
- 출처: LangChain 공식 문서 또는 해당 교재명
- 원본 URL: https://smith.langchain.com/hub/teddynote/summary-stuff-documents

---

`06_문서 로더(Document Loader)`

- *LangChain에서 사용되는 주요 로더*
- *LangChain에서 사용되는 로더 목록*
  - `load()`, `aload()`, `lazy_load()` 등 다양한 기능 제공

- `실습에 활용한 문서`
  - *소프트웨어정책연구소(SPRi) - 2023년 12월호*

> 저자: 유재흥(AI정책연구실 책임연구원), 이지수(AI정책연구실 위촉연구원)
>
> 링크: https://spri.kr/posts/view/23669
>
> 파일명: `SPRI_AI_Brief_2023년12월호_F.pdf`

`Document`

- `LangChain`의 기본 문서 객체

`속성`

- `page_content`: 문서의 `내용`을 나타내는 `문자열`
- `metadata`: 문서의 `매타데이터`를 나타내는 `딕셔너리`

```
# API KEY를 환경변수로 관리하기 위한 설정 파일
import os
from dotenv import load_dotenv

# API KEY 정보로드
load_dotenv()                # true
```

```
from langchain_google_genai import ChatGoogleGenerativeAI
from langchain_core.documents import Document

document = Document(page_content="안녕하세요? 이건 랭체인의 도큐먼트 입니다")
```

```
# 도큐먼트의 속성 확인

document.__dict__
```

- 셀 출력

```
{'id': None,
'metadata': {},
'page_content': '안녕하세요? 이건 랭체인의 도큐먼트 입니다',
'type': 'Document'}
```

- `metadata`에 속성 추가

```
# 메타데이터 추가

document.metadata["source"] = "TeddyNote"
document.metadata["page"] = 1
document.metadata["author"] = "Teddy"
```

```
# 도큐먼트의 속성 확인

document.metadata
```

- 셀 출력

```
{'source': 'TeddyNote', 'page': 1, 'author': 'Teddy'}
```

```
# 메타데이터 추가_연습_2

document.metadata["source"] = "Jay"
document.metadata["page"] = 2
document.metadata["author"] = "Jay"
```

```
# 도큐먼트의 속성 확인
```

```
document.metadata
```

- 셀 출력

```
{'source': 'Jay', 'page': 2, 'author': 'Jay'}
```

Document Loader

- 다양한 파일의 형식으로부터 불어온 문서(`Document`) 객체로 변환하는 역할

주요 Loader

- **`PyPDFLoader`**: `PDF 파일`을 로드하는 로더
- **`CSVLoader`**: `CSV 파일`을 로드하는 로더
- **`UnstructuredHTMLLoader`**: `HTML 파일`을 로드하는 로더
- **`JSONLoader`**: `JSON 파일`을 로드하는 로더
- **`TextLoader`**: `텍스트 파일`을 로드하는 로더
- **`DirectoryLoader`**: `디렉토리`를 로드하는 로더

```
# 예제 파일 경로

FILE_PATH = "../06_Document_Loader/data/SPRI_AI_Brief_2023년12월호_F.pdf"
```

```
from langchain_community.document_loaders import PyPDFLoader


# 로더 설정
loader = PyPDFLoader(FILE_PATH)
```

---

load()

- 문서를 로드하여 반환
- 반환된 결과 = **`List[Document]`**

```
# PDF 로더
docs = loader.load()
```

```
# 로드된 문서의 수 확인
len(docs)                                    # 23
```

```
# 첫번째 문서 확인

docs[0]
```

- 셀 출력

```
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.(
```

`load_and_split()`

- `splitter` 사용 → 문서 분할 / 반환
- 반환된 결과 = **`List[Document]`**

```
from langchain_text_splitters import CharacterTextSplitter

# 문열 분할기 설정
text_splitter = CharacterTextSplitter(chunk_size=200, chunk_overlap=0)

# 문서 분할
docs = loader.load_and_split(text_splitter=text_splitter)

# 로드된 문서의 수 확인
len(docs)                                    # 23

# 첫번째 문서 확인
docs[0]
```

- 셀 출력

```
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.(
```

`lazy_load()`

- `generate` 방식 = 문서 로드

```
# generator 방식으로 문서 로드
for doc in loader.lazy_load():
    print(doc.metadata)
```

- 셀 출력

```
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
{'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.0.0.13462', 'creat
```

`aloda()`

- `비동기(Async)` 방식의 문서 로드

```
# 문서를 async 방식으로 로드

adocs = loader.aload()
```

```
# 문서 로드

await adocs
```

- 셀 출력

```
[Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
Document(metadata={'producer': 'Hancom PDF 1.3.0.542', 'creator': 'Hwp 2018 10.
```

- *next:* ***PDF***
