CASE STUDY

PDF-XML Parsing- Publishing Domain

Client Objective

The purpose of this project is to extract data from PDF files, and further process and convert it into JSON with an option to apply page breaks tags in XML files.

SOLUTION

01 Page Mapping

We provided XML files where we had to map the page breaks from PDF and apply page breaks tags.

02 Text Processing

We used NLP to process text where N-grams is used to extract a sequence of words and search in XML.

03 Data Cleansing

Processed dates, page no. and contents from PDF & created word validator to format text and validate styles

04 Book Formation

Ability to generate multiple reports including both a pictorial export of the family tree into PDF & written reports exported into word.

05 Feature List

Enabling Version controls, Effective communication btw the authors and the stakeholder.

TECHNOLOGIES

- Python
- Pandas
- Pdfplumber
- Python
- Pandas
- Pdfplumber
- Json
- Textacy
- Docx

VALUE TO THE CLIENT

- Our solution provided books generated from our end that helped the client to save a huge number of pages from being printed in vain
- Each page of a book was mapped from its latest edition
- We also provided many other essentials like we created a separate file where we only provided the changes that were made from the previous edition to the latest edition. We also kept a record of what has been changed in all editions of a book