| **빅데이터 분석 정의서** |
| --- |

| **주제 :** | 많은 서류의 내용을 정리하는 작장인들을 위한 객체 인식과 OCR을  통해 제공받은 키워드로 서류의 중요 자료를 정리하는 서비스 |
| --- | --- |

| **2022.12.06** |
| --- |

**Ⅰ. 개요**

1. 아이디어 주제

: 종이 파일 인식하여 디지털화

2. 개발 목표

: PDF 파일을 분석하여 데이터를 뽑아 키워드 별로 검색 가능하게 하는 서비스

3. 개발 내용

: 종이파일 인식하여 디지털화

**Ⅱ. 기능별 빅데이터 분석 명세서**

\*기능별로 정의되어야 함

| **기능명** | 종이 파일 인식하여 디지털화 | |
| --- | --- | --- |
| 1. **데이터 준비** | | |
| **데이터 정의** | 표가 있는 문서 | |
| **데이터 획득 방법** | RoboFlow 안에 있는 bounding Box 처리된 이미지 | |
| **2. 전처리** | | |
| **전처리 과정** | imgaug 라이브러리를 통해 학습 시킬 데이터를 증가 | |
| **3. 모델 생성 및 학습** | | |
| **모델링 목표** | 데이터 추출 및 분석 | |
| **모델링 가능 알고리즘** | 데이터 추출 | Yolov5 모델 |
| 키워드분석 | OCR 모델 |
| **학습** | 1. PDF에서 YOLOv5모델을 사용하여 이미지 파일 추출  2. 이미지 파일을 OCR 모델을 통하여 데이터를 추출하고 엑셀파일로 생성 | |
| **4. 검증** | | |
| **모델링 검증 방안** | Yolov5 학습을 통해 나온 best.pt를 학습에 사용하지 않은 PDF 파일을  통해 검출이 원하는 방향으로 표를 추출하는 지 확인 | |
| **모델링 평가 결과** | 현재 표를 잘 추출하지만 표 안에 그래프를 표로 인식하는 경우 생김  추출된 표를 DF 로 전환하는 과정에서 결측치 처리를 통해 제거 | |