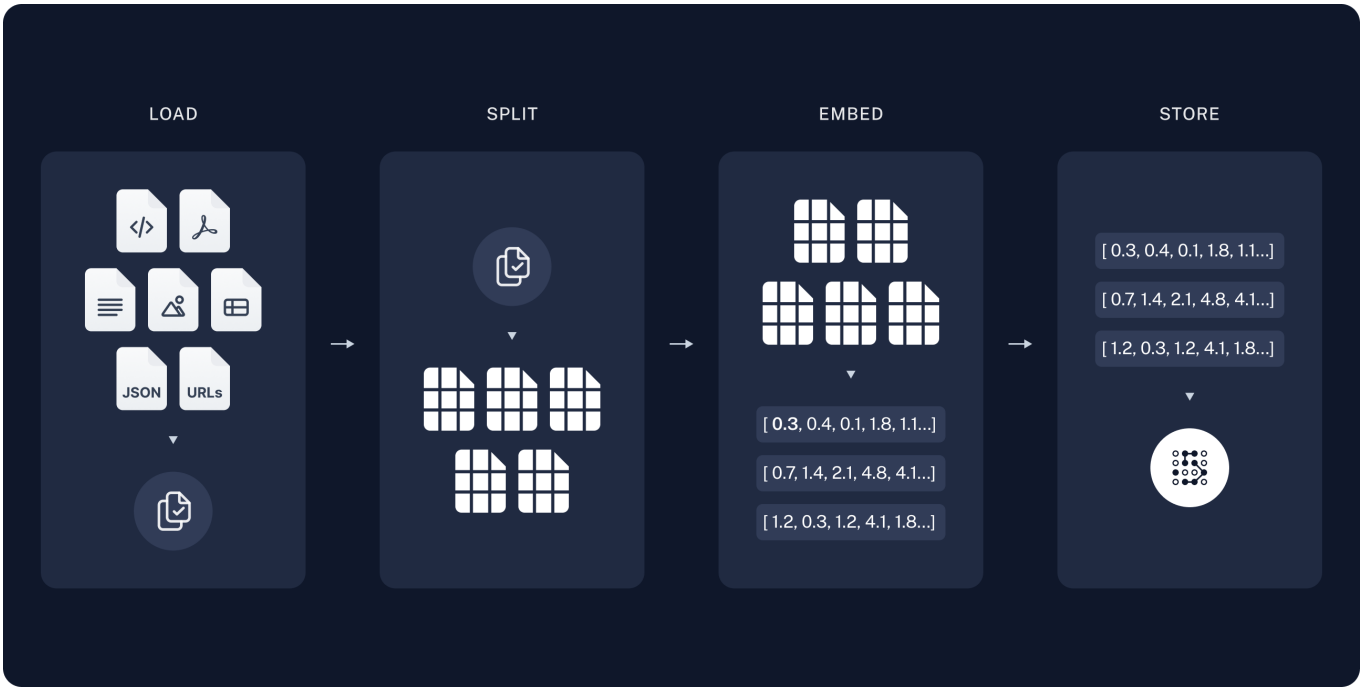
텍스트 분할 (Text Splitter)

