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

---

`Microsoft Word`

- `Microsoft Word`
  - `Microsoft`에서 개발한 워드 프로세서
  - `Word` 문서룰 사용할 수 있는 문서 형식으로 로드하는 방법 다루기

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

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

`Docx2txtLoader`

- `Docx2txt` 사용 → `.docx` 파일을 문서로 로드 가능
  - 사전에 `VS Code` 터미널에 설치할 것

```
pip install -qU docx2txt
pip install msoffcrypto-tool                                    # msoffcrypto 패키지
```

```
from langchain_community.document_loaders import Docx2txtLoader

# 문서 로더 초기화
loader = Docx2txtLoader("../06_Document_Loader/data/sample-word-document.docx")

# 문서 로딩
docs = loader.load()

print(len(docs))                # 1
```

`UnstructuredWordDocumentLoader`

- 사전에 `VS Code` 터미널에 설치할 것

```
pip install python-docx
```

```
from langchain_community.document_loaders import UnstructuredWordDocumentLoader

# 비구조화된 워드 문서 로더 인스턴스화
loader = UnstructuredWordDocumentLoader("../06_Document_Loader/data/sample-word-document.docx")

# 문서 로드
docs = loader.load()

print(len(docs))                # 1
```

- 1개의 단일 `Document`로 로드됨

```
# metadata 출력

print(docs[0].metadata)
```

- 셀 출력

```
{'source': '../06_Document_Loader/data/sample-word-document.docx'}
```

- 내부적으로 비정형은 텍스트 덩어리마다 서로 다른 **요소** 를 만듦
- 기본적으로 이들은 함께 결합되어 있지만 **mode="elements"** 지정 → 쉽게 분리 가능

```
# UnstructuredWordDocumentLoader에서 elements 모드 지정

loader = UnstructuredWordDocumentLoader(
    "../06_Document_Loader/data/sample-word-document.docx",
    mode="elements"
)

# 데이터 로드
docs = loader.load()

# 로드한 문서의 개수 출력
print(len(docs))                        # 128 (셀 출력: 0.8s)
```

```
# 첫번째 문서의 내용 출력

print(docs[0].page_content)
```

- 셀 출력

```
Semantic Search
```

```
# 첫번째 문서의 내용 출력

docs[0].metadata
```

- 셀 출력

```
{'source': '../06_Document_Loader/data/sample-word-document.docx',
'category_depth': 0,
'file_directory': '../06_Document_Loader/data',
'filename': 'sample-word-document.docx',
'last_modified': '2025-09-12T13:42:46',
'page_number': 1,
'languages': ['kor'],
'filetype': 'application/vnd.openxmlformats-officedocument.wordprocessingml.document',
'category': 'UncategorizedText',
'element_id': 'a7703edf875ec776dc2bb839ca335b45'}
```

- *next:* *PPT* (*Power Point*)
