to PDF ▶

Word Documents

Contents

- Retain Elements

This covers how to load Word documents into a document format that we can use downstream.

```python
from langchain.document_loaders import UnstructuredWordDocumentLoader
```

```python
loader = UnstructuredWordDocumentLoader("example_data/fake.docx")
```

```python
data = loader.load()
```

```python
data
```

```
[Document(page_content='Lorem ipsum dolor sit amet.', lookup_str='', metadata=
{'source': 'fake.docx'}, lookup_index=0)]
```

Retain Elements

Under the hood, Unstructured creates different "elements" for different chunks of text. By default we combine those together, but you can easily keep that separation by specifying `mode="elements"`.

```python
loader = UnstructuredWordDocumentLoader("example_data/fake.docx", mode="elements")
```

```python
data = loader.load()
```

Skip to main content

```
data[0]
```

```
Document(page_content='Lorem ipsum dolor sit amet.', lookup_str='', metadata=
{'source': 'fake.docx', 'filename': 'fake.docx', 'category': 'Title'},
lookup_index=0)
```