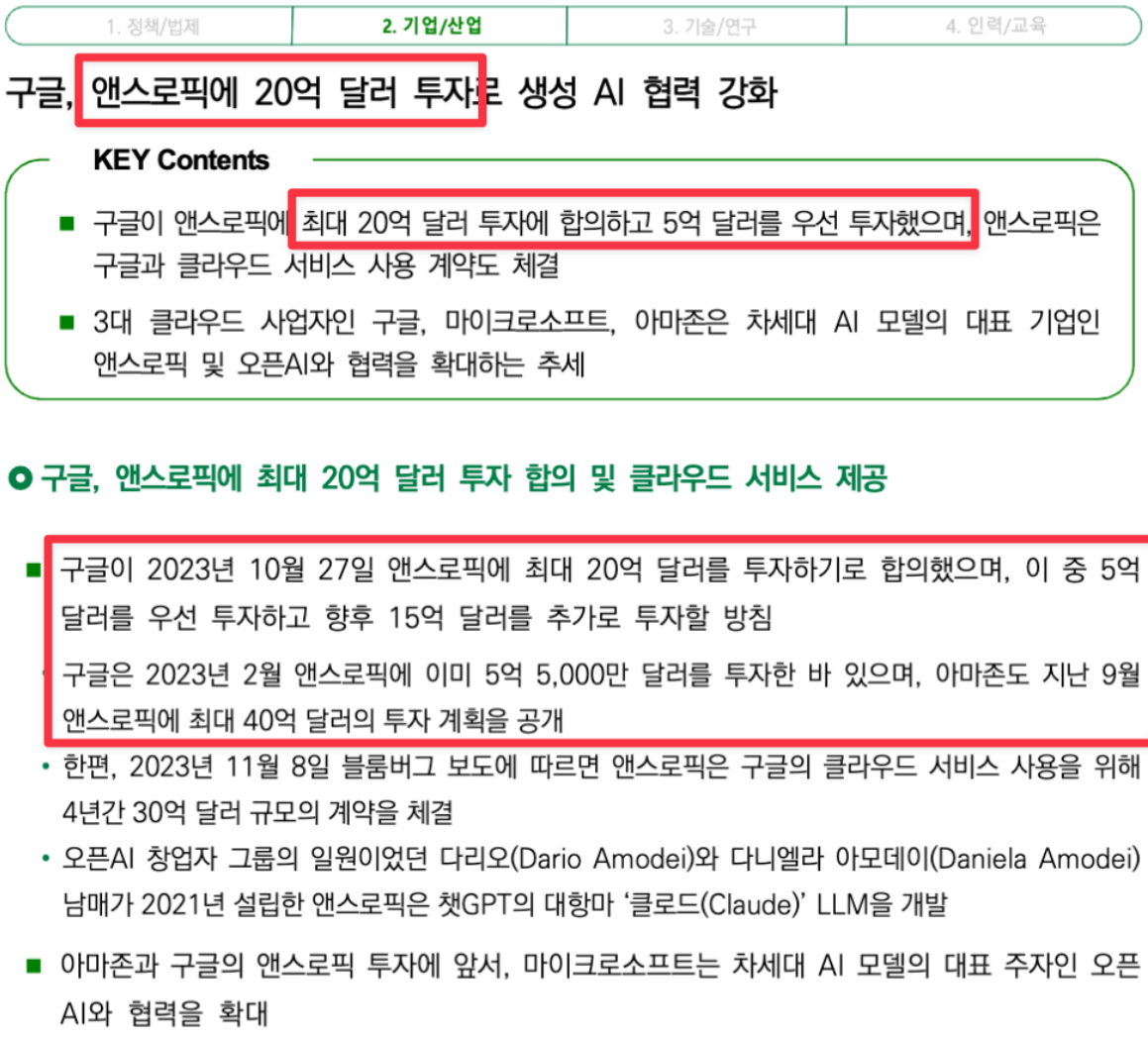
Fast campus



문서분할은 Retrieval-Augmented Generation(RAG) 시스템의 두 번째 단계로서, 로드된 문서들을 **효율적으로 처리**하고, 시스템이 정보를 보다 잘 활용할 수 있도록 준비하는 중요한 과정입니다.

이 단계의 목적은 크고 복잡한 문서를 LLM 이 받아들일 수 있는 **효율적인 작은 규모의 조각으로 나누는 작업**입니다. 나중에 사용자가 입력한 질문에 대하여 보다 효율적인 정보만 압축/선별하여 가져오기 위함입니다.

(예시)
구글이 앤스로픽에 투자한 금액은 얼마야?



1. 정책/법제 | 2. 기업/산업 | 3. 기술/연구 | 4. 인력/교육

구글, 앤스로픽에 20억 달러 투자로 생성 AI 협력 강화

KEY Contents

- 구글이 앤스로픽에 최대 20억 달러 투자에 합의하고 5억 달러를 우선 투자했으며, 앤스로픽은 구글과 클라우드 서비스 사용 계약도 체결
- 3대 클라우드 사업자인 구글, 마이크로소프트, 아마존은 차세대 AI 모델의 대표 기업인 앤스로픽 및 오픈AI와 협력을 확대하는 추세

○ 구글, 앤스로픽에 최대 20억 달러 투자 합의 및 클라우드 서비스 제공

- 구글이 2023년 10월 27일 앤스로픽에 최대 20억 달러를 투자하기로 합의했으며, 이 중 5억 달러를 우선 투자하고 향후 15억 달러를 추가로 투자할 방침
  - 구글은 2023년 2월 앤스로픽에 이미 5억 5,000만 달러를 투자한 바 있으며, 아마존도 지난 9월 앤스로픽에 최대 40억 달러의 투자 계획을 공개
  - 한편, 2023년 11월 8일 블룸버그 보도에 따르면 앤스로픽은 구글의 클라우드 서비스 사용을 위해 4년간 30억 달러 규모의 계약을 체결
  - 오픈AI 창업자 그룹의 일원이었던 다리오(Dario Amodei)와 다니엘라 아모데이(Daniela Amodei) 남매가 2021년 설립한 앤스로픽은 챗GPT의 대항마 '클로드(Claude)' LLM을 개발
- 아마존과 구글의 앤스로픽 투자에 앞서, 마이크로소프트는 차세대 AI 모델의 대표 주자인 오픈AI와 협력을 확대

