

법령 챗봇을 위한 데이터베이스 구축 및 기능 구축

2024.12.19

발표자 : 하정원

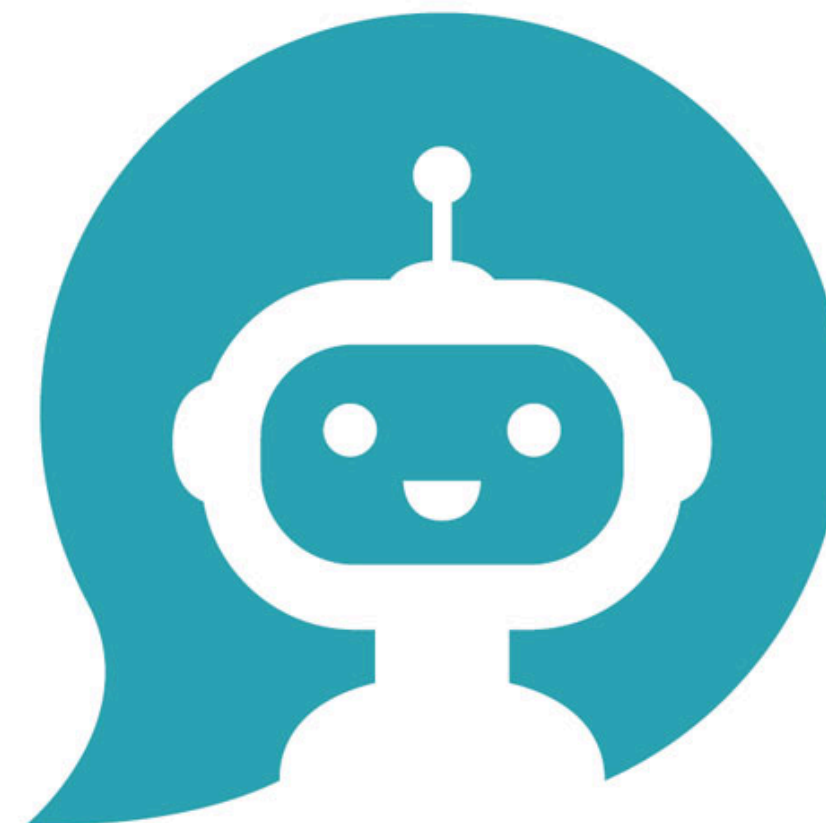


프로젝트 주제

법령 챗봇을 위한 데이터베이스 구축 및 기능 구축

프로젝트 목적

산업체의 산업보건 담당자가 업무 수행시 관련
법령/법규에 대해 궁금한 점을 물어볼 수 있는 챗봇을 구성한다.



데이터베이스화

✓ 법령 PDF를 PostgreSQL DB에 데이터베이스화.

임베딩 벡터

✓ 유사도 검색을 위해 각 조항의 임베딩 벡터를 계산하여 DB에 저장.

RAG Pipeline

✓ DB를 사용하여 검색 증강 기능 구현.



tables 9

combined_table

columns 14

id	integer = nextval('combined_...')
name	varchar(32)
part_num	varchar(20)
part_name	varchar(128)
chap_num	varchar(20)
char_name	varchar(100)
sec_num	varchar(20)
sec_name	varchar(100)
para_num	varchar(20)
para_name	varchar(64)
art_num	varchar(20)
art_name	varchar(128)
content	text
embedding	vector(786)

데이터 특징

안전보건규칙: 편(Part), 장(Chapter), 절(Section), 관(paragraph), 조(article)
그 외: 편(Part), 장(Chapter), 절(Section), 조(article)

데이터 베이스

임베딩 검색을 위한 PostgreSQL 기반 DataBase. 각 항목은 하나의 조를 의미.

기술 스택

pyPDF2로 텍스트화 시킨 후 parsing하여 psycopg2로 INSERT.

질문: 중대재해처벌법에서 말하는 사고의 범위는 어떻게 되나요?

1. 중대재해처벌법에서 정의하는 "중대재해"는 크게 "중대산업재해"와 "중대시민재해"로 구분됩니다.
2. 중대재해처벌법에서 말하는 사고는 원료, 제조물이나 공중이용시설, 공중교통수단의 결함으로 인한 재해에서 사망자가 발생하는 경우입니다.
3. 산업안전보건법에 따른 산업재해 중 사망자가 발생한 경우입니다.

sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	intfloat/multilingual-e5-base	jhgan/ko-sroberta-multitask (786차원)
<div></div> <div> "중대산업재해"와 "중대시민재해"로 </div> <div>0.698</div>	<div></div> <div> "중대산업재해"와 "중대시민재해"로 </div> <div>0.889</div>	<div></div> <div> "중대산업재해"와 "중대시민재해"로 </div> <div>0.836</div>
<div></div> <div> 공중이용시설, 공중교통수단의 . </div> <div>0.760</div>	<div></div> <div> 공중이용시설, 공중교통수단의 </div> <div>0.905</div>	<div></div> <div> 공중이용시설, 공중교통수단의 다. </div> <div>0.762</div>
<div></div> <div> 경우입니다. </div> <div>0.636</div>	<div></div> <div> 경우입니다. </div> <div>0.839</div>	<div></div> <div> 경우입니다. </div> <div>0.662</div>



- 임베딩 모델: jhgan/ko-sroberta-multitask
- top_k=10인 이유:
 - 사전 시나리오를 사용하여 검색하였을 때 예상했던 조항이 평균 6~7순위에 위치.
- 평균값 이상을 사용한 이유:
 - top_k 내부 자료들이 서로 동떨어진 조항은 아니었기에 평균 값 이상을 사용한다면 질문 중심의 Semantic Chunking 효과 기대
- o1 LLM에 답변 템플릿 사용.

