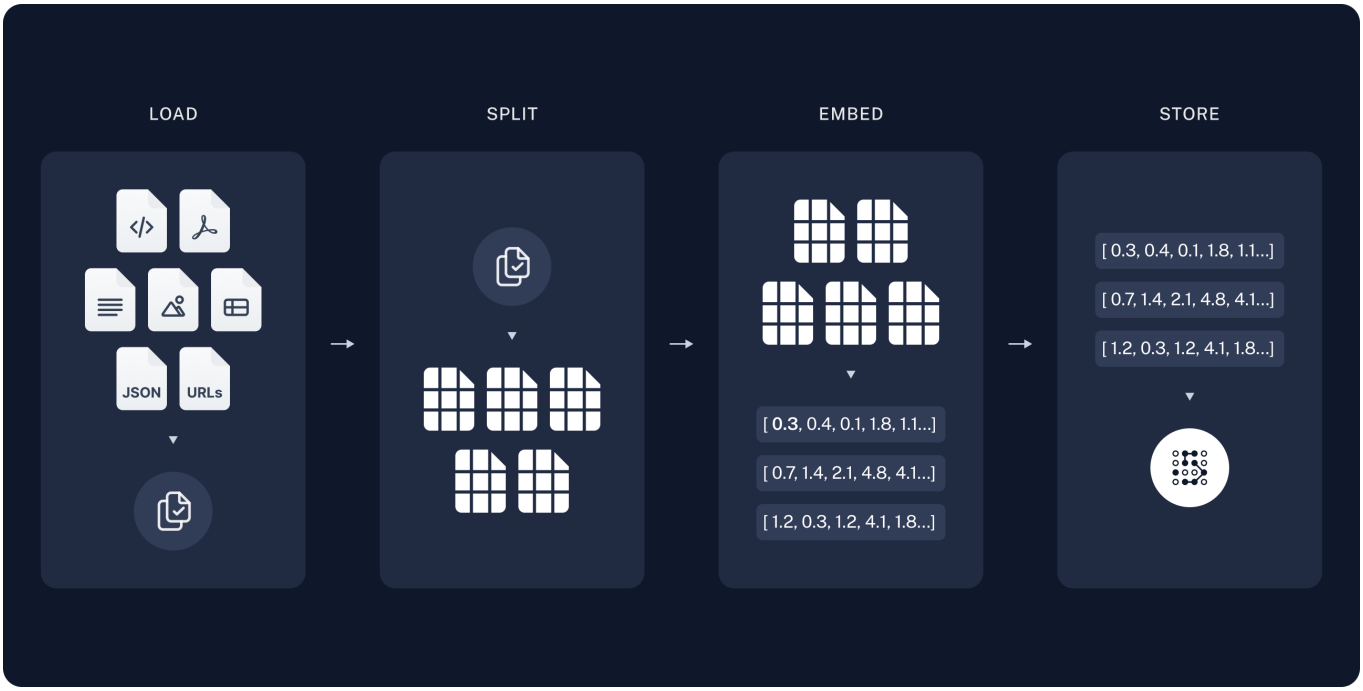
도큐먼트 로드 (Document Loader)

Fast campus



도큐먼트 로드 단계는 Retrieval-Augmented Generation(RAG) 시스템의 첫 번째 단계로, 전체 프로세스에서 매우 중요한 **기초 작업을 수행**합니다. 이 단계의 주된 목적은 RAG 시스템이 필요로 하는 **데이터를 외부 소스로부터 효율적으로 수집하고 준비**하는 것입니다.

1. 데이터 소스의 선택

도큐먼트 로더는 먼저 어떤 종류의 데이터가 필요한지, 그리고 그 데이터를 어디서 어떻게 수집할지 결정해야 합니다. 데이터 소스는 **웹사이트, 데이터베이스, API, 공개 데이터 세트** 등 다양할 수 있습니다. 예를 들어, **최신 뉴스 데이터**가 필요하다면 실시간 뉴스 API나 크롤링(`WebBaseLoader`)를 사용할 수 있고, **과학적 논문이나 전문 지식이 필요하다면 학술 데이터베이스**(`ArxivLoader`)를 사용할 수 있습니다.

2. 데이터 수집

선택된 데이터 소스에서 필요한 데이터를 수집하는 작업을 진행합니다. 이 과정은 API 호출, 웹 스크래핑, 데이터베이스 쿼리 등을 포함할 수 있으며, 때로는 인증이나 접근 권한 설정이 필요할 수도 있습니다.

