

데이터 수집 및 전처리 결과서

- 1. 데이터 출처 및 목록
- 2. 데이터 전처리 (Data Preprocessing)
jaeyong2/ja-rag-cot
법령해석사례집(20년 ~ 23년 데이터 사용)
크롤링 한 네이버 뉴스 기사(2024/12/21~2024/12/24)

1. 데이터 출처 및 목록

데이터	수집 목적	기간	출처	설명	개수
🗨️ jaeyong2/ja-rag-cot (Hugging Face)	CoT 기반 QA 파인튜닝	-	Hugging Face	QA, CoT 형태 레포에서 데이터 다운로드 question, answer, context, cot answer 형식만 다운로드 (외국어로 된 양질의 데이터도 수집 후 번역)	209,496
🗨️ HoJL/qa_no7(Hugging Face)	한국어 QA 파인튜닝	2024/12/21~2024/12/24	네이버 뉴스	특정 날짜의 특정 언론의 기사를 크롤링(조선, 중앙, 동아)	9,665
2018~2023 법령해석 사례집.pdf	도메인 특화 컨텍스트 QA 파인튜닝 및 RAG 데이터	2018~2023년	법제처	법제처의 2018~2023년까지의 법령 해석 사례집 PDF 다운로드.	1,974

2. 데이터 전처리 (Data Preprocessing)

jaeyong2/ja-rag-cot

- 일본어 데이터를 gpt api를 이용해 번역요청
- context, Question, Thinking, Final Answer, 컬럼을 context, question, cot answer, answer컬럼으로 매핑

원본 데이터 예시

context	question	RAW Ground Truth	Thinking	Final Answer
日常生活（にちじょうせいいかつ、）は、毎日繰り返される普段の生活のこと。日常...	禅的思考を日常生活に取り入れることを何と...	["禅ではしばしば日常生活に焦点をあて...	1. Problem Understanding: 問題は、禅的思考を日常...	禅的思考を日常生活に取り入れることを「生...
GNU Free Documentation License（グニユー・フリ...	GFDLが文書の配布を目的としている理由は...	["なお、GPLが主にコンピュータプログラ...	1. Problem Understanding: 本質的に、GFDLが文書...	GFDLが文書の配布を目的としている理由...

전처리 후 데이터 예시

context	question	cot answer	answer
일상생활은 매일 반복되는 평범한 삶을 의미한다.\n\n일상생활이란 사람이 살아가기 ...	"선적 사고를 일상생활에 접목시키는 것을 무엇이라고 부릅니까? 또한, 구체적으로 어...	1. 문제 이해: 문제는 선(禪)적 사고를 일상생활에 접목시키는 것의 명칭과 그 구...	"선(禪)적 사고를 일상생활에 접목시키는 것을 '생활선'이라고 부릅니다. 구체적으로...
GNU 자유 문서 라이선스(GNU Free Documentation License,...	GFDL이 문서의 배포를 목적으로 하는 이유는 무엇인가요? 또한, GPL과 GFDL...	1. 문제 이해: 본질적으로 GFDL이 문서 배포를 목적으로 하는 이유와 GPL과의..	GFDL이 문서의 배포를 목적으로 하는 이유는 문서에 특화된 조항이 정해져 있기 때...

법령해석사례집(20년 ~ 23년 데이터 사용)

- pdf파일에서 '절'별 안전번호를 추출하여 안전번호를 키워드로하여 국가법령정보 api를 이용해
질의요지, 회답, 이유, 안전명 4가지 데이터 수집. 절은 분류로 사용.

pdf예시

파란색 박스는 '절', 빨간색 박스는 안전번호

제1장. 기획재정·금융·공정거래

제1절 기획재정·금융

1. 「관세법」 제89조제6항 제1호 등 관련 (23-0514, 23-0683) 3
2. 「지방자치단체를 당사자로 하는 계약에 관한 법률」 제7조 등 (23-0636) 8

제2장. 산업통상자원·중소기업

제1절 산업통상자원

1. 「전기사업법」 제2조 등 관련 (23-0478) 13
2. 「고압가스 안전관리법 시행규칙」 별표 10 제2호카목5) 등 관련 (23-0544) 16

api요청 결과 예시

```
1  {"ExpService": {
2    "해석기관코드": "1170000",
3    "안전번호": "23-0514",
4    "이유": "법해석의 목표는 어디까지나 법적 안정성을 저해하지 않는 범위 내에서 구체적 타당성을 찾는 데 두어야 하고, 나아가 그러기 위해서는 가능한 한 법률에 사용된 문언의 통상적인 의미에
5    "해석기관명": "법제처",
6    "관리기관코드": "",
7    "해석일자": "20231204",
8    "안전명": "민원인 - 「관세법」 제89조제6항제1호에 따른 관세감면 대상 물품의 범위(「관세법」 제89조제6항제1호 등 관련)",
9    "질의요지": "「관세법」 제89조제1항에서는 세율불균형을 시정하기 위하여 중소기업1)1) 「조세특례제한법」 제6조제1항에 따른 중소기업을 말하며, 이하 같은.이 대통령령으로 정하는 바에 따라
10   "법령해석례일련번호": "337651",
11   "질의기관명": "",
12   "질의기관코드": "",
13   "등록일시": "20231205",
14   "회답": "군용항공기용 부분품과 원재료는 「관세법」 제89조제6항제1호에 따라 관세가 감면되는 물품에서 제외되지 않습니다."
15 }}}
```