SEMANTIC CHUNKING을 하지 않은 이유?

당시 Open Model만을 사용하고 있었기 때문에 Semantic Chunking이 기능적으로 불가능 했음.
차안으로 하나의 행에 하나의 조항이 들어가게 만들었고, 커스터마이징된 top_k를 통해 유동적 Semantic Chunking을 구현되어 오히려 정확도가 올라갔다고 예측.

OPENAI EMBEDDING을 사용하지 않은 이유?

▣OpenAI Embedding 모델도 LLM과 마찬가지로 비용이 지출이 됨.
그러나 그 비용이 인정될 정도의 성능은 아니라고 판단. LLM에 지출되는 것으로 충분.

별표/부칙을 DB에 저장하지 않은 이유?

▣별표와 부칙은 관계형 데이터베이스에 저장되기에 적합하지 않은 구조를 가지고 있음.
만약 별표와 부칙에 관한 안내를 넘어서 내용을 답변해야한다면 DB를 통한 RAG보다는 LlamaIndex안에서 PDF를 그대로 Indexing시키는 방식으로 해결해야한다고 예상.

PDF 텍스트 추출 및 DB 저장 프로그램

어떤 법령 PDF이든 각 조항별로 DB에 저장할 수 있음

임베딩 벡터 생성 및 저장 프로그램

▣ 벡터 검색을 위하여 법령 테이블의 항별 벡터 값을 생성하고 저장할 수 있음.

RAG를 사용한 법령 챗봇 프로그램

- 사전 시나리오 테스트 모두 통과한 정확도
 - 사전 정의된 시나리오 9개(각기 다른 분야에 관한 질문과 답)로 테스트 해보았을 때 모두 자세하고 정확한 답변을 제공.
- 평균 20~30초 답변 시간
 - 추론 모델 o1의 답변에서 추론 시간이 1분이 넘어가는 경우가 많음. 그러나 해당 RAG GPT-4o는 만족되는 정확도에 절반 이하의 시간이 걸림.

질문: 중대재해처벌법에서 말하는 사업자의 의무에는 어떤 것들이 있나요?

관련 법령 정리:

1. **중대재해처벌법 제2장 중대산업재해 제4조 (사업주와 경영책임자등의 안전 및 보건 확보의무)**

– 사업주 또는 경영책임자등은 사업장에서 종사자의 안전·보건상 유해 또는 위험을 방지하기 위해 다음과 같은 조치를 해야 한다:

1. 안전보건관리체계의 구축 및 이행을 위한 인력 및 예산의 확보.
2. 재해 발생 시 재발방지 대책의 수립 및 이행.
3. 법령에 따라 개선, 시정 등이 명해진 사항의 이행.
4. 안전·보건 관계 법령에 따른 의무이행에 필요한 관리상의 조치.

2. **중대재해처벌법 제2장 중대산업재해 제5조 (도급, 용역, 위탁 등 관계에서의 안전 및 보건 확보의무)**

– 사업주 또는 경영책임자등은 제3자에게 도급, 용역, 위탁 등을 행한 경우에도 그 시설, 장비, 장소 등을 실질적으로 지배·운영·관리하는 경우 제3자의 종사자에게 중대산업재해가 발생하지 않도록 법적으로 요구되는 조치를 해야 한다.

3. **중대재해처벌법 제3장 중대시민재해 제9조 (사업주와 경영책임자등의 안전 및 보건 확보의무)**

– 사업주 또는 경영책임자등은 사업장에서 생산·제조·판매·유통 중인 원료나 제조물의 설계, 제조, 관리상의 결함으로 인한 재해를 방지하기 위해 다음의 조치를 해야 한다:

1. 안전보건관리체계의 구축 및 이행을 위한 인력, 예산, 점검.
2. 재해 발생 시 재발방지 대책의 수립 및 이행.
3. 법령에 따라 개선, 시정 등이 명해진 사항의 이행.
4. 안전·보건 관계 법령에 따른 의무이행에 필요한 관리상의 조치.

답변: 중대재해처벌법에 따라 사업자는 종사자의 안전과 보건을 보호하기 위한 체계를 구축하고 효과적으로 운영해야 하며, 재해가 발생했을 경우에는 재발 방지 대책을 수립하여 이를 이행할 의무가 있습니다. 사업주는 또한 관계 법령에 따른 의무를 준수하고, 정부 기관이 요구하는 개선 및 시정 사항을 이행해야 합니다. 이는 주로 사업주가 실질적으로 지배·운영·관리하는 사업장에서 요구되며, 사업의 규모 및 특성에 따라 구체적인 조치가 적용됩니다. 이를 통해 중대재해를 예방하고 시민과 종사자의 생명과 신체를 보호하는 데 목적이 있습니다.

THANK
YOU

감사합니다

