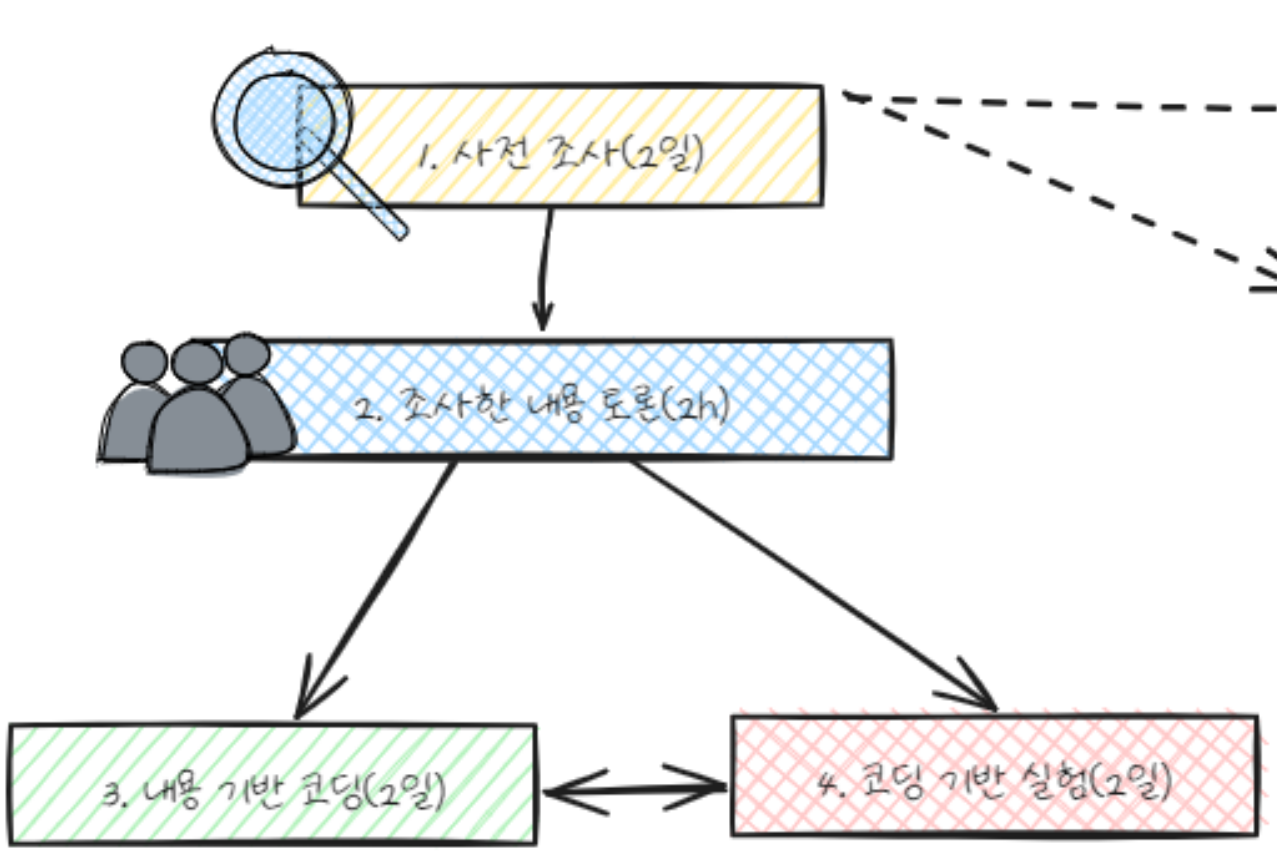
Datacentric week11 주간 회고록

NAVER Boostcamp AI Tech CV-21

김한얼 김보현 김성주 윤남규 정수현 허민석

Mon



강의 듣기

6강, 9강

6강(EDA) 한사람 당 60~70장 영수증리뷰

[이미지를 직접 관찰하는 시간을 가지기]
명도가 낮다
깔끔하다(표기x)
대리석 위에 놓여있다
회전 10도 되어있다

9강 보고 가이드라인 잡기

Tue

EDA_train

EDA_test

0	extractor.zh.in_house.appen2_001047_page0001.j	QR코드나 바코드가 존재함		보현
1	extractor.zh.in_house.appen2_001050_page0001.j	확대해도 흐릿함	박스가 겹침	보현
68	extractor.zh.in_house.appen_000700_page0001.	그림자가 존재함	QR코드나 바	성주
69	extractor.zh.in_house.appen_000701_page0001.	확대해도 흐릿함		성주
135	extractor.ja.in_house.appen_000309_page0001.j	그림자가 존재함		한얼
136	extractor.ja.in_house.appen_000316_page0001.j	그림자가 존재함	텍스트가 로고나 Q	한얼
