PDFs-TextExtract

Python Multiple and Large PDF Documents Text Extraction - Python 3.7

Introduction

As a Data Scientist , You may not stick to data format.

PDFs is good source of data, most of the organization release their data in PDFs only. **As AI is growing, we need more data for prediction and classification**; hence, ignoring PDFs as data source for you could be a blunder.

As you know PDF Processing comes under text analytics.

Most of the Text Analytics Library or frameworks are designed in Python only, this gives a leverage on text analytics. You can never process a pdf directly in exising frameworks of Machine Learning or Natural Language Processing. Unless they are proving explicit interface for this, **we have to convert pdf to text first.**

Problematic

Most Python Liabiries for Pdf Processing such as PyPDF2 and Pdfminer.six perform in text extraction task, but this performance is limited to a small and simple PDF document.

That's why, **PDFs-TextExtract** project developed to **extract text from multiple and large pdf documents.**

Setup Environment

- **Step 1:** Select Version of Python (Python 3.7) to Install from Python.org website.
- **Step 2:** Download Python Executable Installer.
- **Step 3:** Run Executable Installer.
- **Step 4:** Verify Python Was Installed On Windows.
- **Step 5:** Verify Pip Was Installed.
- **Step 6:** Add Python Path to Environment Variables (Optional).
- **Step 7:** Install Python extension for your IDE (Visual Studio Code).

- **Step 8:** Now you'll be able to execute python scripts with your IDE (Visual Studio Code).
- **Step 9:** Execute *Terminal command* inside Python IDE : **pip install pdfminer.six**
- **Step 10:** Execute *Terminal command* inside Python IDE : **pip install PyPDF2**

Usage

- **Step 1:** Open **..\PDFs-TextExtract-master\samples** folder and put your PDF Documents inside.
- **Step 2:** Execute **..\PDFs-TextExtract-master\Scripts\merged.py** script.
- **Step 3:** Execute **..\PDFs-TextExtract-master\Scripts\spliter.py** script.
- **Step 4:** Execute **..\PDFs-TextExtract-master\Scripts\extract_text.py** script.
- **Step 5:** Open **..\PDFs-TextExtract-master\output** and you will find the result there.

Resources

- Overview about PDF Processing with Python
- **pdf2txt** tool forked from pdfminer.six project.
- **merger** and **spliter** tools forked from PyPDF2 project.