

RAG 구현

2025 동계 연구연수생 김현지

Contents

1. RAG - 온디바이스에의 구현

2. 한/영 버전 RAG 구현

- 임베딩 모델 간의 비교를 중점으로

3. RAG 응용 실험

- 임베딩 PDF 기반 질의응답 체계
- 개인정보처리방침 생성기 구현
- 청크와 오버랩 설정에 따른 응답/ 검색결과의 변화

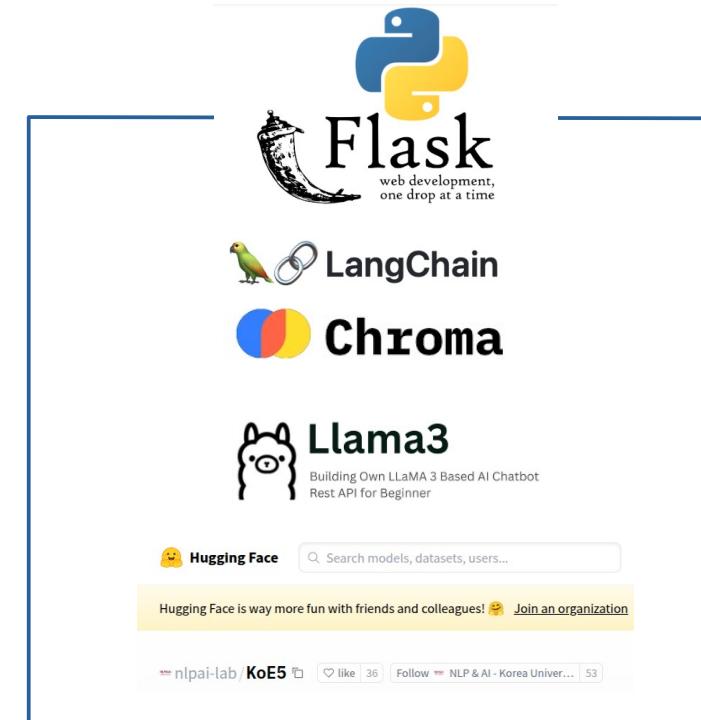
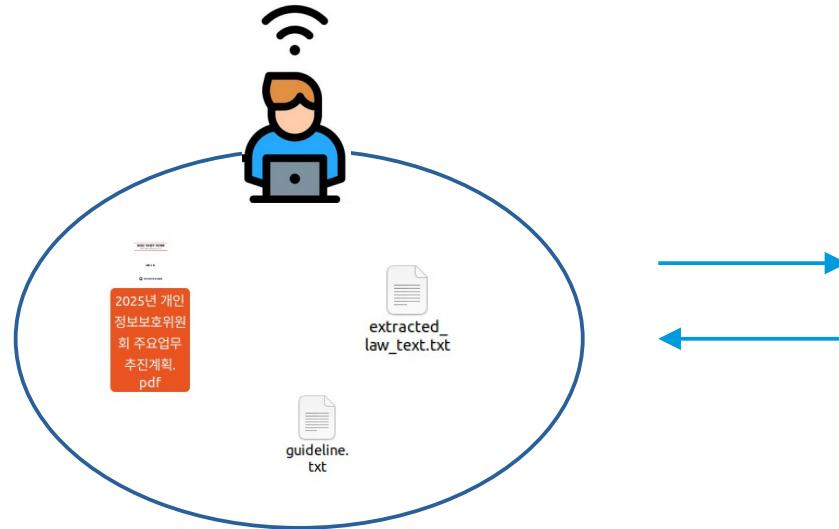
1. RAG- 온디바이스에의 구현

개요

- LLM은 외부망에 위치하고 있어,
외부망을 이용하지 못하는 곳에서 사용하기에 불편함
- 로컬 서버에 원하는 문서나 여러 형태의 파일들을 적재해두고
사용자의 질문에 해당 문서를 바탕으로 답하는 시스템을 만들고자 함.

1. RAG - 온디바이스에의 구현

환경설정 및 로직



2. RAG - 영어 버전 구현

- **FastEmbedding** 이용

- 파이썬 라이브러리의 일종으로, 가장 간단한 Embedding
- 한글에 특화된 모델의 부재로, 한글 이용 시 부자연스러움

```
5 from langchain_community.embeddings.fastembed import FastEmbedEmbeddings
```

```
# LLM 및 임베딩 초기화
cached_llm = Ollama(model="llama3")
embedding = FastEmbedEmbeddings()
```

FastEmbed 지원 모델

Supported Text Embedding Models

	model	dim	description	license	size_in_GB
0	BAAI/bge-small-en-v1.5	384	Text embeddings, Unimodal (text), English, 512...	mit	0.067
1	BAAI/bge-small-zh-v1.5	512	Text embeddings, Unimodal (text), Chinese, 512...	mit	0.090
2	snowflake/snowflake-ecctic-embed-xa	384	Text embeddings, Unimodal (text), English, 512...	apache-2.0	0.090
3	sentence-transformers/all-MiniLM-Lv2	384	Text embeddings, Multilingual (text), English, 256...	apache-2.0	0.090
4	jinaai/jina-embeddings-v2-small-en	512	Text embeddings, Unimodal (text), English, 819...	apache-2.0	0.120
5	BAAI/bge-small-en	384	Text embeddings, Unimodal (text), English, 512...	mit	0.130
6	snowflake/snowflake-ecctic-embed-a	384	Text embeddings, Unimodal (text), English, 512...	apache-2.0	0.130
7	nomic-ai/nomic-embed-text-v1.5-0	768	Text embeddings, Multilingual (text), English, 256...	apache-2.0	0.130
8	BAAI/bge-base-en-v1.5	768	Text embeddings, Unimodal (text), English, 512...	mit	0.210
9	sentence-transformers/paraphrase-multilingual-mpnet-base-v2	384	Text embeddings, Unimodal (text), Multilingual...	apache-2.0	0.220
10	Qdrant/clip-ViT-B-32-text	512	Text embeddings, Multilingual (text,image), Eng...	mit	0.250
11	jinaai/jina-embeddings-v2-base-base	768	Text embeddings, Unimodal (text), Multilingual...	apache-2.0	0.320
12	BAAI/bge-base-en	768	Text embeddings, Unimodal (text), English, 512...	mit	0.420
13	snowflake/snowflake-ecctic-embed-m	768	Text embeddings, Unimodal (text), English, 512...	apache-2.0	0.430
14	nomic-ai/nomic-embed-text-v1.5	768	Text embeddings, Multilingual (text, image), Eng...	apache-2.0	0.520
15	jinaai/jina-embeddings-v2-base-base	768	Text embeddings, Unimodal (text), English, 819...	apache-2.0	0.520
16	nomic-ai/nomic-embed-test-v1	768	Text embeddings, Multilingual (text, image), Eng...	apache-2.0	0.520
17	snowflake/snowflake-ecctic-embed-m-long	768	Text embeddings, Unimodal (text), English, 204...	apache-2.0	0.540
18	mixedread-ai/mmbai-embed-large-v1	1024	Text embeddings, Unimodal (text), English, 512...	apache-2.0	0.640
19	jinaai/jna-embeddings-v2-base-code	768	Text embeddings, Multilingual (text), Multilingual...	apache-2.0	0.640
20	sentence-transformers/paraphrase-multilingual-snowflake/snowflake-ecctic-embed-l	768	Text embeddings, Unimodal (text), English, 512...	apache-2.0	1.000
21	thenlper/gte-large	1024	Text embeddings, Unimodal (text), English, 512...	mit	1.200
22	BAAI/bge-large-en-v1.5	1024	Text embeddings, Unimodal (text), English, 512...	mit	1.200
23	inffloat/multilingual-e5-large	1024	Text embeddings, Unimodal (text), Multilingual...	mit	2.240

Supported Late Interaction Text Embedding Models

	model	dim	description	license	size_in_GB	additional_files
0	answerdot/answerai-colbert-small+1	96	Text embeddings, Unimodal (text), Multilingual...	apache-2.0	0.13	NaN
1	colbert-ir/colbert-v2.0	128	Late interaction embeddings, Multilingual...	mit	0.44	NaN
2	jinaai/jina-colbert-v2	128	New model that expands capabilities of colbert...	cc-by-no-4.0	2.24	[onnx/model.onnx_data]

Supported Image Embedding Models

	model	dim	description	license	size_in_GB
0	Qdrant/resem50-onnx	2048	Image embeddings, Unimodal (image), 2023...	apache-2.0	0.10
1	Qdrant/cocoViT-B-32-vision	512	Image embeddings, Multilingual (text,image), 202...	mit	0.34
2	Qdrant/Unicon-ViT-B-32	512	Image embeddings, Multilingual (text,image), 202...	apache-2.0	0.48
3	Qdrant/Unicon-ViT-B-16	768	Image embeddings (more detailed than Unicon-Vi...	apache-2.0	0.82

Supported Rerank Cross Encoder Models

	model	size_in_GB	description	license
0	Xenova/ms-marco-MiniLM-L-6-v2	0.08	MiniLM-L-6-v2 model optimized for reranking...	apache-2.0
1	Xenova/ms-marco-MiniLM-L-12-v2	0.12	MiniLM-L-12-v2 model optimized for reranking...	apache-2.0
2	jinaai/jina-reranker-v1-tiny-en	0.13	Designed for blazing-fast re-ranking with 8K c...	apache-2.0
3	jinaai/jina-reranker-v1-turbo-en	0.15	Designed for blazing-fast re-ranking with 8K c...	apache-2.0
4	BAAI/bge-reranker-base	1.04	BGE reranker base model for cross-encoder rer...	mit
5	jinaai/jina-reranker-v2-base-multilingual	1.11	A multilingual reranker model for cross-encoder...	cc-by-no-4.0

https://qdrant.github.io/fastembed/examples/Supported_Models/

2. RAG- 영어버전 구현

- FastEmbed로 영어문서 업로드 시

```
(venv) etri12@etri12:~/flask_app$ gedit app1.py &
[1] 94014
(venv) etri12@etri12:~/flask_app$ curl -X POST http://localhost:8080/pdf
  -F "file=@/home/etri12/pdfs/CELEX_32016R0679_EN_TXT.pdf"
{
    "chunk_len": 437,
    "doc_len": 128,
    "filename": "CELEX_32016R0679_EN_TXT.pdf",
    "status": "Successfully uploaded"
}
```

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://localhost:8080/ask_pdf -H
"Content-Type: application/json" -d '{"query": "define personal data in GDPR"}'
{
    "answer": "According to Article 4 Definitions in the GDPR (General Data Protection Regulation), personal data refers to any information relating to an identified or identifiable natural person (data subject). This includes information that can be used to identify a person directly or indirectly, such as by reference to an identifier like a name, ID number, location data, online identifier, or one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person.\n\nIn other words, personal data is any information that can be linked to a specific individual, either directly (e.g., their name) or indirectly (e.g., through additional information).",
    "metrics": {
        "bleu_score": null,
        "map_score": null,
        "ndcg_score": null,
        "rouge_scores": null,
        "similarity_score": 0.8986865923172433
    }
}
```

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://localhost:8080/ask_p
df -H "Content-Type: application/json" -d '{"query": "define GDPR"}'
{'
```

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://localhost:8080/ask_p
df -H "Content-Type: application/json" -d '{"query": "define GDPR"}'
{
    "answer": "Based on the provided information, GDPR (General Data Protection Regulation) is a regulation that aims to protect fundamental rights and freedoms of natural persons in relation to the processing of their personal data. The principles of protection are applicable regardless of nationality or residence, ensuring respect for the right to the protection of personal data.\n\nIn summary, GDPR:\n1. Respects the right to the protection of personal data.\n2. Applies to the processing of personal data by controllers and processors in all member states.\n3. Sets out rules on the protection of natural persons with regard to the processing of their personal data.\n4. Ensures that individuals have control over their own personal data, including the ability to consent, access, correct, and erase it.\n\nThe regulation also outlines specific definitions, such as:\n\"Personal data\" refers to any information relating to an identified or identifiable natural person (data subject).\n\"Processing\" means any operation or set of operations performed on personal data, whether by automated means or not.\n\"Controller\" means the natural or legal person, public authority, agency or body which, alone or jointly with others, determines the purposes and means of processing of personal data.\n\"Processor\" means a natural or legal person, public authority, agency or body which processes personal data on behalf of the controller.\n\nThe regulation also includes provisions for special categories of personal data, such as sensitive information (e.g., racial or ethnic origin, political opinions, religious beliefs), and sets out rules for processing this type of data."
}
```

The screenshot shows a search interface with a text input field containing the Korean query "GDPR 정의". Below the input, there is a large block of English text providing the definition of GDPR. The text is a direct translation of the JSON response shown in the previous code block. At the bottom of the interface, there are buttons for "번역 수정" (Translation Modify) and "번역 평가" (Translation Review).

2. RAG- 영어버전 구현

- FastEmbed로 영어문서 업로드 시

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://127.0.0.1:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "What is the most important thing in GDPR?"}'
```

```
{  
    "answer": "Based on the provided information from the GDPR regulation, I would say that the most important thing in GDPR is the right to the protection of personal data. This is stated in Article 8(1) of the Charter of Fundamental Rights of the European Union and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), which provides that everyone has the right to the protection of personal data concerning him or her.\n\nThis right is reiterated throughout the GDPR regulation, emphasizing the importance of protecting natural persons' personal data and ensuring that they have control over their own data. The regulation sets out various principles for processing personal data, including transparency, fairness, and proportionality, with the ultimate goal of ensuring that individuals are protected from harm or prejudice as a result of data processing.\n\nIn addition, the GDPR places a strong emphasis on the importance of data protection by design and by default, which means that controllers and processors must take into account the right to the protection of personal data when designing and implementing their data processing activities. This requires a proactive approach to data protection, with controllers and processors being required to implement appropriate technical and organizational measures to ensure that personal data is processed in a way that respects individuals' rights and freedoms.\n\nOverall, the right to the protection of personal data is at the heart of the GDPR, and it serves as a foundation for all other principles and requirements set out in the regulation.",  
    "metrics": {  
        "similarity_score": 0.8387498735921068  
    }  
}
```

