**데이터 수집 및 전처리 안내**

1. **수집 데이터**

**•** **ESG 모범규준 21년\_개정판1** (ESG 기준원)

• **ESG 경영 전략 보고서2,3,4,5** (각 기업별 홈페이지)

1. **데이터 형식 및 구조**

* 수집된 파일은 모두 PDF 형식

1. **데이터 전처리 과정**
2. **OPENAI 파인튜닝**

* 특수문자 정리 (단, 한글, 영문, 숫자, 기본 문장부호는 유지)
* 문단 구분을 위한 패턴 추가
* 문단 시작과 끝 부분의 불필요한 공백 제거
* 헤더/푸터로 보이는 반복되는 텍스트 제거 (예: 페이지 번호)
* 문장 단위로 줄바꿈 추가 (마침표, 물음표, 느낌표 뒤에)
* 전처리 된 파일을 QA 형식의 jsonl 파일로 변환

패턴, 보라색, 라일락, 스크린샷이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림 1. 2021 모범규준 QA.jsonl

1. **RAG 결합용 데이터**

* PDF 파일을 PPT로 변환
* 텍스트나 정보가 없는 순수 이미지 제거
* 상단 및 좌측에 있는 불필요한 텍스트(로마숫자, 영어대문자로 된 목차) 제거
* 메타데이터 jsonl 파일로 변환

스크린샷, 텍스트, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림 2. 신한 라이프 전처리.jsonl

텍스트, 스크린샷, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

**그림 3. 삼표 페인트 피피티 전처리**.jsonl

스크린샷, 블랙이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.

그림 4. CJ 프레시 피피티 전처리.jsonl

**스크린샷, 텍스트, 폰트이(가) 표시된 사진

AI가 생성한 콘텐츠는 부정확할 수 있습니다.**

그림 5. KT&G 피피티 전처리

**자료 출처**

1. <https://www.cgs.or.kr/business/best_practice.jsp>
2. <https://www.shinhanlife.co.kr/hp/cdhh0160.do>
3. <https://www.sampyocement.co.kr/kor/esg/esg_05.php>
4. <https://www.cjfreshway.com/sustainability/reports.jsp>
5. <https://www.ktng.com/policies>