236	extractor.th.in_house.appen_000368_page0001.j	텍스트가 로고나 QR코드 안...		남규
237	extractor.th.in_house.appen_000377_page0001.j	그림자가 존재함	박스가 잘못됨	남규
267	extractor.th.in_house.appen_000660_page0001.j	스캔본임		수현
268	extractor.th.in_house.appen_000666_page0001.j	그림자가 존재함	QR코드나 바	수현
398	extractor.th.in_house.appen_000184_page0001.j	스캔본임		민석
399	extractor.th.in_house.appen_000110_page0001.j			민석

데일리 스크럼에서 Annotation rule을 세웠다.
전원 10강까지 다듣고 train,test 이미지를 인당 80장씩 나누어서 EDA를 했다.

Wed

2.데이터셋을 어떻게 쓸지

(public-ufo로 가져오기, synthetic)

public: 언어 비중에 대한 EDA만 해서 몇개 있는지 파악해서 가져오

synthetic: 문서를 이미지파일로 저장해보고싶다..

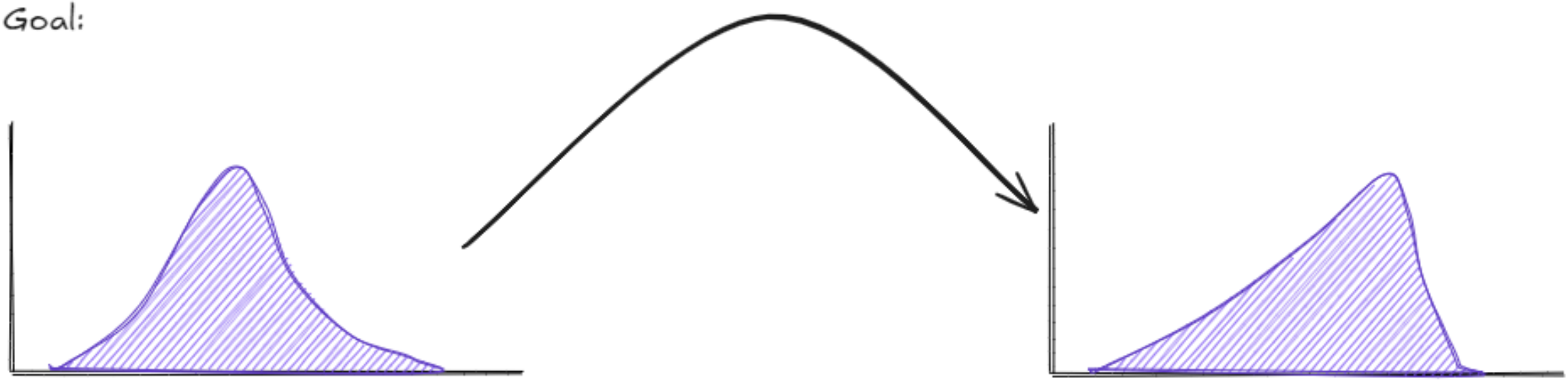
🙋‍♀️영수증에 들어갈 단어를 어떻게 생성할지

3.증강을 추가적으로 어떻게 할지 고민(Test Data 보기)

- 깨끗하게
- 더럽게

Thu

Goal:



Methods

- 10~15도 회전
- 그림자 부여 또는 밝기 조절
- Normalization: 테스트 데이터의 평균과 분산에 맞춰야 함

Thu

Methods

6명이 공통으로 써봄직한 방법

- 10~15도 회전
- 그림자 부여 또는 밝기 조절
- Normalization: 테스트 데이터의 평균과 분산에 맞춰야 함
- 가우시안 노이즈, 솔트앤페퍼vs디노이즈?
- 배경을 추가 vs 배경을 제거?
- Superesolution vs 일부러 Blur?
- 잘린 글씨
- 작은 이미지 그냥 확대

Thu

6명이 각자 써봄직한 방법

- 학습을 언어 종류 별로 진행하기
  - 중국어:
  - 일본어: 가로 선을 제거??
    세로쓰기가 없으면 세로데이터 제거
  - 태국어:
  - 베트남어: 스캔본이 많고 공백이 있으므로 padding

} 결과별로 CSV 통합

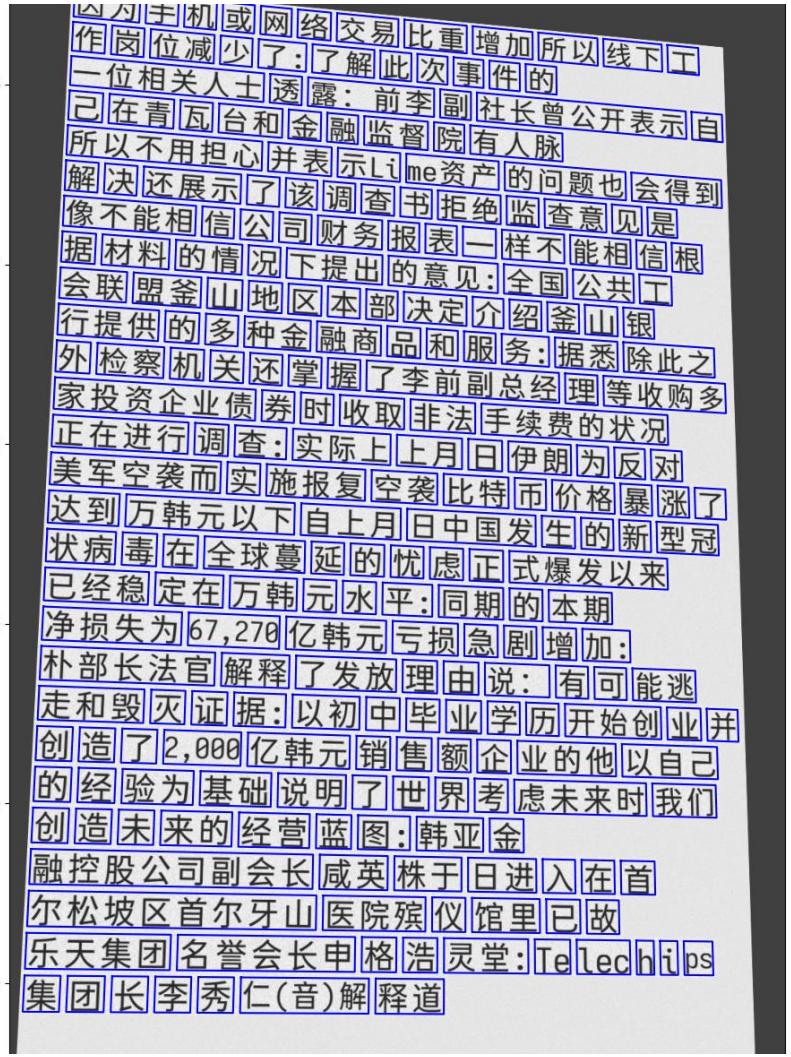
- 공개 데이터 가져오기
  - 퍼블릭 데이터
  - 데이터 크롤링: 슈도 라벨링 사용
- 합성 데이터 만들기: 휘어진 글자 추가, 회전된 글자 등

} Ensemble

그 외

- 특수한 데이터 처리 고민: 흰색 줄 처리, 보안코드 처리, 분홍 선, 투명 배경, 붉은 영수증 접힌 자국. --oOo--, 손가락, 영수증 외부 글씨, 인식되지 않은 경계선? 등
- 앙상블: 앙상블도 데이터를 보면서 잘 해보려는 시도

Fri



수현님의 공공데이터 해체show

감사합니다.