GDPR 규정에서 제공된 정보를 바탕으로 GDPR에서 가장 중요한 것은 개인정보 보호 권리라고 말씀드리고 싶습니다. 이는 유럽연합 기본권 헌장 제8조 제1항과 유럽연합 기능 조약(TFEU) 제16조 제1항에 명시되어 있으며, 이는 모든 사람이 자신과 관련된 개인정보를 보호할 권리가 있음을 규정하고 있습니다.\n\n이 권리(GDPR 규정 전반에 걸쳐 반복되며, 자연인의 개인 데이터를 보호하고 자신의 데이터에 대한 통제권을 보장하는 것의 중요성을 강조합니다. 이 규정은 투명성, 공정성, 비례성 등 개인 데이터 처리를 위한 다양한 원칙을 제시하며, 궁극적인 목표는 데이터 처리 결과로 개인이 피해나 편견으로부터 보호받을 수 있도록 하는 것입니다.)\n\n또한 GDPR은 설계 및 기본적으로 데이터 보호의 중요성을 강조하고 있으며, 이는 컨트롤러와 프로세서가 데이터 처리 활동을 설계하고 구현할 때 개인 데이터 보호 권리를 고려해야 한다는 것을 의미합니다. 이를 위해서는 데이터 보호에 대한 적극적인 접근이 필요하며, 컨트롤러와 프로세서는 개인의 권리와 자유를 존중하는 방식으로 적절한 기술적 및 조직적 조치를 이행해야 합니다.\n\n전반적으로 개인정보 보호 권리는 GDPR의 핵심이며, 규정에 명시된 다른 모든 원칙과 요구 사항의 기초가 됩니다.

유사도도 약 0.84로 나타나며, 쿼리에 대한 답변을 잘 하고 있음.

2. RAG-영어버전 구현

• 기사 업로드

GDPR, which changed global industry... Google is also fined 65.3 billion won

[Security News Reporter Yang Won-mo] The European Personal Information Protection Regulation (GDPR) will mark its first year in effect on the 25th. The regulation, which is mandatory for companies and organizations that handle the personal information of EU citizens, has changed the global industrial landscape significantly. As advertisers' marketing money is concentrated on large ICT platforms with GDPR-compliant capabilities, a new market called the Personal Information Protection Officer (DPO) has been created.

[Image = fixabay]

The GDPR is scary because of the high fines. The GDPR is largely imposed on two types of fines on 11 criteria, including the nature, duration, and severity of the breach. They are 'general violations' and 'severe violations'. General violations are subject to fines of '10 million euros (about W13.2 billion)' or '2 percent or more of global annual sales'. For serious violations, either '4 percent of global annual sales' or '20 million euros (about W26.4 billion)', the higher of '4 percent of global annual sales' will be fined. For global companies that generate tens of trillions of dollars in sales, 2 percent or 4 percent can be astronomical. We have summarized five major GDPR violations that occurred in the past year in order of fines.

Google - EUR 50 Million
In January, Google, a global company, was fined 50 million euros (about W65.3 billion) by the French National Information Freedom Commission (CNIL) for violating the GDPR's "Transparency (Article 5 Privacy Principles)" policy. △ No information on the purpose of data processing and storage period was provided in one place. △ No clear and comprehensive information was provided with general and ambiguous explanations. △ Failure to obtain the "valid consent" required for personalized advertisements also violated Article 6 of the GDPR's "processing legitimacy." Google has expressed its intention to appeal.

Barreiro Montizo - €400,000
In November last year, Portuguese healthcare provider Barreiro Montizo was fined 400,000 euros (about W500 million) for granting employees other than △ doctors access to patients' personal information, failing to apply technical and administrative measures to prevent △ illegal access to personal information, and failing to maintain the confidentiality, integrity and availability of △ privacy processing systems and services. This is known as the first GDPR violation to occur in Portugal.

Vsnid - €220,000
In March, the online advertising company "Vsnid" was fined 220,000 euros (about 300 million won) by the Polish government for violating Article 14 of the GDPR "information to be provided when personal information is not collected from the data subject."

구글의 위반 사례를 타깃으로 질의를 날립

Google – 5천만 유로 지난 1월, 글로벌 기업 '구글'은 GDPR의 '투명성(제5조 개인정보 보호 원칙)' 정책을 위반한 혐의로 프랑스 국가정보자유위원회 (CNIL)로부터 5천만 유로(약 653억 원)의 벌금을 부과받았습니다. △데이터 처리 목적과 저장 기간에 대한 정보가 한곳에 제공되지 않은 점 △일반적이고 모호한 설명과 함께 명확하고 포괄적인 정보가 제공되지 않은 점 △개인 맞춤형 광고에 필요한 '유효한 동의'를 받지 못한 점도 GDPR의 '처리 적법성' 제6조를 위반한 것으로 판단했습니다. 구글은 항소 의사를 밝혔습니다.

2025 GDPR 위반사례를 영어로 번역해 PDF파일 형태로 저장한 후
Chroma DB에 적재

2.RAG - 영어 버전 구현

- GDPR과 구글에 대한 쿼리 응답

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://localhost:8080/ask_pdf  
-H "Content-Type: application/json" -d '{  
    "query": "What is the GDPR clause that Google violated?"}  
  
{  
    "answer": "According to the text, Google was fined \u20ac50 million by the French National Information Freedom Commission (CNIL) for violating Article 6 of the GDPR's \"processing legitimacy\" requirement. Specifically, the violation occurred due to Google's failure to obtain valid consent for personalized advertisements.\n\nAdditionally, it is mentioned that the fine was related to a lack of transparency in Google's privacy policies and practices, as no information on the purpose of data processing and storage period was provided in one place, making it unclear what users' personal data was being used for.",  
    "metrics": {  
        "similarity_score": 0.8460861299578097  
    }  
}
```

한국어 ▾

높임말 응어집

"텍스트에 따르면, 구글은 프랑스 국가정보자유위원회(CNIL)로부터 GDPR의 \"처리 정당성\" 요건 제6조를 위반한 혐의로 5천만 유로의 벌금을 부과받았습니다. 특히, 구글이 개인 맞춤형 광고에 대한 유효한 동의를 얻지 못했기 때문에 이러한 위반이 발생했습니다.\n\n또한, 데이터 처리 및 저장 기간의 목적에 대한 정보가 한 곳에 제공되지 않아 사용자의 개인정보가 무엇에 사용되고 있는지 불분명하기 때문에 구글의 개인정보 보호 정책과 관행의 투명성 부족과 관련이 있다고 언급되었습니다,

번역 수정 | 번역 평가



쿼리: Google이 위반한 GDPR 조항은 무엇인가?

2. RAG - 영어버전 구현

- 시나리오성 질의응답

```
(venv) etri12@etri12:~/Flask_app$ curl -X POST http://localhost:8080/ask_pdf  
-H "Content-Type: application/json" -d '{  
    "query": "explain the way to transmit EU privacy to China"}'  
  
{  
    "answer": "According to Article 46 of the General Data Protection Regulation (GDPR), when transferring personal data from the European Union (EU) to a third country or international organization, the following conditions must be met:  
    . The transfer shall take place only if one of the following conditions is met:  
    n* The third country or international organization ensures an adequate level of protection for personal data essentially equivalent to that ensured within the EU.  
n* The data subject has given explicit consent to the proposed transfer after having been informed of the possible risks of such a transfer due to the lack of adequate protection and appropriate safeguards in the third country or international organization.  
n* The transfer is necessary for the performance of a contract between the data subject and the controller, or the conclusion or performance of a contract between two or more controllers.  
n* The transfer is necessary for the establishment, exercise, or defense of legal claims.  
n* The transfer is necessary to protect the vital interests of the data subject.  
n\nTo transmit EU privacy to China, you would need to meet one of these conditions. It's important to note that the GDPR also requires companies to provide appropriate safeguards and measures to ensure the protection of personal data when transferred to a third country or international organization.",  
    "metrics": {  
        "similarity_score": 0.8219073343125417  
    }  
}
```

한국어 ▾ 높임말 🔍 용어집 ⓘ

"일반 데이터 보호 규정(GDPR) 제46조에 따르면, 유럽연합(EU)에서 제3국 또는 국제기구로 개인 데이터를 전송할 때 다음 조건을 충족해야 합니다:
1. 다음 조건 중 하나가 충족되는 경우에만 전송이 이루어집니다:
n* 제3국 또는 국제 기구는 EU 내에서 보장된 개인 데이터와 본질적으로 동등한 수준의 보호를 보장합니다.
n* 정보 주체는 제3국 또는 국제기구에서 적절한 보호와 적절한 안전장치가 부족하여 이러한 전송의 위험 가능성을 통보받은 후 제안된 전송에 명시적으로 동의했습니다.
n* 양도는 정보 주체와 컨트롤러 간의 계약 이행 또는 두 명 이상의 컨트롤러 간의 계약 체결 또는 이행을 위해 필요합니다.
n* 양도는 법적 청구의 성립, 행사 또는 방어를 위해 필요합니다.
n* 전송은 데이터 주체의 중요한 이익을 보호하기 위해 필요합니다.
n\nEU 개인정보를 중국으로 전송하려면 다음 조건 중 하나를 충족해야 합니다. GDPR은 또한 기업이 제3국 또는 국제기구로 이전할 때 개인 데이터 보호를 보장하기 위한 적절한 보호 조치와 조치를 제공하도록 요구한다는 점에 유의하는 것이 중요합니다."

번역 수정 | 번역 평가

🔊 📄 ⭐ 🔍

쿼리: EU의 개인정보를 중국에 이전하는 방법을 설명하라.

FastEmbed와 한국어 호환의 한계

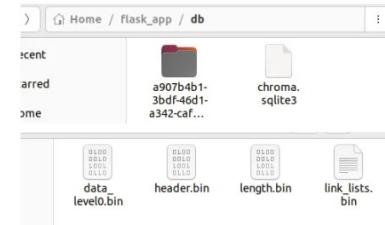
- FastEmbed로 한글 문서 업로드 시

```
^c(venv) etri12@etri12:~/flask_app$ curl -X POST -F "file=@/home/etri12/pdfs/개인정보 보호법(법률)(제19234호)(20240315).pdf" http://127.0.0.1:8080/pdf
{
  "chunk_len": 94,
  "doc_len": 41,
  "filename": "\uac1c\uc778\uc815\ubcf4 \ubcf4\ud638\ubc95(\ubc95\ub960)(\uc81c19234\ud638)(20240315).pdf",
  "status": "Successfully uploaded"
}
```

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://127.0.0.1:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "What is the most important thing in privacy law?"}'
{
  "answer": "As a technical assistant good at searching documents, I can help you find the information you need.\n\nAfter reviewing the provided text, I noticed that it appears to be related to the Personal Information Protection Act (PIPA) in Korea. If you're looking for specific information or want me to summarize certain sections, please let me know and I'll do my best to assist you!\n\nSome potential areas of focus could include:\n1. **Data Processing**: The text mentions various provisions related to data processing, such as consent requirements, data minimization, and security measures.\n2. **Protection Committee**: The Act establishes a Protection Committee (\ubcf4\ud638\uc704\uc6d0\ud68c) responsible for overseeing personal information protection efforts.\n3. **Anonymized Information**: Sections 28-28-4 discuss the handling of anonymized information (\ucac0\uba85\uc815\ubcf4), including its processing, combination, and preservation.\n\nPlease let me know if there's anything specific you'd like me to focus on or if you have any questions!"
}
```

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://127.0.0.1:8080/ask_pdf \
-H "Content-Type: application/json" \
-d '{"query": "개인정보 보호법에 따른 정보 제공 범위는 무엇인가?"}'
{
  "answer": "As a technical assistant good at searching documents, I can help you find the relevant information in this large text.\n\nCan you please specify what kind of information you're looking for? Are you looking for specific regulations related to personal data protection, or perhaps some guidance on how to handle certain situations?\n\nAlso, would you like me to summarize the main points of the document, or help you locate a specific section or clause within the text?\n\nLet me know, and I'll do my best to assist you!"
}
```

문서 검색에 능숙한 기술 어시스트트로서 필요한 정보를 찾는데 도움을 드릴 수 있습니다.\n\n제공된 텍스트를 검토한 결과, 한국의 개인정보보호법 (PIPA)과 관련이 있는 것으로 보입니다. 구체적인 정보를 찾고 있거나 특정 섹션을 요약하고 싶으시면 알려주시면 최선을 다해 도와드리겠습니다!\n\n다음과 같은 잠재적인 초점 영역이 포함될 수 있습니다:\n1. 데이터 처리: 테스트에는 동의의 요건, 데이터 최소화, 보안 조치 등 데이터 처리와 관련된 다양한 조항이 언급되어 있습니다.\n2. 보호위원회: 이 법은 개인정보 보호 노력을 감독할 책임이 있는 보호위원회(\ubcf4\ud638\uc704\udc60\ud68c)를 설립합니다.\n3. 익명화된 정보: 섹션 28-28-4에서는 익명화된 정보(\ucac0\uba85\uc815\ubcf4)의 처리, 조합 및 보존에 대해 논의합니다.\n\n특별히 집중하고 싶은 것이 있거나 궁금한 점이 있으면 알려주세요!



FastEmbed와 한국어 호환의 한계

- 동일한 질의를 AI에게 날렸을 때

