Text file format

To Extract the Text file formats

1. PDF
2. DOC
3. HTML
4. XML
5. JSON
6. CSV

Pdf to text conversion python libraries

1.Approach

PYPDF

1. Extracts text from pdf
2. reader.pages[page_number]
3. reader.numPages
4. metadata
5. Rotating Pdf pages

1.writer.getNumPages()
2.addMetadata
3.writer.getPage(pageNumber)

1. Merging pdfs
2. Splitting the Pdfs

2. Approach

PYMUPDF

1. Document.page_count : the number of pages
2. Document.metadata : the metadata
3. Document.get_toc() : get the table of contents
4. Document.load_page() : read a page

It supports multiple file formats and computationally faster.

3.Approach

FITZ

1.doc.page_count
2.doc.metadata
3.doc.load_page

DOC TO TEXT CONVERSION:

1.Approach:

Spire

Extract a specific paragraph

1. # Get a specific section

Document.Sections[index]

1. # Get a specific paragraph

Section.Paragraphs[index]

1. Extract. a text from a entire Word document
2. # Get text from the entire document

str = doc.GetText()

2.Approach

Docx2txt

1. docx2txt.process
2. Docx2txt.tables

3.Approach

Docx2pdf

1. convert.path

HTML TO TEXT CONVERSION

1. Approach:

- html2text
- Beautifulsoup

2.Approach

pyhtml2pdf

XML TO TEXT CONVERSION

1.Approach

1. ElementTree
2. Beautifulsoup

2.Approach

pyxml2pdf

JSON TO TEXT CONVERSION

1. Approach
    - pandas
    - Json
2. Approach
    - Convert json to python
    - convert json to pdf
    - convert json to civ
3. Approach

Testsigma free online tool

CSV to text EXtraction

1. Approach
- Converting csv file to html file using pandas framework
- PDFkit Python API to convert our HTML file to the PDF file format.
2. Approach
    - Zamzar
    - Aspose