원인을 파악**하는데는 단 2번의 질문으로도 가능할 수 있습니다.
- 사람의 분석 능력이나 해당 분야의 노하우를 얼마나 알고 있느냐 등에 달려있습니다.

2. 실전 데이터 분석

2.1 문제의 원인을 찾아라

현업에서 겪었던 문제

- **왜** 국의 섭취율이 기대하는 것보다 낮게 나올까?
  - 특정 끼니의 국의 섭취율이 낮게 나와 전체 섭취율을 낮춥니다.
- **왜** 특정 끼니의 국의 섭취율이 낮게 나올까?
  - 배식 받은 국의 양과 퇴식 때의 국의 양 결과값이 비슷하기 때문입니다.
- **왜** 실제 사진에서는 배식과 퇴식 때 받은 국의 양이 차이가 있는데 결과값은 비슷하게 측정이 될까?
  - 카메라가 설치된 곳 위쪽에 형광등이 있어 국에 빛반사가 있었고 3D 데이터에 결측값이 발생했기 때문입니다.

-> (예상) 결측값을 보정하면 국의 섭취율이 어느 정도 기대하는 바대로 나올 것이다!!

2. 실전 데이터 분석

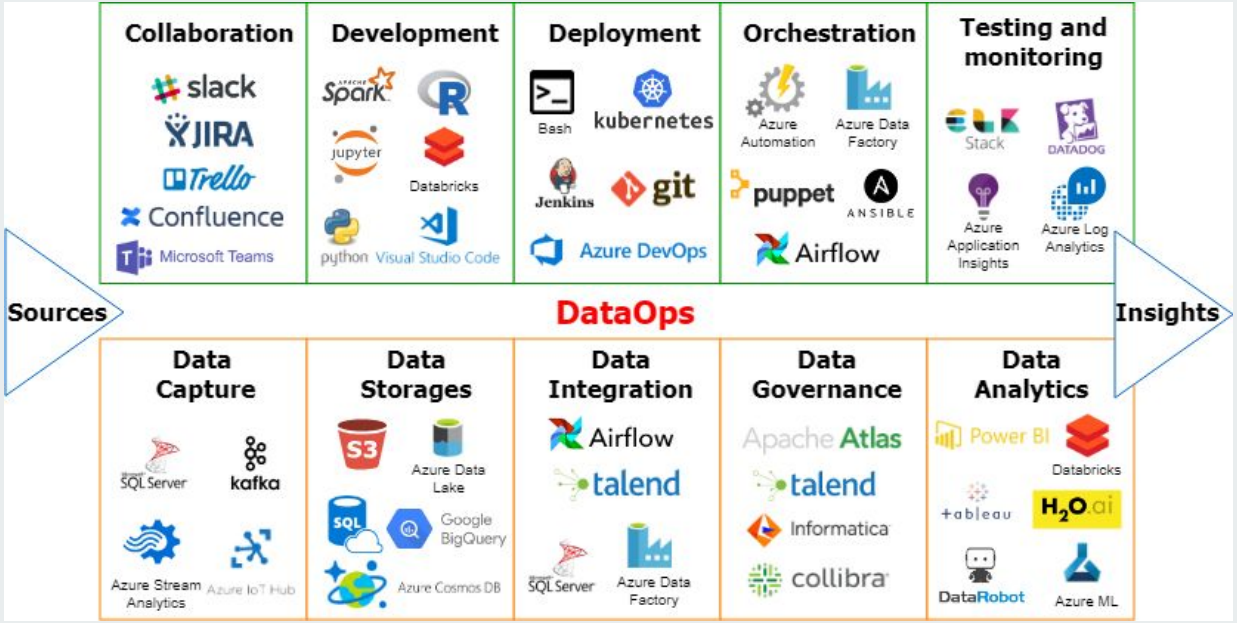
2.2 노이즈가 있는 3D 데이터 분석하기

노이즈가 있는 3D 데이터 분석하기

3. 주니어로 입사 전 알면 좋을 것들

3. 주니어로 입사 전 알면 좋을 것들

3.1 DataOps



3. 주니어로 입사 전 알면 좋을 것들

3.2 Database

MySQL

3. 주니어로 입사 전 알면 좋을 것들

3.3 Github

- git cheat sheet
- 배달의 민족

Q & A