```
(venv) etri12@etri12:~/flask_app$ curl -X POST http://127.0.0.1:8080/ai -H "Content-Type: application/json" -d '{"query": "What is the most important thing in privacy law?"}'  
{  
    "answer": "In privacy law, there are several key principles and concepts that are essential to understanding how personal information should be protected. However, if I had to pinpoint a single most important aspect of privacy law, it would be:  
    **Consent**: Informed consent is the foundation of privacy law. It means that individuals must knowingly and voluntarily agree to share their personal information with others. This concept is rooted in the idea that individuals have a right to control their own personal data and should only disclose it when they understand how it will be used and shared.  
    **Transparency**: The party seeking to collect or use personal information must clearly explain its purposes, methods, and limits.  
    **Voluntariness**: The individual's decision to provide personal information must not be coerced or influenced by external factors.  
    **Informed**: The individual must have sufficient knowledge about the risks associated with sharing their data.  
    Consent is crucial because it helps ensure that individuals are aware of how their personal information will be used, stored, and shared. This principle is reflected in various privacy laws and regulations, such as:  
    1. The General Data Protection Regulation (GDPR) in the European Union  
    2. The Health Insurance Portability and Accountability Act (HIPAA) in the United States  
    3. The Personal Information Protection and Electronic Documents Act (PIPEDA) in Canada  
    In summary, consent is the most important aspect of privacy law because it ensures that individuals have control over their personal information and are aware of how it will be used, which is essential for protecting their privacy and dignity."  
}
```

"개인정보 보호법에서는 개인정보를 어떻게 보호해야 하는지 이해하는 데 필수적인 몇 가지 주요 원칙과 개념이 있습니다. 그러나 개인정보 보호법의 가장 중요한 측면을 하나만 정확히 지적해야 한다면 다음과 같습니다:
동의: 사전 동의는 개인정보 보호법의 기초입니다. 이는 개인이 자신의 개인정보를 다른 사람과 공유하기 위해 알고 자발적으로 동의해야 한다는 것을 의미합니다. 이 개념은 개인이 자신의 개인정보를 통제할 권리가 있으며, 개인정보가 어떻게 사용되고 공유될지 이해할 때면 공개해야 한다는 생각에 뿌리를 두고 있습니다.
효과적인 동의: 필요합니다.
1. **투명성**: 개인정보를 수집하거나 사용하려는 당사자는 그 목적, 방법 및 한계를 명확히 설명해야 합니다.
2. **자발성**: 개인의 개인정보 제공 결정은 외부 요인에 의해 강요되거나 영향을 받아서는 안 됩니다.
3. **정보 제공**: 개인은 데이터 공유와 관련된 위험에 대해 충분한 지식을 가지고 있어야 합니다.
동의는 개인이 자신의 개인정보가 어떻게 사용, 저장 및 공유될지 알 수 있도록 돕기 때문에 매우 중요합니다. 이 원칙은 다음과 같은 다양한 개인정보 보호법 및 규정에 반영되어 있습니다:
1. 유럽 연합의 일반 데이터 보호 규정(GDPR)
2. 미국의 건강보험 휴대성 및 책임성에 관한 법률(HIPAA)
3. 캐나다의 개인정보 보호 및 전자문서법(PIPEDA)
요약하자면, 동의는 개인정보 보호법의 가장 중요한 측면입니다. 이는 개인이 자신의 개인정보를 통제할 수 있고, 개인정보가 어떻게 사용될지 인식할 수 있도록 보장하기 때문입니다. 이는 개인정보 보호와 존엄성을 위해 필수적입니다"

AI (llama3)가 pdf 문서 기반 검색보다 더 답을 잘 내고 있다.

3. RAG – 한국어 버전 구현

시도해본 모델

: [heegyu/EEVE-Korean-Instruct-10.8B-v1.0-GGUF](#)

- 파인튜닝이 필요하고, 그냥 실행 했을 때는 되지만 세션 분리가 잘 안돼서 앞 질문에 대해 계속해서 대답하고, 사람의 개입이 안되는 문제가 종종 발생함. (Human in the Loop 가 제대로 실행되지 않음)
- 임베딩 기능을 제공하지 않아서 임베딩 모델을 따로 써야 됨.
- llama.cpp 형식으로 경량화 되어 제공되고 있음.-> CPU 실행 시 적합한 모델이라는 생각이 들었음.
- CMAKE를 빌드하고 올려서 그냥 돌렸을 때는 채팅이 아주 잘됨.
하지만 QA용으로 쓰기에는 대화형 챗봇 특성상 특정 용도에 따라 Fine-tunning이 필요함.
- QA 용 RAG로는 대화형 모델은 적합하지 않은 이유:
캐시와 세션을 매 쿼리 때마다 계속 비우고 해도 세션이 유지 되기 때문에 앞 질문에 대한 대화가 연속됨.
출력값을 늘려도 똑같은 문제 발생함.
- QA 용으로 파인튜닝된 모델 사용의 필요성을 느끼게 됨.

3. RAG - 한국어 버전 구현

- 임베딩 모델을 가져와서 씀. - jhgan 모델

<https://huggingface.co/jhgan/ko-sroberta-multitask>

- llm 캐시를 대화할 때마다 초기화 함.

→ cached_llm.clear_cache()

```
}

(faiss) etri12@etri12:~/faiss_app$ curl -X POST -H "Content-Type: application/json" -d '{"message": "GDPR은 뭔가요?"}' http://localhost:8000/chat | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload   Total Spent  Left Speed
100  335  100  302  100   33  120k  13502 --::--- --::--- --::--- 163k
{
  "response": "물론이죠! 한국에서는 정보통신망법에 따라 개인정보보호법 위반 시 기업에 대한 벌금 및 처벌이 명시되어 있습니다. 벌금은 위반의 성격과 심각성에 따라 달라질 수 있으며, 다음과 같은 범위로 부과될 수 있습니다."
}

(faiss) etri12@etri12:~/faiss_app$ curl -X POST -H "Content-Type: application/json" -d '{"message": "GDPR은 뭔가요?"}' http://localhost:8000/chat | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload   Total Spent  Left Speed
100  197  100  164  100   33  68907  13865 --::--- --::--- --::--- 98500
{
  "response": "한국에서 개인정보보호법이 위반될 경우 기업의 벌금 및 처벌에 대해 더 자세한 정보를 제공해 주실 수 있나요?"
}
```

```
100  244  100  211  100   33  85947  13441 --::--- --::--- --::--- 119k
{
  "response": "개인 정보 보호법이 한국에서 어떻게 집행되는지, 그리고 위반 시의 가능한 결과, 예를 들면 벌금과 처벌에 대해 더 자세히 설명해 주실 수 있나요?"
}

[ERROR] 모델 호출 실패: 모델이 빈 응답을 반환했습니다.
127.0.0.1 - - [20/Jan/2025 16:59:51] "POST /chat HTTP/1.1" 500 -
[DEBUG] Sent prompt to llama: <|im_start|>user
개인정보보호법이 한국에서 어떻게 적용하나요?
<|im_end|>
<|im_start|>assistant
<|im_end|>

[ERROR] 모델 호출 실패: 모델이 빈 응답을 반환했습니다.
127.0.0.1 - - [20/Jan/2025 16:59:53] "POST /chat HTTP/1.1" 500 -
[DEBUG] Sent prompt to llama: <|im_start|>user
개인정보보호법이 한국에서 어떻게 적용하나요?
<|im_end|>
<|im_start|>assistant
<|im_end|>

[ERROR] 모델 호출 실패: 모델이 빈 응답을 반환했습니다.
127.0.0.1 - - [20/Jan/2025 16:59:55] "POST /chat HTTP/1.1" 500 -
[DEBUG] Sent prompt to llama: <|im_start|>user
개인정보보호법이 한국에서 어떻게 적용하나요?
<|im_end|>
<|im_start|>assistant
<|im_end|>

[ERROR] 모델 호출 실패: 모델이 빈 응답을 반환했습니다.
127.0.0.1 - - [20/Jan/2025 16:59:58] "POST /chat HTTP/1.1" 500 -
[DEBUG] Sent prompt to llama: <|im_start|>user
개인정보보호법이 한국에서 어떻게 적용하나요?
<|im_end|>
<|im_start|>assistant
<|im_end|>

[ERROR] 모델 호출 실패: 모델이 빈 응답을 반환했습니다.
127.0.0.1 - - [20/Jan/2025 17:00:01] "POST /chat HTTP/1.1" 500 -
[DEBUG] Sent prompt to llama: <|im_start|>user
개인정보보호법이 한국에서 어떻게 적용하나요?
<|im_end|>
<|im_start|>assistant
```

3. RAG - 한국어 버전 구현

- 다국어 지원 임베딩 모델 : google-bert/bert-base-multilingual-cased

```
query: 개인정보 보호법이 뭐야?  
Loading vector store  
/home/etri12/flask2/app.py:97: LangChainDeprecationWarning: The class 'Chroma' was deprecated in LangChain 0.2.9 and will be removed in 1.0. An updated version of the class exists in the :class:`~langchain.chroma` package and should be used instead. To use it run `pip install -U :class:`~langchain.chroma` and import as 'from :class:`~langchain_chroma import Chroma''.  
    vector_store = Chroma(persist_directory=folder_path, embedding_function=embedding_function)  
/home/etri12/flask2/venv/lib/python3.10/site-packages/langchain_core/vectorstore/base.py:1086: UserWarning: Relevance scores must be between 0 and 1, got [Document(metadata={'CreationDate': 'D:20250113113542+09' '00', 'ModDate': 'D:20250113113542+09' '00', 'Producer': 'iText 2.1.7 by 1T3XT', 'file_path': 'pdf/개인정보 보호법(법률)(제19234호)(20240315).pdf', 'page': 12, 'source': 'pdf/개인정보 보호법(법률)(제19234호)(20240315).pdf', 'total_pages': 41}, page_content='하여서는 아니 된다. 다만, 인명의 구조·구급 등을 위하여 필요한 경우로서 대통령령으로 정하는 경우에는 그러하니 아니하다.\n③ 제1항 각 호에 해당하여 이동형 영상정보처리기기로 사람 또는 그 사람과 관련된 사물의 영상을 촬영하는 경우에는 불빛, 소리, 안내판 등 대통령령으로 정하는 바에 따라 촬영 사실을 표시하고 알려야 한다.\n법제처 13 국가법령정보센터'), -48.70324301533385], (Document(metadata={'CreationDate': 'D:20250113113542+09' '00', 'ModDate': 'D:20250113113542+09' '00', 'Producer': 'iText 2.1.7 by 1T3XT', 'file_path': 'pdf/개인정보 보호법(법률)(제19234호)(20240315).pdf', 'page': 12, 'source': 'pdf/개인정보 보호법(법률)(제19234호)(20240315).pd
```

