Critical Findings

PyPDF2 Library

PyPDF2 is a Python library that allows the manipulation of PDF documents. It can be used to create new PDF documents, modify existing ones and extract content from documents. PyPDF2 is a pure Python library.

- It can read, parse and write PDFs. It can be used as a command line tool or as a library.
- It's written in pure Python with no external dependencies (except for Python itself).
- It supports Unicode strings so that it can handle non-English characters.

Importance and Uses

As we know PDF is the most widely used document format, with over 73 million new PDF files saved every day on Gmail & Drive. It can be an eBooks, digitally signed agreements, password-protected documents, or scanned documents like passports.

PyPDF2 helps to work with PDF files and perform the following tasks:

- Extract text from PDF file
- Converting PDF files into images (png or jpeg) or text files
- Encrypt a PDF file.
- Rotate, merge and split PDF files
- Adding a watermark to a PDF file
- Editing existing PDFs by adding, removing, replacing, or modifying pages

Installation

There are several ways to install PyPDF2. The most common option is to use pip.

```
pip install PyPDF2
```

Real time example:

If we want to extract data from PDF file to convert it into any other format, we simply use PyPDF2 Library instead of typing lengthy codes. Here is an easy simple code for that.

```
from PyPDF2 import PdfReader
```

```
reader = PdfReader("example.pdf")
page = reader.pages[0]
print(page.extract_text())
```

Hurdles:

But due to some reasons, it is hard to extract data from PDF files. PDF documents can contain images and text. PDF files don't store text in a semantically meaningful way, but in a way that makes it easy to show the text on screen or print it. For this reason text extraction from PDFs is hard.

AWS TEXTRACT

Amazon Textract is a Machine Learning (ML) service that automatically extracts text, handwriting, and data from scanned documents. Amazon Textract makes it easy to text detection and analysis to applications. Amazon Textract provides synchronous operations for processing small, single-page, documents and with near real-time responses. It analyzes the input using the power of **NLP algorithms** to extract key phrases, entities, and sentiments automatically.

In this artificially driven digital world, Artificial Intelligence (AI) is playing an impressive role in Natural Language Processing (NLP). Under the umbrella of Machine Learning, all leading communication languages are vigorously nurturing this data driven world.

Main Modules

Form extraction

We can detect key-value pairs in document images automatically and retain the context without manual intervention.

Table extraction

Amazon Textract preserves the composition of data stored in tables during extraction. This is helpful for documents that are largely composed of structured data, such as financial reports or medical records with tables in columns and rows.

Handwriting recognition

Documents, such as medical intake forms and employment applications, include both handwritten and printed text.

Invoices and receipts

Amazon Textract uses machine learning (ML) to understand the context of invoices and receipts and automatically extracts relevant data such as vendor name, invoice number, item prices, total amount, and payment terms.