

Part 1

트러블 슈팅 (Issue & Solution)

문제 정의 :
이미지 데이터와 어노테이션 정보 간
미 매칭

원인 추론
: 케글 제공 데이터 자체가 미 매칭

- 조사 및 해결 과정:
- 원본 데이터에서 미매칭 데이터 선별
 - AI 허브를 통한 어노테이션 재작업 진행
 - 미매칭 외 bbox 겹침 등 추가 이슈 발견 → 데이터 전수 수정

결과:
데이터 품질 및 신뢰도 확보 완료



Part 2 데이터 통계

▶ 전체 통계

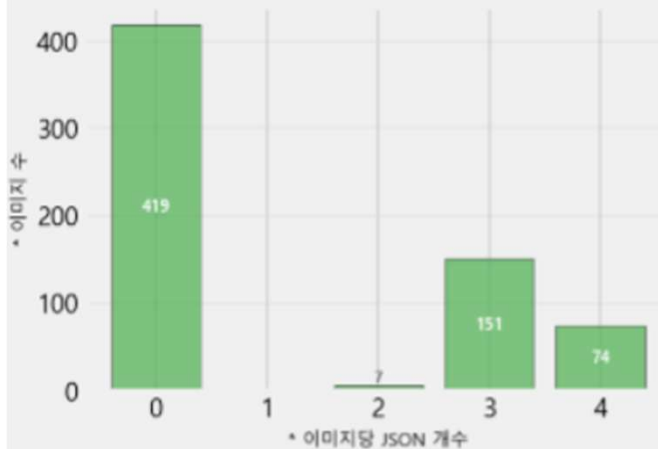
■ 전체 통계 요약

- * 총 이미지: 651개
- * JSON 있는 이미지: 232개
- * JSON 없는 이미지: 419개

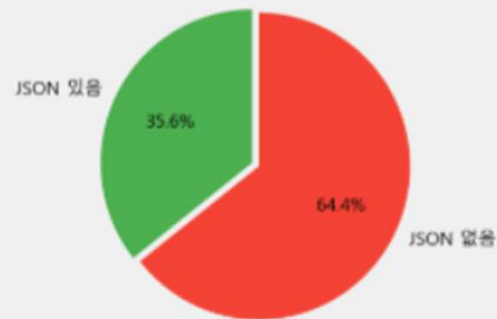
총 JSON: 1,001개
이미지 있는 JSON: 763개
이미지 없는 JSON: 238개

총 매핑 수: 1,182개

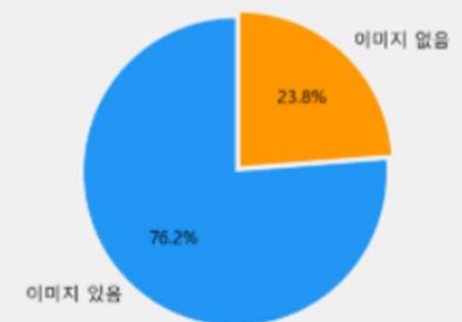
▶ 이미지당 JSON 개수 분포



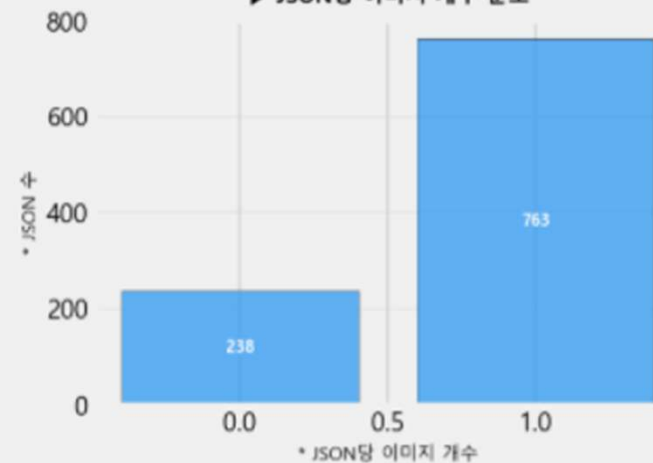
▶ 이미지별 JSON 매핑 비율



▶ JSON별 이미지 매핑 비율



▶ JSON당 이미지 개수 분포



Part 3

Image-Annotation 매핑 분석

▣ 이미지당 JSON 개수 통계

- 최소: 0개
- 최대: 4개
- 평균: 1.17개

▣ 분포

- 0개: 419개 (64.4%)
- 2개: 7개 (1.1%)
- 3개: 151개 (23.2%)
- 4개: 74개 (11.4%)

▣ JSON당 이미지 개수 통계

- 최소: 0개
- 최대: 1개
- 평균: 0.76개

▣ 분포

- 0개: 238개 (23.8%)
- 1개: 763개 (76.2%)

▣ 각 상태별 이미지/어노테이션 숫자

- 이미지(O)/어노테이션(O)

image : 232
anno : 763

- 이미지(X)/어노테이션(O)

image : 137
anno : 365

→ AI HUB 데이터로 부터 같은 이름의 이미지를 복사 후 매칭.
→ 또한 이 이미지에 해당하는 추가 annotation을 복사.

- 이미지(O)/어노테이션(X)

image : 419
anno : 1676

→ MD5 checksum으로 이미지를 확인 후 이에 해당하는 annotation 복사 후 매칭

▣ 예외 분석

- category_id가 1인 (dl_idx 와 category_id가 다른 경우)
1950 개 → 모두 category_id 를 dl_idx 로 수정.

- 같은 image에서 bbox의 좌표가 같은 경우.

137개. 이미지(X)/어노테이션(O) 에서 추가 annotation 복사 중 생긴 것으로, category_id와 bbox가 동일한 경우 하나는 삭제.

- box 좌표가 잘못된 경우

6 개

K-003351-003832-020238_0_2_0_2_70_000_200.json, BBox: []

K-003351-013900-020238_0_2_0_2_70_000_200.json, BBox: []

K-003351-016688-018357_0_2_0_2_90_000_200.json, BBox: [70, 438, 584]

K-003351-031863-036637_0_2_0_2_90_000_200.json, BBox: []

K-003351-031863-038162_0_2_0_2_90_000_200.json, BBox: [428, 199, 295]

K-003544-006563-016548-033878_0_2_0_2_75_000_200.json, BBox: [133.183, 202, 200]