```
(venv) etri12@etri12:~/flask2$ curl -X POST http://localhost:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "개인정보 보호법이 뭐야?"}'  
{  
    "answer": "\ud83d\udcda\nAccording to my search results, the Personal Information Protection Act (PIPA) in Korea is a law that aims to protect the personal information of individuals. The full name of the act is the \"Personal Information Protection Act\" (), and it was enacted on February 29, 2012.\n\nThe PIPA defines personal information as \"any information related to a individual's identity or behavior\", such as names, dates of birth, addresses, phone numbers, credit card numbers, etc. The law requires organizations that process personal information to establish measures to ensure the secure handling and management of this information.\n\nSome key provisions of the PIPA include:  
1. Consent: Organizations must obtain the consent of individuals before collecting and using their personal information.  
2. Transparency: Organizations must inform individuals about how their personal information will be used and processed.  
3. Purpose limitation: Personal information can only be collected and used for a specific, legitimate purpose.  
4. Data minimization: Only necessary personal information should be collected and processed.  
5. Accuracy: Personal information must be accurate and up-to-date.\n\nThe PIPA also establishes penalties for organizations that violate the law, including fines and even criminal charges in some cases.\n\nPlease note that this is a summary of the main points of the PIPA, and it's not intended to be an exhaustive or definitive interpretation of the law. If you have specific questions about the PIPA or need more detailed information, I recommend consulting the actual text of the law or seeking legal advice from a qualified professional."  
}
```

잘 되는듯 하지만 no relevant document 문구가 뜨는 것을 발견함
→ LLM이 PDF 반영 없이 백그라운드 놀리지로 답변하는 것

제 검색 결과에 따르면, 한국의 개인정보보호법(PIPA)은 개인의 개인정보를 보호하기 위한 법입니다. 법의 정식 명칭은 "개인정보보호법"이며, 2012년 2월 29일에 제정되었습니다. PIPA는 개인정보를 이름, 생년월일, 주소, 전화번호, 신용카드 번호 등과 같이 "개인의 신원이나 행동과 관련된 모든 정보"로 정의합니다. 이 법은 개인정보를 처리하는 기관이 이 정보의 안전한 처리 및 관리를 보장하기 위한 조치를 수립하도록 요구합니다.

PIPA의 몇 가지 주요 조항은 다음과 같습니다:

- 동의: 조직은 개인 정보를 수집하고 사용하기 전에 개인의 동의를 얻어야 합니다.
- 투명성: 조직은 개인의 개인정보가 어떻게 사용되고 처리될지 개인에게 알려야 합니다.
- 목적 제한: 개인 정보는 특정 정당한 목적을 위해서만 수집되고 사용될 수 있습니다.
- 데이터 최소화: 필요한 개인 정보만 수집하고 처리해야 합니다.
- 정확성: 개인 정보는 정확하고 최신이어야 합니다.

또한 벌금과 경우에 따라 형사 고발을 포함한 처벌을 규정하고 있습니다.

이것은 PIPA의 주요 요점을 요약한 것이며, 법에 대한 철저하거나 확정적인 해석을 의도한 것은 아닙니다. PIPA에 대해 구체적인 질문이 있거나 더 자세한 정보가 필요한 경우, 법의 실제 본문을 참조하거나 자격을 갖춘 전문가에게 법률 자문을 구하는 것을 권장합니다."

3. RAG - 한국어 버전 구현

- 다국어 지원 임베딩 모델 : google-bert/bert-base-multilingual-cased

Vector DB에서 어떤 문서를 검색하고, 어떤 내용을 LLM에게 넘겨주는지 알고 싶어서 context로 출력하도록 코드 추가함 → 발췌문이 잘 전달되고 있으나 LLM은 다른 답을 하는 문제 발생

```
127.0.0.1 - [31/Jan/2025 09:31:30] "POST /ask_pdf HTTP/1.1" 200 -
query: 개인정보 파일 운영에 대해 한국어로 설명해줘.
Loading vector store
Retrieved documents and scores:
Document content (excerpt): 하여서는 아니 된다. 다만, 인명의 구조·구급 등을 위하여 필요한 경우로서 대통령령으로 정하는 경우에는 그러하n지 아니하다.n③제1항 각 호에 해당하여 이동형 영상정보처리기기로 사람 또는 그 사람과 관련된 사물의 영상을 촬영하는 경우n에는 불빛, 소리, 안내판 등 대통령령으로 정하는 바에 따라 촬영 사실을 표시하고 알려야 한다.n법제처 13 국가법령정보센
③제1항 각 호에 해당하여 이동형 영상정보처리기기로 사람 또는 그 사람과 관련된 사물의 영상을 촬영하는 경우n에는 불빛, 소리, 안내판 등 대통령령으로 정하는 바에 따라 촬영 사실을 표시하고 알려야 한다.
법제처 13 국가법령정보센
Document content (excerpt): ①공공기관 외의 개인정보처리자는 개인정보파일 운영으로 인하여 정보주체의 개인정보 침해가 우려되는 경우에는 영향평가를 하기 위하여 적극 노력하여야 한다.<개정 2023. 3. 14.>
법제처 20 국가법령정보센터
Document content (excerpt): 한 주의와 감독을 게을리하지 아니한 경우에는 그러하지 아니하다.
제74조의2(몰수·추징 등) 제76조부터 제73조까지의 어느 하나에 해당하는 죄를 지은 자가 해당 위반행위와 관련하여 취득한 금품이나 그 밖의 이익은 몰수할 수 있으며, 이를 몰수할 수 없을 때에는 그 가액을 추징할 수 있다. 이 경우n법제처 37 국가법령정보센터
제74조의2(몰수·추징 등) 제76조부터 제73조까지의 어느 하나에 해당하는 죄를 지은 자가 해당 위반행위와 관련하여 취득한 금품이나 그 밖의 이익은 몰수할 수 있으며, 이를 몰수할 수 없을 때에는 그 가액을 추징할 수 있다. 이 경우n법제처 37 국가법령정보센터
Document content (excerpt): 2020. 2. 4.>
제63조(자료제출 요구 및 검사) ①보호위원회는 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보처리자에게 관n게 물품·서류 등 자료를 제출하게 할 수 있다. <개정 2013. 3. 23., 2014. 11. 19., 2017. 7. 26., 2020. 2. 4.>n법제처 32 국가법령정보센터
제63조(자료제출 요구 및 검사) ①보호위원회는 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보처리자에게 관
```

```
"context": [
    "하여서는 아니 된다. 다만, 인명의 구조·구급 등을 위하여 필요한 경우로서 대통령령으로 정하는 경우에는 그러하n지 아니하다.n③제1항 각 호에 해당하여 이동형 영상정보처리기기로 사람 또는 그 사람과 관련된 사물의 영상을 촬영하는 경우n에는 불빛, 소리, 안내판 등 대통령령으로 정하는 바에 따라 촬영 사실을 표시하고 알려야 한다.n법제처 13 국가법령정보센
    "①공공기관 외의 개인정보처리자는 개인정보파일 운영으로 인하여 정보주체의 개인정보 침해가 우려되는 경우에는 영향평가를 하기 위하여 적극 노력하여야 한다.<개정 2023. 3. 14.>n법제처 20 국가법령정보센터",
    "한 주의와 감독을 게을리하지 아니한 경우에는 그러하지 아니하다.n제74조의2(몰수·추징 등) 제76조부터 제73조까지의 어느 하나에 해당하는 죄를 지은 자가 해당 위반행위와 관련하여 취득한 금품이나 그 밖의 이익은 몰수할 수 있으며, 이를 몰수할 수 없을 때에는 그 가액을 추징할 수 있다. 이 경우n법제처 37 국가법령정보센터",
    "2020. 2. 4.>n제63조(자료제출 요구 및 검사) ①보호위원회는 다음 각 호의 어느 하나에 해당하는 경우에는 개인정보처리자에게 관n게 물품·서류 등 자료를 제출하게 할 수 있다. <개정 2013. 3. 23., 2014. 11. 19., 2017. 7. 26., 2020. 2. 4.>n법제처 32 국가법령정보센터",
    "하여 사람 또는 사물의 영상 등을 촬영하거나 이를 유·무선망을 통하여 전송하는 장치로서 대통령령으로 정하n는 장치를 말한다.n8. “과학적 연구”란 기술의 개발과 실증, 기초연구, 응용연구 및 민간 투자 연구 등 과학적 방법을 적용하는 연구를 말한다.n법제처 1 국가법령정보센터"
]
```

```
(venv) etrii2@etrii2:~/flask$ curl -X POST http://localhost:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "개인정보 파일 운영에 대해 한국어로 설명해줘."}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time     Time  Current
          Dload  Upload   Total   Spent  Left  Speed
100  5712  100  5636  100    76   91      1  0:01:16  0:01:01  0:00:15  1323
{
  "answer": "As a technical assistant, I can help you understand the Korean documents provided. However, since I don't have direct information about '개인정보 파일 운영' (operation of personal information files), I will provide an overview of what I found.nFrom the given context and laws, it seems that '개인정보 파일 운영' refers to the management or operation of personal information files in Korea. The laws mentioned are from the National Law Information Center (법제처) and appear to be related to data protection and privacy.nArticle 1-3 mentions that when collecting images or videos through movable image processing equipment, a notice must be given as specified by presidential decree. This is to ensure transparency and accountability in handling personal information.nArticle 14 emphasizes the importance of conducting impact assessments before implementing new systems or processes that may affect individual privacy. This is to prevent potential risks and ensure compliance with data protection laws.nArticle 74-2 discusses fines and penalties for violating data protection laws, including confiscation of assets or property related to the violation.nLastly, Article 63 allows the Personal Information Protection Committee (보호위원회) to request information or documents from personal information processors in certain situations, such as when a complaint is filed.nIn summary, '개인정보 파일 운영' appears to be related to managing and protecting personal information files in Korea, which involves ensuring transparency, conducting impact assessments, and complying with data protection laws.nPlease note that my understanding may not be perfect due to the limitations of the model's training data."}
```

3. RAG - 한국어 버전 구현

- 이외에도 skt, kykim, jhgan, Beomi 등 수많은 모델로 실험해봄.

1. sentence-transformers의 encode() 함수는 리스트를 입력받고, Numpy 배열을 반환함.

- 하지만 ChromaDB에서는 리스트 형태의 리스트 (List[List[float]])를 요구함.

→ sentence-transformers는 np.array(넘파이 배열)를 반환하기 때문에 .tolist()를 통해 리스트형태 변환하는 것이 필요하다.

2. chromadb를 생성할 때 보통 768의 차원으로 생성하지만, 384차원 모델이 들어오면 임베딩 차원이 일치하지 않는다.

그런데 이때까지 FastEmbed로 했을 때는 기본 모델이 384니까 이것도 그대로 해도 되는거라고 생각해서 384모델을 계속 썼다.

→ FastEmbed에서는 내부적으로 ChromaDB와의 호환성을 위해 임베딩 차원과 데이터 형식을 자동 변환해주기 때문에 문제 없이 동작하는 거라고 함.

→ 벡터 저장소를 생성할 때 dimension 값을 384로 변환했다.

```
embedding = FastEmbedEmbeddings()

vector_store = Chroma(
    persist_directory="db",
    embedding_function=embedding
)

# 문장 임베딩 생성 (Numpy 배열 → 리스트 변환)
embeddings = model.encode(texts).tolist()

# ChromaDB 벡터 저장소 생성 (dim=384 설정 필요)
vector_store = Chroma(
    persist_directory="db",
    embedding_function=lambda docs: model.encode(docs).tolist(),
    collection_metadata={"dimension": 384}
)
```

3. RAG - 한국어 버전 구현

- nlpai-lab/KoE5

- sentence-transformers 모델



Model Versions

Model Name	Dimension	Sequence Length	Introduction
KURE-v1	1024	8192	Fine-tuned BAAI/bge-m3 with Korean data via CachedGISTEmbedLoss
KoE5	1024	512	Fine-tuned intfloat/multilingual-e5-large with ko-triplet-v1.0 via CachedMultipleNegativesRankingLoss

다국어 E5 모델

- 텍스트 임베딩을 생성하는 대규모 사전 훈련 모델.
 - 기본적으로 BERT 계열의 Transformer 아키텍처를 기반으로 함.
 - 텍스트를 벡터로 변환해 의미적으로 유사한 문장을 효과적으로 검색 가능
 - 대규모 데이터셋을 활용해 학습되었고, 문서 검색 성능이 뛰어남,
- E5는 단순한 텍스트 분류가 아니라 텍스트 간 의미적 유사도를 학습하는 모델임.
- 유사도 검색에 용이한 모델임.
- ### E5를 사용한 이유
- : 웹 문서, 논문, 뉴스 등으로 대규모 사전 훈련이 되어 있었음.
- 라벨링이 쿼리와 negative_hard, document로 이루어져 있어 QA에 적합하다고 생각했음.
 - Zeroshot, Fewshot이 가능해서 작은 데이터로도 검색 성능이 뛰어남.

3. RAG - 한국어 버전 구현

• Padding, truncation 설정

1. padding: 길이가 다른 텍스트들이 동일한 길이로 맞춰져서 모델에 입력될 수 있음.
2. Trunction: 모델이 최대 길이를 넘지 않도록 텍스트를 잘라서 입력할 수 있음.
 - chroma db는 리스트 형태로 데이터를 받기 때문에 .tolist()를 통해 변환함.
 - 앞서 chroma db 에 임베딩이 안됐던 이유가 모델 자체의 차원이 안 맞음도 문제가 되지만, 입력 데이터와 문서의 차원이 알맞으면 올바른 값이 도출 된다는 것을 알게 됐고, padding, truncation 값을 넣어줌으로써 문제를 해결할 수 있었음.

