Extract Text From PDF

16th September 2020

OVERVIEW

The idea of the project is, Given a PDF file, the API is supposed to extract the text out of it and return the text as the response.

Packages Used

1. **Pdfminer3**
2. **Flask framework**

Live Project:

https://pdf-text-extracter.herokuapp.com/extractText

Note: The usage for this API is explained in the Readme.

Approach

1. Used the pdfminer3 package to extract text from the pdf page by page and store them in a list.
2. Various classes used during the task are explained inside the code.
3. **Images and other multimedia present in the PDF's are skipped** and only the text is considered. (See line 38 in the ExtractText.py file)
4. A simple flask server is set up and the file is accepted at an '**extractText**' route, which accepts a **POST request.**
5. The function to extract the pdf is run and the array is returned from the function which is in turn returned as the value to the key "**pageWiseExtractedText**"
6. If a user sends a GET Request or the key for the file sent is not proper, the route returns appropriate errors.
7. The API Structure and response structure are defined in the code as well as the readme for the project.

Raghav Maheshwari