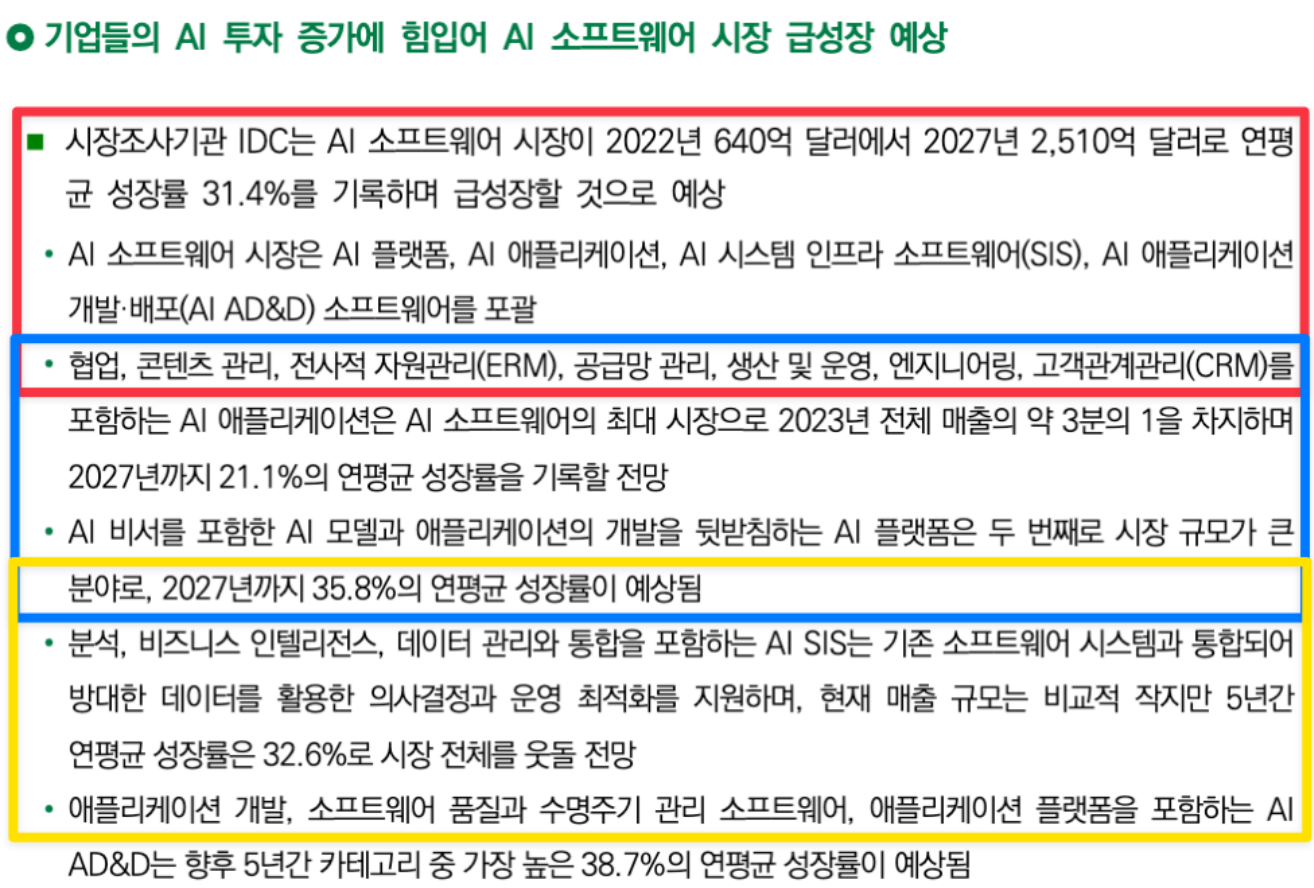
분할의 필요성

1. **핀포인트 정보 검색(정확성)**: 문서를 세분화함으로써 **질문(Query) 에 연관성이 있는 정보**만 가져오는데 도움이 됩니다. 각각의 단위는 특정 주제나 내용에 초점을 맞추므로, **관련성이 높은 정보를 제공**합니다.
2. **리소스 최적화(효율성)**: 전체 문서를 LLM 으로 입력하게 되면 비용이 많이 발생할 뿐더러, 효율적인 답변을 많은 정보속에 발췌하여 답변하지 못하게 됩니다. 때로는 이러한 문제가 **할루시네이션**으로 이어지게 됩니다. 따라서, 답변에 필요한 정보만 발췌하기 위한 목적도 있습니다.

문서분할 과정

1. **문서 구조 파악**: PDF 파일, 웹 페이지, 전자 책 등 다양한 형식의 문서에서 구조를 파악합니다. 이는 문서의 헤더, 푸터, 페이지 번호, 섹션 제목 등을 식별하는 과정을 포함할 수 있습니다.
2. **단위 선정**: 문서를 어떤 단위로 나눌지 결정합니다. 이는 페이지별, 섹션별, 또는 문단별일 수 있으며, 문서의 내용과 목적에 따라 다릅니다.
3. **단위 크기 선정(chunk size)**: 문서를 몇 개의 토큰 단위로 나눌 것인지를 정합니다.
4. **청크 오버랩(chunk overlap)**: 분할된 끝 부분에서 맥락이 이어질 수 있도록 일부를 겹쳐서(overlap) 분할하는 것이 일반적입니다.

청크 크기 & 청크 오버랩



코드

```
from langchain_text_splitters import
RecursiveCharacterTextSplitter
```

```
# 단계 2: 문서 분할(Split Documents)
text_splitter =
RecursiveCharacterTextSplitter(chunk_size=1000,
chunk_overlap=50)
splits = text_splitter.split_documents(docs)
```

참고

- 텍스트 분할기
- LangChain TextSplitters