```
def embed_documents(self, texts):
    inputs = self.tokenizer(texts, return_tensors="pt", padding=True, truncation=True)
    with torch.no_grad():
        outputs = self.model(**inputs)
    embeddings = outputs.last_hidden_state[:, 0, :].numpy()
    return normalize(embeddings, norm="l2").tolist()

def embed_query(self, query):
    inputs = self.tokenizer(query, return_tensors="pt", padding=True, truncation=True)
    with torch.no_grad():
        outputs = self.model(**inputs)
    embeddings = outputs.last_hidden_state[:, 0, :].numpy()
    return normalize(embeddings, norm="l2").squeeze().tolist()
```

3. RAG - 한국어 버전 구현

• Pdf 업로드 후 쿼리 날리기

```
(venv) etri12@etri12:~/flask2$ curl -X POST http://localhost:8080/pdf -F "file=@/home/etri12/다운로드/2025년 개인정보보호위원회 주요업무 추진계획.pdf" | jq
```

```
{
  "chunk_len": 14,
  "doc_len": 12,
  "filename": "2025년 개인정보보호위원회 주요업무 추진계획.pdf",
  "status": "success"
}
```

```
(venv) etri12@etri12:~/flask2$ curl -X POST http://localhost:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "주요업무 추진계획에 대해 한국어로 말해줘."}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload Total Spent   Left Speed
100  1557  100  1485  100    72      57   2  0:00:36  0:00:25  0:00:11  414
{
  "answer": "👉 두 업무 추진 방향을 요약하자면 다음과 같습니다.\n\n**업무1. 주요업무 추진계획**\n\n* 주요업무의 목표를 달성하기 위해 필요한 조치를 정의하고, 이를 실행할 계획을 세웁니다.\n* 이 계획은 연간/분기별로 업무의 진행 상황을 모니터링하여 조정할 수 있도록 합니다.\n\n**업무2. 프로젝트 추진 방향**\n\n* 새로운 프로젝트를 기획하고, 이를 수행하는 팀원과 함께 업무 계획을 수립합니다.\n* 이 계획은 프로젝트의 목표, 일정, 예산 등을 정의하고, 이를 지키는 조치를 정의합니다.\n\n이러한 두 업무 추진 방향으로 주요업무를 추진하여 성과를 내보겠다는 것입니다.",
  "context": [
    "2 업무 추진 방향\n- 3 -"
  ]
}
```

```
(venv) etri12@etri12:~/flask2$ curl -X POST http://localhost:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "마이데이터에 대해 한국어로 말해줘."}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload Total Spent   Left Speed
100  3899  100  3837  100    62      66   1  0:01:02  0:00:58  0:00:04  787
{
  "answer": "😊\n\n마이데이터 시대 개막으로 국민 체감 성과 창출이 본격화 되었다. 국민들은 이제 자신의 정보를 이동시켜 원하는 서비스를 누리도록 지원받을 수 있다. 마이데이터 안착을 위한 선도서비스 5종 단계적으로 출시되어 의료·통신·에너지 등 국민생활 밀접분야에서부터 시작하여 모든 분야에 걸쳐 확산될 예정이다.\n\n또한, 선행부문(금융·공공등)과 신규부문(의료·통신etc.) 간 데이터 융합 지원을 통해 시너지 창출이 가능하게 되며, 합리적인 전송비용 분담체계 구축을 통해 국민들이 자신의 정보를 안전하게 보호할 수 있도록 한다.\n\n한편, 투명하고 안전한 마이데이터 생태계 구축을 위해 '마이데이터 지원 플랫폼'을 오픈하여 전송요구 이력 조회, 전송 철회 등 국민의 개인정보 전송요구권 행사를 지원할 수 있도록 한다. 또한, 개인정보관리 전문기관 안내서 발간을 통해 안전한 데이터 활용을 위한 지침을 제공하며, 엄격한 지정요건 심사 및 안전실태 점검을 통해 지속적으로 관리할 예정이다.\n\n따라서, 마이데이터 시대는 국민들이 자신의 정보를 자유롭게 활용하고, 새로운 서비스를 창출하는 기회를 제공할 뿐만 아니라, 안전하고 투명한 데이터 생태계 구축을 통해 국민의 삶을 개선할 수 있는 중요한 시기라고 할 수 있다.",

  "context": [
    "• 마이데이터 시대 개막, 성과 창출 본격화\n• 마이데이터 본격 시행으로 국민 체감 성과 창출\n• 의료·통신·에너지 등 국민생활 밀접분야부터 마이데이터 시행으로\n• 국민이 자신의 정보를 이동시켜 원하는 서비스를 누리도록 지원(3월)\n• 마이데이터 안착을 위한 마중물이 될 선도서비스 5종 단계적 출시\n• 전 분야 마이데이터 단계적 확산 추진\n• 의료·통신 분야의"
  ]
}
```

3. RAG - 한국어 버전 구현

• Pdf 업로드 후 쿼리 날리기

```
(venv) etri12@etri12:~/flask2$ curl -X POST http://localhost:8080/ask_pdf -H "Content-Type: application/json" -d '{"query": "영상정보를 가명처리 해야하는데, 불특정 다수에게 사전 동의를 구할 수가 없어. 어떻게 해야하는지 한국어로 알려줄래?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload   Total   Spent    Left  Speed
100  3621  100  3448  100    173      54      2  0:01:26  0:01:02  0:00:24  736
{
  "answer": "[INST] Ah, 좋은 질문입니다! 🧐 영상정보를 가명처리해야 하는 경우, 불특정 다수가 활용되어 사전 동의가 곤란한 경우, 「가칭영상정보처리기기 설치·운영 등에 관한 법률*」 제정 추진하여야 합니다. 이 법률에는 영상정보 활용 기준 마련, 대규모 영상관제시설 안전성 강화, 개인 권리행사 범위 확대 등의 내용이 포함됩니다. 또한, 생체인식기술 확산에 대응하여 생체인식정보 처리 원칙 및 정보주체 권리보호 방안 마련 등 생체인식정보 규율체계 개선도 필요합니다. 📊"
}
```

```
"context": [
  {
    "category": "Privacy Policy",
    "content": "▪️ 지속가능한 신산업 혁신 기반 마련\n▪️ 급격한 신기술 변화에 상응하는 법체계 마련\n▪️ 불특정 다수가 활용되어 사전 동의가 곤란한 영상정보의 특수성을 고려\n▪️ 「가칭영상정보처리기기 설치·운영 등에 관한 법률*」 제정 추진\n* 영상정보 활용 기준 마련, 대규모 영상관제시설 안전성 강화, 개인 권리행사 범위 확대 등\n▪️ 얼굴·지문 등 생체인식기술 확산에 대응하여 생체인식정보 처리 원칙 및 정보주체 권리보호 방안 마련 등 생체인식정보 규율체계 개선도 필요합니다.\n  "
    "source": "2025년 개인정보보호위원회 주요업무 추진계획.pdf"
  }
]
```

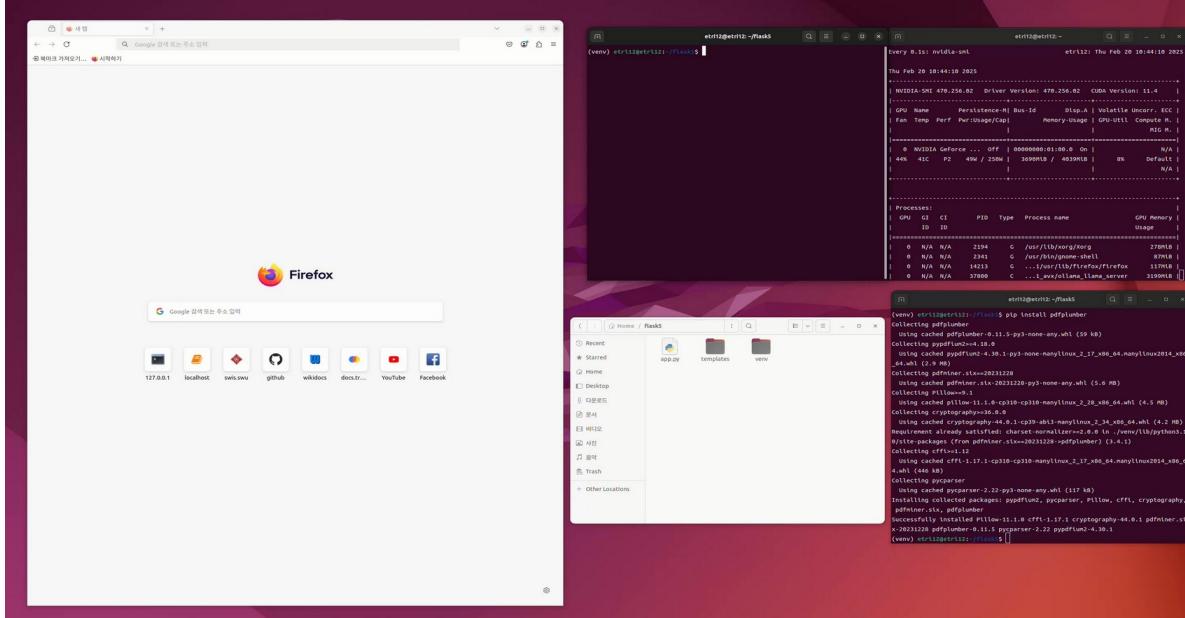
□ 급격한 신기술 변화에 상응하는 법체계 마련

- 불특정 다수가 활용되어 사전 동의가 곤란한 영상정보의 특수성을 고려하여 「가칭영상정보처리기기 설치·운영 등에 관한 법률*」 제정 추진
* 영상정보 활용 기준 마련, 대규모 영상관제시설 안전성 강화, 개인 권리행사 범위 확대 등
- 얼굴·지문 등 생체인식기술 확산에 대응하여 생체인식정보 처리 원칙 및 정보주체 권리보호 방안 마련 등 생체인식정보 규율체계 개선

□ 가명정보 활용 확산으로 신산업 성장 지원

- AI 연구자 등이 안전한 환경에서 유연한 가명처리가 가능하도록 지원하는 「가칭개인정보 이노베이션존」 추가 지정·활성화 및 사용자 편의성 개선
* 폐쇄망 시스템에서 클라우드 환경까지 확장, 기관 간 연계 활성화 등
- 가명처리 적정성 심의위원회 법제화 및 심의 지원 방안 마련, 가명처리 절차상 제출문서 감축 등 가명처리 기준·절차 간소화
- 지역 가명정보 활용 지원센터(총7개소) 협의회 및 산업체별 가명처리 전문가 협의체 도입으로 가명처리 사례 공유·확산 등 가명정보 활용 저변 확대
- '가명정보 지원 플랫폼'에 AI 등 신기술 개발의 핵심재료인 비정형 데이터(영상·이미지·음성 등) 가명처리 기능 추가 등 데이터 활용 지원

시연영상 – PDF 질의응답



PDF 기반 질의응답 RAG

쿼리: 14세 미만 아동의 정보를 수집할 때 주의할 점은?

PDF에 대해 질문을 입력하세요

질문을 입력하세요:

[에 대해 주의해야 할 점이 뭐야?]

질문 제출

질문에 대한 답변이 생성되었습니다.

PDF 질문에 대한 답변

14세 미만 아동의 개인정보 수집에 대해 주의해야 할 점은 다음과 같습니다. 1. **법정대리인의 동의 필요**: 14세 미만 아동의 개인정보를 처리하려면 법정대리인의 동의를 받아야 합니다. 따라서, 개인정보처리자는 법정대리인이 동의하였는지를 확인하여야 합니다(제22조의2 제1항). 2. **사용자의 직접 수집 허용**: 하지만, 대동령령으로 정하는 최소한의 정보에 해당되는 경우에는 법정대리인의 동의 없이 해당 아동으로부터 직접 수집할 수 있습니다(제22조의2 제2항). 3. **고지 등 필요**: 개인정보처리자는 14세 미만의 아동에게 개인정보 처리와 관련한 사항을 고지 등을 할 때에는 이 해하기 쉬운 양식과 명확하고 알기 쉬운 언어를 사용하여야 합니다(제22조의2 제3항). 이러한 조치들은 14세 미만 아동의 개인정보 보호를 강조하고 있습니다.

개인정보 처리방침 생성 질문을 입력하세요:

[개인정보처리방침을 생성해줘]

pdf 질문에 대한 답변은 위와 같다.

개인정보 보호법
[사행 2024. 3. 15.] | 법률 제19334호, 2023. 3. 14., 일부개정]
국민신문고 바로가기
아래 소프트웨어로 문서재보내요!

⑥ 삭제 <2023. 3. 14.>
⑦ 제1장부터 제5장까지에서 규정한 사항 외에 정보주체의 동의를 받는 세부적인 방법에 관하여 필요한 사항은 개인정보의 수집·매체 등을 고려하여 별도로로 정한다. <개정 2017. 4. 18., 2023. 3. 14.>

+ 최근 상당 규칙

④ 제22조의2(어동의 개인정보 보호) ① 개인정보처리자는 만 14세 미만 아동의 개인정보를 처리하기 위하여 이 법에 따른 동의를 받아야 할 때에는 그 법정대리인의 동의를 받아야 하며, 법정대리인이 동의하였는지를 확인하여야 한다.
② 제1항에도 불구하고 법정대리인의 동의를 받기 위하여 필요한 최소한의 정보로서 별도로로 정하는 정보는 법정대리인의 동의 없이 해당 아동으로부터 직접 수집할 수 있다.
③ 개인정보처리자는 만 14세 미만의 아동에게 개인정보 처리와 관련한 사항의 고지 등을 할 때에는 이해하기 쉬운 양식과 명확하고 알기 쉬운 언어를 사용하여야 한다.
④ 제1항부터 제3항까지에서 규정한 사항 외에 동의 및 동의 확인 방법 등에 필요한 사항은 별도로로 정한다.

[본조신설 2023. 3. 14.]

조문과 일치하는 법적 근거로 답하고 있음을 볼 수 있다.

문서 기반 질문하기

언하게 국외이전 할 수 있는 방법 | 질문하기

답변: 사용자는 금융 마이데이터 사업을 하려는 것을 목표로 하고 있습니다. 따라서, 사용자가 안전하게 국외 이전 할 수 있는 방법을 알려드리겠습니다.

첫째로, 금융분야 마이데이터 서비스 가이드라인에 따르면 마이데이터 전송 절차를 엄격히 준수해야 합니다. 이를 통해 고객의 개인정보가 보호되며, 사용자는 국외 이전시 문제가 발생하지 않도록 방지할 수 있습니다.

둘째로, 사용자는 PDS(개인 데이터 저장소)와 신용정보원의 전송요구앱을 구축하여 운영해야 합니다. 이를 통해 고객의 개인정보정보를 안전하게 보관하고, 국외 이전시 문제가 발생하지 않도록 방지할 수 있습니다.

셋째로, 사용자는 고객의 개인정보정보 자기결정권을 존중하여야 합니다. 이를 통해 고객이 자신의 개인정보정보를 관리하고, 국외 이전시 문제가 발생하지 않도록 방지할 수 있습니다.

따라서, 금융 마이데이터 사업을 하려는 경우 안전하게 국외 이전 할 수 있는 방법은 다음과 같습니다.

1. 마이데이터 전송 절차 엄격 준수
2. PDS와 신용정보원의 전송요구앱 구축 운영
3. 고객의 개인정보정보 자기결정권 존중

이러한 방안을 통해 금융 마이데이터 사업을 하려는 경우 안전하게 국외 이전 할 수 있습니다.

3.5. 정보전송 방법 및 절차

가. 마이데이터전송절차

- 마이데이터전송절차는 금융보안원 발간 「금융분야 마이데이터 기술 가이드라인」을 따른다.

나. 일반전송절차

정보제공자가 ① 고객에게 정보를 전송하는 절차와 ② 기관(신용정보제공·이용자, 개인정보평가회사, 개인사업자신용평가회사)에게 정보를 전송하는 절차를 구분

- ① 고객이 정보수신자인 경우 : PDS에 정보 전송

- 고객이 정보 전송을 요구하는 경우 고객의 PDS에 정보를 전송하고, 고객은 PDS를 통하여 정보를 수신하고 정보를 관리(PDS = 고객)

- (운영방법) 고객의 신용정보를 보유한 모든 기관이 PDS 개별 구축 및 운영에 대한 애로사항을 고려하여, 한국신용정보원(마이데이터 지원센터)에서 전송요구앱 및 PDS 통합운영 시스템을 구축하여 운영

* 워킹그룹 대상 '전송요구권 운영 방안에 대한 설문조사'('20.12.22)' 결과 반영

개인정보 처리방침 생성기

• 개인정보처리방침 프롬프트