랭체인 도큐먼트 로더

주요 작업

- PDF 문서를 로드
- Word, 한글 문서 로드
- Excel, CSV, SQL Table 로드
- 마크다운(.md) 파일 로드
- HTML 문서 로드

3. 데이터 필터링과 전처리

수집된 데이터 중에서 RAG 시스템에 필요한 정보만을 추출하고, 필요에 따라 데이터를 정제하는 과정입니다. 이 단계에서는 **불필요한 포맷팅을 제거**하거나, 언어나 문맥에 맞는 필터를 적용할 수 있습니다.

주요 작업

- 불필요한 이미지 제거
- 그래프 제거
- License 표기 등 제거
- 그 외 답변에 포함되지 말아야할 정보 제거

4. 데이터 로드

전처리된 데이터를 RAG 시스템 내부적으로 사용할 수 있는 형식으로 변환하고 로드합니다. 이는 보통 메모리 내 데이터 구조로의 변환을 포함하며, 필요에 따라 데이터베이스나 파일 시스템에 저장할 수도 있습니다.

예시

뉴스 데이터 로드

예를 들어, RAG 시스템이 세계 뉴스에 대한 질문에 답변하기 위해 사용되는 경우, 도큐먼트 로더는 아래와 같은 작업을 수행할 수 있습니다.

- **데이터 소스 선택**: 신뢰할 수 있는 뉴스 서비스(뉴스 사이트)를 선택합니다.
- **데이터 수집**: API 혹은 크롤링을 통해 최신 뉴스 기사 데이터를 수집합니다.
- **데이터 필터링 및 전처리**: 수집된 기사 중 특정 주제(예: 국제 정치, 경제)에 해당하는 기사만을 선택하고, 필요없는 광고나 메타데이터를 제거합니다.
- **데이터 로드**: 전처리된 뉴스 기사를 시스템의 데이터베이스에 저장하거나, 직접 메모리로 로드하여 다음 처리 단계인 문서분할로 넘어갑니다.

PDF 문서 데이터 로드

PDF 문서를 데이터 소스로 사용하는 경우의 도큐먼트 로더 작업은 특히 문서가 정형화되어 있지 않거나 다양한 포맷을 포함하고 있을 때 복잡할 수 있습니다. 아래는 PDF 문서를 이용해 특정 학술 주제에 대한 데이터를 로드하는 과정을 단계별로 설명합니다.

- **데이터 소스 선택**: 도큐먼트 로더는 사용할 PDF 문서의 출처를 결정합니다. 이는 대학교의 디지털 라이브러리, 연구 기관의 데이터베이스, 또는 공개적으로 접근 가능한 학술 문서 저장소일 수 있습니다.(예. arxiv)
- **데이터 수집**: 선택된 출처에서 필요한 주제와 관련된 PDF 문서를 수집합니다. 이 과정은 직접 다운로드하거나, 필요한 경우 API를 통해 자동으로 문서를 수집할 수 있습니다.
- **데이터 필터링과 전처리**: 수집된 PDF 문서들은 다양한 형태의 정보(텍스트, 이미지, 표, 등)를 포함할 수 있으므로, 이 정보를 RAG 시스템이 활용할 수 있는 형태로 변환하는 것이 필요합니다. 이를 위해 PDF 문서를 텍스트로 변환하는 작업이 필요하며, 이 과정에서 **OCR(Optical Character Recognition)** 기술을 사용할 수도 있습니다. **텍스트로 변환된 데이터는 불필요한 요소를 제거하고, 필요에 따라 문장을 재구성**하여 더 명확하게 합니다.

- **데이터 로드**: 전처리된 텍스트 데이터를 시스템이 사용할 수 있는 형태로 메모리에 로드하거나, 데이터베이스에 저장합니다. 이 데이터는 후속 단계인 문서 분할, 임베딩을 위한 기초 자료로 사용됩니다.

코드

```
# 단계 1: 문서 로드(Load Documents)
from langchain_community.document_loaders import
PyMuPDFLoader

loader = PyMuPDFLoader("data/문서.pdf")
docs = loader.load()
```

이러한 과정을 통해 RAG 시스템은 필요한 데이터를 효과적으로 활용할 준비를 마치게 됩니다. 데이터 로드는 시스템의 성능과 직결되므로 매우 중요하며, 이 단계에서의 효율과 정확성이 전체 시스템의 출력 품질에 큰 영향을 미칩니다.