PDF Text Extractors

Tools for Extracting Text from PDFs

Danny S. Vargas
Viviane Moreira

08/05/2019

Tools

1. Tika
2. PDFBox
3. mupdf
4. pdfminer
5. pdftotext
6. poopler(popplerqt4)
7. textract
8. pypdf2xml

Tika

- Baseline
- Java

PDFBox

- Tika
- Java

muPDF

- PDF, XPS, and E-book viewer
- C

pdfminer

- Exact location of text in a page
- Python

pdftotext

- xpdf
- poppler

poppler(popplerqt4)

- PDF viewer
- xpdf3.0
- python

Textract

- Any type of file
- pdftotext and pdfminer
- python

pypdf2xml

- Python implementation of pdftoxml
- pdfminer(unicode characters)

Common Features

- PDFBox & Tika (Java)
- pdfminer, pdftotex & textract
- pdftotext & poppler(poppler-utils)
- pypdf2xml(pdftoxml in python)

Results

- Baseline(Tika)

PDFBox

- No spaces between pages
- Similar results (structure and content)

muPDF

- Additional break-lines
- Similar results(structure and content)

pdfMiner

- Results depend on the input
  - (+) Follow the correct order
  - (-) One character per line
- Delayed

pdftotext

- Similar performance to Tika
- It does not follow the same order

poppler(popplerqt4)

- It maintains the structure (position of the text in the pdf document)

textract

- pdftotext and pdfminer(methods)
- Results depend on the method used

Tika vs Others & pdf2xml

- Organized
- How to interpret and clean the tags?
- Used as reference of time performance

Summary

- Dependency on the input
- Correct order
- Configuration
- Documentation
- Time performance

Finally

- PDFBox
- Tika
- mupdf

Repository

- Github repository:
  `https://github.com/dannysv/extract_textpdf`.