```
context_text = "<br><br>".join([doc.page_content[:1024] for doc in retrieved_docs])

prompt_template = f"""
[개인정보 처리방침 자동 생성 시스템]

사용자가 요청한 사항: {query}

### 참고 문서:
{context_text}

출처: 해당 문서들에서 제공된 정보.

---

[개인정보 처리방침 생성기]

당신은 **개인정보 보호 전문가**입니다.
사용자가 요청한 사항과 제공된 참고 문서를 기반으로, 해당 산업 및 서비스에 적합한 **개인정보 처리방침**을 작성하세요.
외래어를 제외하고는 반드시 한국어만 사용해야 합니다.
다음 항목을 포함하여 개인정보 처리방침을 생성하십시오:

### **1. 개인정보의 수집 및 이용 목적**
- 해당 서비스에서 개인정보를 수집하는 목적을 명확히 설명하세요.
- 사용자의 요청이 특정 산업과 관련된 경우, 해당 산업에 적합한 목적을 반영하세요.

### **2. 수집하는 개인정보 항목**
- 필수적으로 수집해야 하는 개인정보 항목을 명확히 구분하세요.
- 선택적으로 수집하는 개인정보 항목도 안내하세요.
- 사용자의 요청 내용에 맞춰 해당 산업군에서 일반적으로 수집하는 항목을 반영하세요.

### **3. 개인정보의 보유 및 이용기간**
- 개인정보보호법에서의 법적 근거와 서비스 운영 정책에 따라 개인정보를 보유하는 기간을 설명하세요.
- 보유 기간이 종료되면 개인정보가 어떻게 처리되는지 안내하세요.

### **4. 개인정보의 제3자 제공 및 공유 여부**
- 사용자의 사전 동의를 받은 경우와 법령상 제공이 요구되는 경우를 구분하여 설명하세요.
- 제3자 제공 시,
- 제공받는 자:
- 제공 목적:
- 제공 항목:
```

```
### **5. 개인정보 보호 조치 및 안전성 확보 방안**
- 개인정보 보호를 위한 기술적·관리적·물리적 보호 조치를 안내하세요.
- 해당 서비스의 보안 정책이 있다면 반영하세요.

### **6. 사용자의 권리 및 행사 방법**
- 사용자가 자신의 개인정보에 대해 열람, 정정, 삭제 등을 요청할 수 있는 방법을 안내하세요.
- 요청 방법, 절차, 처리 기한 등을 명확히 설명하세요.

### **7. 영상정보처리기기 운영 및 관리**
- 고정형 및 이동형 CCTV, 기타 영상정보처리기기의 운영 목적과 관리 방안을 안내하세요.
- 해당 사항이 없을 경우 "해당 사항 없음"으로 작성하세요.

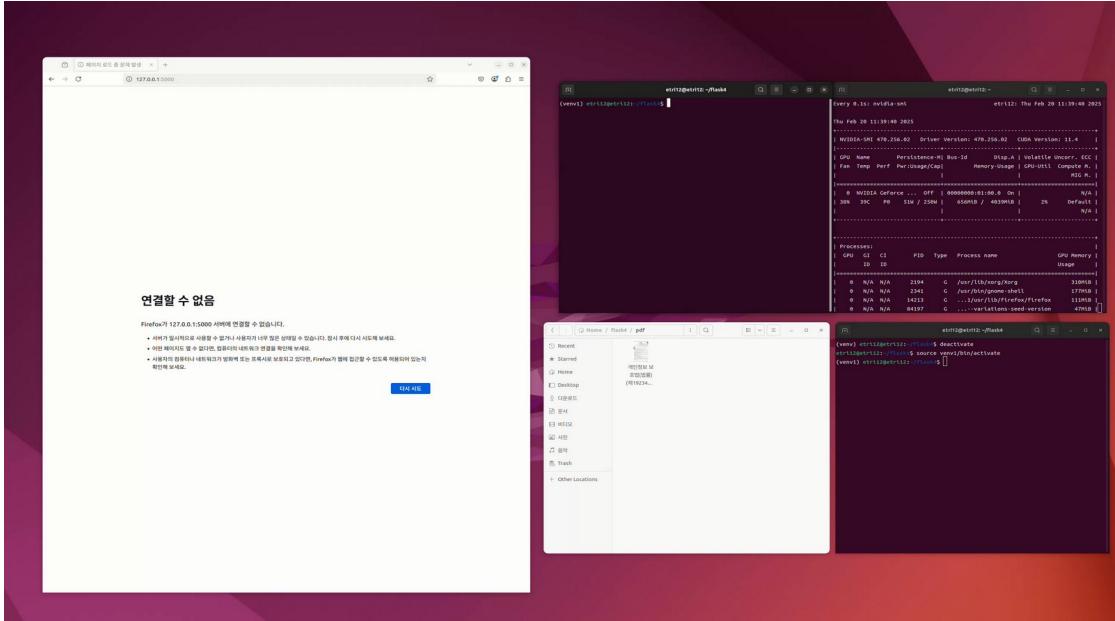
### **8. 개인정보 처리방침의 변경 및 고지 방법**
- 개인정보 처리방침 변경 시, 사용자가 어떻게 변경 사항을 알 수 있는지 안내하세요.
- 고지 방법(예: 웹사이트 공지, 이메일 통지 등)을 명확히 설명하세요.

### **9. 개인정보 보호 책임자 및 문의처**
- 개인정보 보호 관련 문의 및 불만을 접수할 담당자의 정보를 제공하세요.
- **담당자 이름, 소속 부서, 전화번호, 이메일 주소**를 포함하세요.

---

**작성 방식:**
- 반드시 **한국어**로 작성하세요.
- 참고 문서를 기반으로 정확하고 명확한 내용을 작성하세요.
- 주어진 문서에 해당 내용이 없는 경우 "해당 사항 없음"으로 답하세요.
- "등"과 같은 표현처럼 모호한 정보는 제공하지 말고, 사실에 기반한 답변만 작성하세요.
- 법적 용어와 공식적인 문장 구조를 사용하여 **전문적인 개인정보 처리방침**을 생성하세요.
```

시연영상 - 개인정보처리방침 생성



개인정보 처리방침 생성기

PDF에 대해 질문을 입력하세요

질문을 입력하세요:

어려면 어떤 법 조항에 주목해야

질문 제출

질문에 대한 답변이 생성되었습니다.

PDF 질문에 대한 답변

만 14세 미만 아동의 개인정보를 얻으려면 다음과 같은 법 조항을 주목해야 합니다. 제22조의2(아동의 개인정보 보호)에 따르면, 개인정보처리자는 만 14세 미만 아동의 개인정보를 처리하기 위하여 이 법에 따른 동의를 받아야 하며, 법정대리인의 동의를 받은 후 그에 대한 확인을 해야 합니다. 따라서, 만 14세 미만 아동의 개인정보를 얻으려면まず 법정대리인의 동의를 받은 후 그에 대한 확인을 해야 하며, 또한 그에 대한 이해하기 쉬운 양식과 명확하고 알기 쉬운 언어를 사용하여야 합니다.

생성된 개인정보 처리방침

키즈카페 연결 플랫폼 개인정보처리방침 1. **개인정보의 수집 및 이용 목적** 키즈카페 연결 플랫폼은 서비스 제공을 위하여 사용자의 개인정보를 수집합니다. 이는 **개인정보 보호법 제15조(개인정보의 수집·이용의 제한)**에 근거하여, 사용者の 동의 없이 수집·이용하지 않습니다. 개인정보는 다음의 목적을 위해 사용됩니다. - **회원가입 및 관리**: 서비스 이용을 위한 본인 확인 - **고객지원**: 문의사항 응대 및 서비스 개선 - **서비스 제공**: 맞춤형 서비스 제공 및 운영 2. **수집하는 개인정보 항목** 키즈카페 연결 플랫폼은 아래와 같은 개인정보를 수집합니다. - **필수항목**: 이름, 이메일 - **선택항목**: 주소, 전화번호, 생년월일 3. **개인정보의 보유 및 이용 기간** 키즈카페 연결 플랫폼은 수집된 개인정보를 **이용 목적이 달성된 후 즉시 파기**합니다. 다만, 법령에 따라 다음과 같이 일정 기간 보관될 수 있습니다. - **회원가입 정보**: 탈퇴 시 즉시 삭제 또는 익명화 - **계약 및 청약 철회 기록**: 5년간 보관 (전자상거래법 제6조 및 시행령 6조) - **소비자 불만 및 분쟁 처리 기록**: 3년간 보관 (전자상거래법 제6조) 4. **개인정보의 제3자 제공 여부 및 공유** 키즈카페 연결 플랫폼은 사용자의 동의 없이 개인정보를 제3자에게 제공하지 않습니다. 다만, 다음의 경우에 한해 예외적으로 제공될 수 있습니다. - **사용자의 사전 동의를 받은 경우 (개인정보 보호법 제17조)** - 제공 대상: 키즈카페 파트너사 - 제공 목적: 키즈카페 예약 및 결제 서비스 제공 - 제공 항목: 이름, 예약 정보 - **법령에 따라 제공이 요구되는 경우 (개인정보 보호법 제18조)** - 수사기관, 법원 등의 요청이 있을 경우 제공 가능 5. **개인정보 보호 조치 및 안전성 확보** 키즈카페 연결 플랫폼은 다음과 같은 방법으로 개인정보를 보호합니다. - **암호화**: 암호화를 사용하여 개인정보의 안전성을 보장합니다. - **보안프로그램 설치**: 보안프로그램을 설치하여 개인정보에 대한 접근을 제한합니다. - **개인정보 교육**: 모든 직원에게 개인정보 보호 교육을 제공하여 개인정보를 적절하게 처리할 수 있도록 합니다. 6. **사용자의 권리 및 행사 방법** 사용자는 언제든지 자신의 개인정보에 대해 다음과 같은 권리를 행사할 수 있습니다. - **열람 및 정정 권리**: 개인정보 보호법 제36조** - 부정확한 개인정보의 정정 또는 삭제 요청 권리 행사는 개인정보 보호책임자에게 이메일 또는 고객센터를 통해 요청할 수 있습니다. 7. **개인정보 보호법 제36조** - 부정확한 개인정보의 정정 또는 삭제 요청 권리 행사는 개인정보 보호책임자에게 이메일 또는 고객센터를 통해 요청할 수 있습니다. 8. **설치 목적**: 고객의 안전을 보호하는 목적으로 설치되는 개인정보 처리방침입니다. 9. **운영 기간**: 개인정보 보호법 제25조(영상정보처리기기 설치·운영 제한) - **보유 기간**: 30일 보관 후 자동 삭제 8. **개인정보 처리방침 변경 및 고지**: 본 개인정보 처리방침은 법률 개정 및 내부 정책에 따라 변경될 수 있으며, 변경 시 홈페이지 및 이메일을 통해 사전 공지합니다. 9. **개인정보 보호책임자 및 문의처**: 사용자는 개인정보 보호 관련 문의 및 불만을 아래 담당자에게 제기할 수 있습니다. - **개인정보 보호책임자**: 홍길동 - **이메일**: privacy@kidsplatform.com - **고객센터**: 02-1234-5678 본 방침은 2025년 *월*일부터 적용됩니다.

참고 문서

처리방침 예시'를 참고하여 '아동용 개인정보 처리방침'을 별도로 마련하여

있도록 개인정보 처리방침 수립 · 공개의 취지를 기재하는 것을 권장함

#14세 미만 아동의 개인정보를 처리하여 이를 처리방침에 안내할 때에는 이해하기

14세 미만 아동의 개인정보 처리에 관한 사항

공개하였다고 하더라도 개인정보 처리방침 전문을 공개하여야 함

개인정보의 처리와 보호에 관한 절차 및 기준을 안내하고, 이와 관련된 고충을

① #<개인정보처리자명>은(는) 14세 미만 아동의 개인정보를 처리하기 위하여 동의가

유지 · 관리, 서비스 부정이용 방지, 만 14세 미만 아동의 개인정보 처리 시

따라 작성 내용이 다를 수 있으므로, 각자의 상황에 맞게 개인정보 처리방침을 수립하여야 함

② #<개인정보처리자명>은(는) 14세 미만 아동의 개인정보 처리에 관하여 그

청크화와 오버랩에 따른 응답 변화

- 유사도 임계값 : 0.5, 문서검색수:5

쿼리: 2025년 국민의 삶은 어떻게 바뀌나요?

문서의 검색값 평균

: 가장 높은 값과 가장 낮은 벡터값을 뺀 중간의 평균을 계산

청크, 오버랩, splitter 설정에 따른 응답 차이 실험

<https://velog.io/@heyggun/LLM-LangChain-Chat-with-Your-Data-2-Data-Splitting>

청크와 오버랩 사이즈 설정에 따른 llm의 응답 차이 논문

https://www.manuscriptlink.com/society/kips/conference/ask2024/file/downloadSoConfManuscript/abs/KIPS_C2024A0276

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 100, 오버랩:10