- 이유 데이터내에 관계법령이 같이 속해있어 관계법령 데이터를 이유에서 분리.
관계법령이 없거나 분리 할 수 없는 데이터는 제외함.
- 안전명에서 법령이름을 추출하여 국가법령정보 사이트를 크롤링 하여 법령 본문 데이터 수집
검색할 수 없는 법령은 제외함.
- 베이스라인 모델(llama 3.2 3B)의 최대 Max sequence length가 128k 이기 때문에
128k 이상의 토큰을 가진 데이터도 제외함.
- 최종적으로 만들어진 데이터는 년도, 분류, 질의요지, 회답, 이유, 관계법령, 법령 의 컬럼을 가지고있는
데이터 768개.

법령해석 데이터 예시

년도	분류	질의요지	회답	이유	관계법령	법령
2022	국토개발건 건축설	「건축법」 제52조제2항에서는 대통령령으로 정하는 건축물의 외벽에 사용하는 마감재료...	가. 질의 가에 대해이 사안의 경우 건축물 외벽의 수선이 수반되지 않는 용도변경 허...	가. 질의 가에 대해개정 건축법 시행령 부칙 제2조제8호에서는 같은 영 제61조제2...	「건축법」 제52조(건축물의 마감재료 등) ① (생략)② 대통령령으로 정하는 건축...	제1조(목적) 이 법은 건축물의 대지·구조·설비 기준 및 용도 등을 정하여 건축...
2020	농림	「농어업경영체 육성 및 지원에 관한 법률」(이하 “농어업경영체법”이라 함) 제6조의...	이 사안의 경우 등록정보가 정정된 날을 농어업경영체법 제4조제2항에 따른 변경등록 ...	농어업경영체법 제4조제1항에서는 농어업·농어촌에 관련된 용자·보조금 등을 지원받으려..	농어업경영체 육성 및 지원에 관한 법률제4조(농어업경영정보의 등록) ① 농어업·농어...	제1조(목적) 이 법은 경쟁력 있는 농어업경영체를 육성하고 농어업의 공동경영을 ...

크롤링 한 네이버 뉴스 기사(2024/12/21~2024/12/24)

- 네이버 뉴스의 기사 크롤링
 - target: 조선, 중앙, 동아
- 불필요한 기사 제거
 - 본문 길이가 400보다 적은 기사 필터링
 - 연관성이 떨어지는 기사 필터링
 - ex) 오늘의 운세, 만화 등
- 텍스트 전처리
 - 불필요한 특수문자, 공백 등 제거
 - 청킹 chunk_size=500, chunk_overlap=50
- context, question, cot answer, answer 형식을 갖춰서 gpt api를 이용해 한 기사 당 두 개의 Q&A 생성
사용 모델: gpt-4o-mini

네이버 뉴스 기사 QA 예시

```
1  {
```

```
2  "context": "프랑스가 정치적 혼란에 빠진 주요 원인도 경제 때문이다. 지난 9월 취임한 미셸 바르니에 총리는 프랑스의 재정 적자 축소를 위해 대기업과 고소득자에 대한 과세, 사회복지 지출 감소",
3  "question": "미셸 바르니에 총리가 사회복지장 재정 법안을 직권으로 통과시킨 이유는 무엇인가?",
4  "cot_answer": "1. 바르니에는 재정 적자 축소를 위한 긴급한 조치를 필요로 했으며, 이는 프랑스 경제의 불안정성을 해결하기 위한 핵심 요소였다.\n2. 의회에서의 반대가 예상되었고, 특히 좌파",
5  "answer": "바르니에 총리는 재정 적자 축소와 경제 안정성을 확보하기 위해 헌법 조항을 발동하여 사회복지장 재정 법안을 직권으로 통과시켰다."
6 },
7 {
8  "context": "프랑스가 정치적 혼란에 빠진 주요 원인도 경제 때문이다. 지난 9월 취임한 미셸 바르니에 총리는 프랑스의 재정 적자 축소를 위해 대기업과 고소득자에 대한 과세, 사회복지 지출 감소",
9  "question": "야권의 불신임이 바르니에 총리의 내각 붕괴에 미친 영향은 무엇인가?",
10 "cot_answer": "1. 야권은 바르니에 총리의 권한 행사에 대해 강한 반감을 표시하며 불신임을 추진함으로써 정부에 대한 신뢰를 저하시켰다.\n2. 불신임 투표는 정치적 갈등을 심화시켰고, 이는 정",
11 "answer": "야권의 불신임은 바르니에 총리의 내각 붕괴를 초래하여 정부에 대한 신뢰를 저하시켰고, 이는 프랑스의 정치적 혼란을 더욱 심화시켰다."
12 },
```