```
(venv) etri12@etri12:~/flask5$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload   Total   Spent    Left  Speed
100  2084  100  2023  100    61      36      1  0:01:01  0:00:54  0:00:07  542
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n2025년에는 개인정보 수요가 증가할 것이며, 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께 새로운 형태의 직업과 직무가 등장할 것입니다. 이에 따라 업무 추진 여건 또한 변화하게 될 것입니다.\n\n문서의 내용을 바탕으로 분석해보면, 2025년에는 AI 시대의 개인정보 규율체계 혁신이 필요합니다. 이를 위해 AI 시대에 부합하는 개인정보 법제 정비가 필요한 것입니다. 또한 지속 가능한 신산업 혁신 기반 마련이 필요할 것입니다.\n\n이러한 변화는 국민의 삶을 크게 영향을 미칠 수 있습니다. 예를 들어, 데이터 활용 수요 확대와 함께 새로운 직업과 직무가 등장할 것이므로, 국민들은 새로운 형태의 직업을 찾거나 역량을 강화하여야 할 것입니다. 또한 AI 시대의 개인정보 규율체계 혁신으로 인해 국민들은 자신의 개인정보를 더 잘 보호해야 할 것입니다.\n\n따라서 2025년 국민의 삶은 새로운 형태의 직업과 직무, 개인정보保护 등 다양한 측면에서 변화하게 될 것입니다.",
  "context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....\n10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><b>r>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n1 업무 추진 여건\n(개인정보수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,<br><br>유사도: 0.6811 - III 2025년 6대 핵심 추진과제\nAI 시대 개인정보 규율체계 혁신\nAI 시대에 부합하는 개인정보 법제 정비\n유사도: 0.6688 - III 2025년 6대 핵심 추진과제\n.....\n4 AI 시대 개인정보 규율체계 혁신\n지속 가능한 신산업 혁신 기반 마련<br><br>"}
}
```

```
...,
  "context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....\n10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><b>r>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n1 업무 추진 여건\n(개인정보수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,<br><br>유사도: 0.6811 - III 2025년 6대 핵심 추진과제\nAI 시대 개인정보 규율체계 혁신\nAI 시대에 부합하는 개인정보 법제 정비\n유사도: 0.6688 - III 2025년 6대 핵심 추진과제\n.....\n4 AI 시대 개인정보 규율체계 혁신\n지속 가능한 신산업 혁신 기반 마련<br><br>"}
```

문서의 유사도 검색값

0.9087

0.8922

0.6903

0.6811

0.6688

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 300, 오버랩:30

```
(venv) etri12@etri12:~/flask$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Current
          Dload  Upload   Total Spent  Left  Speed
100  2084  100  2023  100    61      36      1  0:01:01  0:00:54  0:00:07  542
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n2025년에는 개인정보 수요가 증가할 것이며, 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께 새로운 형태의 직업과 직무가 등장할 것입니다. 이에 따라 업무 추진 여건 또한 변화하게 될 것입니다.\n\n문서의 내용을 바탕으로 분석해보면, 2025년에는 AI 시대의 개인정보 규율체계 혁신이 필요합니다. 이를 위해 AI 시대에 부합하는 개인정보 법제 정비가 필요한 것입니다. 또한 지속가능한 신산업 혁신 기반 마련이 필요할 것입니다.\n\n이러한 변화는 국민의 삶을 크게 영향을 미칠 수 있습니다. 예를 들어, 데이터 활용 수요 확대와 함께 새로운 직업과 직무가 등장할 것이므로, 국민들은 새로운 형태의 직업을 찾거나 역량을 강화하여야 할 것입니다. 또한 AI 시대의 개인정보 규율체계 혁신으로 인해 국민들은 자신의 개인정보를 더 잘 보호해야 할 것입니다.\n\n따라서 2025년 국민의 삶은 새로운 형태의 직업과 직무, 개인정보保护 등 다양한 측면에서 변화하게 될 것입니다.",

  "context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....
10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><br>
```

```
  "context_text": "유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><br>유사도: 0.8226 - ■■ 춤출하고 탄탄한 개인정보 안전망 구축\nIV 2025년, 국민의 삶이 이렇게 바뀝니다 ..... 10<br><br>유사도: 0.6996 - 순 서\nI. 정부 전반
기 성과 및 평가 ..... 1\nII. 2025년 업무추진 여건 및 방향 ..... 2\nIII. 2025년 6대 핵심 추진과제 ..... 4
■■ AI 시대 개인정보 규율체계 혁신\n■■ 지속가능한 신산업 혁신 기반 마련\n■■ 글로벌 개인정보 규범 주도권 확보\n■■ 마이데이터 시대 개막, 성과 창출 본격화\n■■ 개인정보 보호 컨트롤타워 역할 강화\n■■ 춤출하고 탄탄한 개인정보 안전망 구축<br><br>유사도: 0.6472 - - 안전한 개인정보, 신뢰받는 인공지능 시대 -\n2025. 1. 13.\n 개인정보보호위원회<br><br>유사도: 0.6258 - ■■ 마이데이터 시대 개막, 성과 창출 본격화\n■■ 마이데이터 본격 시행으로 국민 체감 성과 창출\n○의료·통신·에너지 등 국민생활 밀접 분야부터 마이데이터 시행으로\n국민이 자신의 정보를 이동시켜 원하는 서비스를 누리도록 지원(3월)\n○마이데이터 안착을 위한 마중물이 될 선도서비스 5종 단계적 출시\n■■ 全 분야 마이데이터 단계적 확산 추진\n○의료·통신 분야의 정보 전송자 및 전송항목 확대, 국민 편의성 등을 고려하여 교육·고용·여가 등 신규 부문 단계적 확대 논의
<br><br>
}
```

문서의 유사도 검색값

0.8922

0.8226

0.6996

0.6472

0.6258

→ 0.723133333

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 512, 오버랩:50

```
(venv) etri12@etri12:~/flask5$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Time  Current
                                         Dload  Upload   Total   Spent    Left  Speed
100  6398  100  6337    0    61      53    0  0:01:59  0:01:58  0:00:01  1422
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n2025년에 개인정보 보호와 신뢰받는 인공지능 시대를 맞이할 것입니다. 이를 위해서는 AI 시대에 부합하는 개인정보 보호 법제 정비가 필요합니다. 이에 따라 원본 데이터 활용을 개인정보위 심의·의결 하에 허용하는 특례를 마련할 계획입니다.\n\n또한, 신기술·신산업 혁신 촉진을 위해 AI 개발 사업자 등의 '정당한 이익'이나 '공익' 등을 고려하여 개인정보 적법 처리 근거를 확대할 것입니다. 이를 통해 국내 AI 생태계의 비약적인 발전 지원을 기대할 수 있습니다.\n\n이러한 노력으로 2025년에는 국민의 삶이 다음과 같은 변화에 직면하게 될 것입니다.\n\n1. **안전한 개인정보**: AI 시대의 등장으로 개인정보 보호가 매우 중요해질 것입니다. 이를 위해서는 AI 개발 사업자 등의 '정당한 이익'이나 '공익' 등을 고려하여 개인정보 적법 처리 근거를 확대할 계획입니다.\n\n2. **신뢰받는 인공지능**: 2025년에는 신뢰받는 인공지능 시대를 맞이하게 될 것입니다. 이를 위해서는 AI 개발 사업자 등의 '정당한 이익'이나 '공익' 등을 고려하여 개인정보 적법 처리 근거를 확대할 계획입니다.\n\n3. **데이터 기반 비즈니스·행정**: 2025년에는 데이터 기반 비즈니스·행정의 일상화가 늘어날 것입니다. 이를 위해서는 AI 개발 사업자 등의 '정당한 이익'이나 '공익' 등을 고려하여 개인정보 적법 처리 근거를 확대할 계획입니다.\n\n4. **프라이버시 침해 위험**: 2025년에는 프라이버시 침해 우려가 증대할 것입니다. 이를 위해서는 AI 개발 사업자 등의 '정당한 이익'이나 '공익' 등을 고려하여 개인정보 적법 처리 근거를 확대할 계획입니다.\n\n따라서 2025년에는 국민의 삶이 다음과 같은 변화에 직면하게 될 것입니다. 국민은 안심하고 살아갈 수 있는 안전한 개인정보 환경을 누리게 될 것이며, 신뢰받는 인공지능과 데이터 기반 비즈니스·행정을 경험할 수 있을 것입니다.",
```

문서의 유사도 검색값

0.8922

0.7517

0.6098

0.6030

0.5940

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 1024, 오버랩: 100

```
(venv) etri12@etri12:~/flask$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Time  Current
                                         Dload  Upload   Total  Spent   Left  Speed
100  5878  100  5817     0    61      64      0  0:01:30  0:01:30 --:--:--  1545
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n首先, 개인정보 보호는 새로운 패러다임에 놓여 있습니다. AI 시대 개인정보 규율체계 혁신은 우리나라를 포함한 주요국가들이 추진하고 있는 방향입니다. 정부는 프라이버시 침해 위협을 최소화하는 방향으로 다양한 접근법을 채택하고 있습니다.\n둘째, 데이터 경제 패권 경쟁이 심화되고 있습니다. 각 국가는 전략적 자산인 데이터 활용을 지원하면서 국가 안보와 자국민 보호를 목표로 개인정보 보호를 위한 법·제도적 기반도 강화하고 있습니다. 그러나 표준화된 AI·데이터 분야 국제규범 미비로 높은 규제비용을 감당할 수 있는 글로벌 벅테크 기업 등에 유리한 디지털 통상 환경 조성은 국가 안보와 자국민 보호를 고려해야 합니다.\n셋째, 중소·스타트업 대상 우수기술 개발 지원과 기술 수요처와의 1:1 매칭 컨설팅 등의 프로그램이 있습니다. 이러한 프로그램은 국제표준 선점을 위한 핵심 분야별(AI·자율주행·블록체인 등) 개인정보 기술표준 개발 및 표준 채택 지원에 기여할 것입니다.\n따라서, 2025년 국민의 삶이 어떻게 바뀌나요?それは 프라이버시 보호와 데이터 경제 패권 경쟁을 적절하게 균형 짜는 새로운 패러다임으로 향해 있습니다. 또한 중소·스타트업의 우수기술 개발 지원과 글로벌 벅테크 기업과의 협력을 통해 새로운 기술혁신을 촉진할 것입니다.",

  "context_text": "유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n유사도: 0.7517 - 순서\n정부 전반기 성과 및 평가 ..... 1\n2025년 업무추진 여건 및 방향 ..... 2\n2025년 6대 핵심 추진과제 ..... 4\nAI 시대 개인정보 규율체계 혁신
```

문서의 유사도 검색값

0.8922

0.7517

0.6098

0.5761

0.5635

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 100, 오버랩:30

```
(venv) etri12@etri12:~/flask$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq % Total    % Received % Xferd  Average Speed   Time     Time      Current  
          Dload  Upload   Total Spent  Left Speed  
100  1804  100  1743  100    61      44       1  0:01:01  0:00:38  0:00:23  369  
{
```

"answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n2025년에는 개인정보 보호 컨트롤타워 역할 강화가 이루어질 것이며, 촘촘하고 탄탄한 개인정보 안전망 구축이 필요합니다. 이를 통해 개인의 프라이버시를 보호하고, 데이터 활용 수요 확대에 부응할 수 있습니다.\n\n또한, 2025년에는 AI 시대 개인정보 규율체계 혁신이 이루어질 것이며, AI 시대에 부합하는 개인정보 법제 정비가 필요합니다. 이러한 혁신은 지속 가능한 신산업 혁신 기반 마련을 가능하게 하며, 국민의 삶을 크게 바꿀 것입니다.\n\n따라서 2025년에는 데이터 활용 수요 확대와 개인정보 보호 강화 등으로 인해 국민의 삶이 새로운 도약을 이루는 계기가 될 것 같습니다.",

```
"context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....  
10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><br>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n- 업무 추진 여건\n(개인정보 수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,<br><br>유사도: 0.6811 - III 2025년 6대 핵심 추진과제\n- AI 시대 개인정보 규율체계 혁신\n- AI 시대에 부합하는 개인정보 법제 정비<br><br>유사도: 0.6688 - III 2025년 6대 핵심 추진과제 ..... 4\n- AI 시대 개인정보 규율체계 혁신\n- 지속 가능한 신산업 혁신 기반 마련<br><br>"}
```

문서의 유사도 검색값

0.8922

0.7726

0.6903

0.6811

0.6688

→ 0.714666667

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 100, 오버랩:50

```
(venv) etri12@etri12:~/flask$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Time  Current
                                         Dload  Upload   Total   Spent   Left  Speed
100  2699  100  2638    0   61     40      0  0:01:05  0:01:05  --:--:--  684
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?는 매우 중요한 질문입니다. 문서의 내용을 기반으로 논리적인 답변을 제공하려 합니다.\n\n첫째, 개인정보 보호 컨트롤타워 역할 강화와 촘촘하고 탄탄한 개인정보 안전망 구축이 있습니다. 이러한 제정은 국민의 삶에 있어 매우 중요합니다. 2025년, 인공지능(AI) 시대의 도래로 개인정보 보호가 더욱 중요한 관심사로 떠오르게 되겠으며, 이러한 제정들은 AI 시대 개인정보 규율체계 혁신을 촉진할 것입니다.\n\n둘째, 업무 추진 여건 및 방향이 있습니다. 2025년, 신기술 분야를 중심으로 데이터 활용 수요가 확대됩니다. 이러한 데이터 활용은 국민의 삶에 있어 다양한 oportunities를 제공하게 됩니다. 그러나 이러한 데이터 활용에 있어 개인정보 보호가 매우 중요해질 것입니다.\n\n셋째, 6대 핵심 추진과제 중 하나는 AI 시대 개인정보 규율체계 혁신입니다. 이러한 혁신은 국가의 안정성과 국민의 삶을 지탱하는 데 중요한 역할을 수행하게 될 것입니다.\n\n따라서, 2025년 국민의 삶이 어떻게 바뀌나요?는 다음과 같습니다. AI 시대의 도래로 개인정보 보호가 더욱 중요해질 것이며, 촘촘하고 탄탄한 개인정보 안전망 구축이 중요해질 것입니다. 또한, 데이터 활용 수요 확대와 함께 업무 추진 여건 및 방향이 중요해질 것입니다.\n\n결론적으로, 2025년 국민의 삶은 AI 시대의 도래로 개인정보 보호를 중시하고, 촘촘하고 탄탄한 개인정보 안전망 구축을 강조할 것입니다.",
```

"context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다
10

유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -

r>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n- 11 업무 추진 여건\n□(개인정보
보수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,

유사도:
0.6811 - III 2025년 6대 핵심 추진과제\n- AI 시대 개인정보 규율체계 혁신\n□AI 시
대에 부합하는 개인정보 법제 정비

유사도: 0.6688 - III 2025년 6대 핵심 추진
과제 4\n- AI 시대 개인정보 규율체계 혁신\n- 지속가능한
신산업 혁신 기반 마련

"
}

문서의 유사도 검색값

0.8922

0.7726

0.6903

0.6811

0.6688

www.ijerpi.org

→ 0.714666667

청크화와 오버랩에 따른 검색 결과 변화

- 청크 사이즈: 100, 오버랩:0

```
.1:8000/uask_pdf -H "Content-Type: application/json" -d '{"query": "2025년
국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time      Time      Current
                                 Dload  Upload Total Spent   Left Speed
100  1826  100  1765  100    61     38      1  0:01:01  0:00:45  0:00:16  450
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n기준 문헌에서 제시한 정보를
기반으로, 2025년에는 다음과 같은 변화들이 예상됩니다.\n**개인정보수요증대**:
신기술 분야 중심으로 데이터 활용 수요가 확대됩니다. 따라서 AI 시대에 부합하는 개
인정보 법제 정비가 필요하게 됩니다. 이에 따라 개인정보 규율체계 혁신이 있을 것입
니다.\n**업무 추진 여건**:
업무 추진 여건이 개선되게 되며, 신산업 혁신 기반 마련이 될 것입니다. 이러한 변화들은 국민의 삶을 지속적으로 향상시킬 것이며, 새로운 기회를 제공할 것입니다.\n따라서 2025년에는 개인정보보호 및 데이터 활용, 업무
추진 여건 개선 등 다양한 변화들이 예상됩니다. 이러한 변화들은 국민의 삶을 다양하
게 영향을 미칠 것이며, 새로운 시대를 열어놓게 될 것입니다.",
  "context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....
10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><b
r>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n1 업무 추진 여건\n(개인정보
수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,<br><br>유사도:
0.6811 - III 2025년 6대 핵심 추진과제\nAI 시대 개인정보 규율체계 혁신\nAI 시
대에 부합하는 개인정보 법제 정비<br><br>유사도: 0.6688 - III 2025년 6대 핵심 추
진과제 ..... 4 AI 시대 개인정보 규율체계 혁신\n지속 가능한
신산업 혁신 기반 마련<br><br>"}
```

```
"context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....
10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><b
r>유사도: 0.6903 - II 2025년 업무 추진 여건 및 방향\n1 업무 추진 여건\n(개인정보
수요증대) 신기술 분야를 중심으로 데이터 활용 수요 확대와 함께,<br><br>유사도:
0.6811 - III 2025년 6대 핵심 추진과제\nAI 시대 개인정보 규율체계 혁신\nAI 시
대에 부합하는 개인정보 법제 정비<br><br>유사도: 0.6688 - III 2025년 6대 핵심 추
진과제 ..... 4 AI 시대 개인정보 규율체계 혁신\n지속 가능한
신산업 혁신 기반 마련<br><br>"}
```

문서의 유사도 검색값

0.9087

0.8922

0.6903

0.6811

0.6688

→ 0.754533333

청크화와 오버랩에 따른 응답 변화

- 청크 사이즈: 100, 오버랩:50 / 0 - 응답 비교

```
(venv) etri12@etri12:~/flasks$ curl -X POST http://127.0.0.1:8000/ask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Time  Current
          Dload  Upload Total Spent   Left Speed
100  2699  100  2638    0   61      40      0  0:01:05  0:01:05  --:--:--  684
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?는 매우 중요한 질문입니다. 문서의 내용을 기반으로 논리적인 답변을 제공하려 합니다.\n\n첫째, 개인정보 보호 컨트롤타워 역할 강화와 촘촘하고 탄탄한 개인정보 안전망 구축이 있습니다. 이러한 제정은 국민의 삶에 있어 매우 중요합니다. 2025년, 인공지능(AI) 시대의 도래로 개인정보 보호가 더욱 중요한 관심사로 떠오르게 되겠으며, 이러한 제정들은 AI 시대 개인정보 규율체계 혁신을 촉진할 것입니다.\n\n둘째, 업무 추진 여건 및 방향이 있습니다. 2025년, 신기술 분야를 중심으로 데이터 활용 수요가 확대됩니다. 이러한 데이터 활용은 국민의 삶에 있어 다양한 opportunities를 제공하게 됩니다. 그러나 이러한 데이터 활용에 있어 개인정보 보호가 매우 중요해질 것입니다.\n\n셋째, 6대 핵심 추진 과제 중 하나는 AI 시대 개인정보 규율체계 혁신입니다. 이러한 혁신은 국가의 안정성과 국민의 삶을 지탱하는 데 중요한 역할을 수행하게 될 것입니다.\n\n따라서, 2025년 국민의 삶이 어떻게 바뀌나요?는 다음과 같습니다. AI 시대의 도래로 개인정보 보호가 더욱 중요해질 것이며, 촘촘하고 탄탄한 개인정보 안전망 구축이 중요해질 것입니다. 또한, 데이터 활용 수요 확대와 함께 업무 추진 여건 및 방향이 중요해질 것입니다.\n\n결론적으로, 2025년 국민의 삶은 AI 시대의 도래로 개인정보 보호를 중시하고, 촘촘하고 탄탄한 개인정보 안전망 구축을 강조할 것입니다.",
```

```
.1:8000/uask_pdf -H "Content-Type: application/json" -d '{"query": "2025년 국민의 삶이 어떻게 바뀌나요?"}' | jq
% Total    % Received % Xferd  Average Speed   Time     Time      Time  Current
          Dload  Upload Total Spent   Left Speed
100  1826  100  1765  100    61      38      1  0:01:01  0:00:45  0:00:16  450
{
  "answer": "2025년 국민의 삶이 어떻게 바뀌나요?\n\n기존 문헌에서 제시한 정보를 기반으로, 2025년에는 다음과 같은 변화들이 예상됩니다.\n\n**개인정보수요증대**: 신기술 분야 중심으로 데이터 활용 수요가 확대됩니다. 따라서 AI 시대에 부합하는 개인정보 법제 정비가 필요하게 됩니다. 이에 따라 개인정보 규율체계 혁신이 있을 것입니다.\n\n**업무 추진 여건**: 업무 추진 여건이 개선되며, 신산업 혁신 기반 마련이 될 것입니다. 이러한 변화들은 국민의 삶을 지속적으로 향상시킬 것이며, 새로운 기회를 제공할 것입니다.\n\n따라서 2025년에는 개인정보보호 및 데이터 활용, 업무 추진 여건 개선 등 다양한 변화들이 예상됩니다. 이러한 변화들은 국민의 삶을 다양하게 영향을 미칠 것이며, 새로운 시대를 열어놓게 될 것입니다.",\n  "context_text": "유사도: 0.9087 - IV 2025년, 국민의 삶이 이렇게 바뀝니다 ....\n10<br><br>유사도: 0.8922 - IV 2025년, 국민의 삶이 이렇게 바뀝니다.\n- 10 -<br><br>
```

청크화와 오버랩에 따른 응답 변화

4. 연구 결과

세팅 1: Chunk size = 400, Overlap size = 150

Query: 운수 좋은 날의 결말은 뭐야?

Answer

죄송합니다만, 제공된 정보에는 "운수 좋은 날"의 결말에 대한 정보가 포함되어 있지 않습니다.

세팅 2: Chunk size = 600, Overlap size = 50

Query: 운수 좋은 날의 결말은 뭐야?

Answer

"운수 좋은 날"의 결말은 김첨지가 그의 아내가 죽은 것을 발견하고 그녀의 얼굴에 대고 중얼거리는 장면으로 끝납니다. 그는 아내에게 설령탕을 사다놓았는데 왜 먹지를 못하냐고 묻습니다. 이는 김첨지가 아내의 죽음을 받아들이지 못하고 있는 상황을

2. 문서 분할

주어진 문서에 대해 융용 프로그램 혹은 답변 생성 모델에 더 적합하도록 문서를 변환하고 싶은 경우는 빈번하다. 이를 위한 가장 간단한 방법은 긴 문서를 모델의 컨텍스트 창에 맞게 작은 사이즈로 분할 하는 방법이다. 하지만 긴 텍스트 문서를 여러 작은 덩어리로 분할하는 것도 잠재적인 복잡성이 많이 있다. 이상적으로는 의미상 관련된 텍스트 조각을 함께 유지하는 것이 좋다. 하지만 "의미상 관련"이 의미하는 바는 텍스트 유형에 따라 달라질 수 있다. 통상적으로 텍스트 분할기는 다음과 같이 작동

https://www.manuscriptlink.com/society/kips/conference/ask2024/file/downloadSoConfManuscript/abs/KIPS_C2024A0276

청크화와 오버랩에 따른 응답 변화

결론 :

오버랩이 클 때 (30,50) → 평균 유사도가 감소함 (약 0.71)

- 서로 겹치는 내용이 많이 생기면서 하나의 청크에 여러 맥락이 포함될 가능성이 높아짐
- 검색 시 비슷한 내용이 반복적으로 포함된 청크들이 검색될 가능성이 증가한다.
- 특정 키워드나 문맥과의 유사도가 상대적으로 고르게 분산됨 → 평균 유사도의 하락

→ 맥락은 더 풍부해질 수 있으나, 특정 쿼리와의 직접적인 유사도가 낮아질 수 있다.

오버랩이 작을 때(0,10) → 평균 유사도가 증가함 (약 0.75)

- 청크 간 중복이 적기 때문에 각 청크가 독립적인 의미를 가질 가능성이 높다
- 검색 시 해당 쿼리와 가장 일치하는 청크가 독립적으로 선택된다.
- 유사도가 높은 특정 청크가 검색될 확률이 높아지면서 평균 유사도가 증가할 수 있다.

→ 더 직접적으로 쿼리와 연관될 가능성이 높아지고, 유사도가 높아질 가능성이 있다.

즉, 문서의 성질에 따라 검색 결과가 달라짐을 알 수 있었다

감사합니다.