

Engine Works


Under the hood of Alteryx: tips, tricks and how-tos.

Community > Learn > Blogs > Engine Works > PDF Spatial Parse with the Python Tool

Share:   

PDF Spatial Parse with the Python Tool



 gawa 16 - Nebula

...01-22-2025 08:43 AM

Why is Parsing PDFs Difficult?

Parsing unstructured data like PDFs is generally a very challenging task. With structured data, data is easily accessible by specifying column names and row numbers, much like a grid on graph paper. In contrast, unstructured data is like a blank sketchbook, where identifying data relies heavily on geometric positioning.

How to Read PDFs in Alteryx Designer

When you need to import a PDF into Alteryx Designer, the first option you might consider is using the Intelligence Suite add-on tools. [This article](#) introduces examples of how to use these tools. Alternatively, it's also doable to read PDF file using the Python tool. In this blog, I will introduce a PDF parsing method that uses the Python tool and the `pdfminer.six` library.

Reading Data from PDFs with the Python Tool Using pdfminer.six

While I don't go into a detailed explanation of the `pdfminer.six` library as it is not the main topic of this article, the key point is that this library enables you to extract not only text data from a PDF but also their coordinate information. Please note that `pdfminer.six` does not support OCR functionality, so PDF text shall be readable in advance.

For example, let's try reading a PDF of a typical resume like this:



Name	Jason Lee		
Date of Birth	July 8, 1988	Gender	Male
Address	88 Brookside Avenue, Austin, TX 78704		
Phone Number	(512) 555-4910		

Job History

Year	Description
2016-2020	Software Developer TechNova Inc.
2020–Present	Senior Software Engineer CloudBridge Technologies

Resume

By using the `pdfminer.six` library within the Python tool, some attribute of each text element are extracted as shown in the figure below. The text in the PDF is recognized as rectangular objects, and the (X, Y) coordinates of the four vertices of each rectangle are captured as **X0, X1, Y0, Y1**. The **Angle** represents the text orientation: **0°** indicates horizontal text, while **90°/-90°** indicates vertical text.

Coordinate information is utmost important when parsing PDF files. For example, text elements with the same **X0 (or Y0)** value can be deemed vertically (or horizontally) aligned, which may help in reconstructing tables.

Record	X0	X1	Y0	Y1	Text	Angle	Pages
1	0.09072	0.13862	0.715266	0.729306	Name	0	0
2	0.21804	0.345768	0.711492	0.731532	Emily Carter	0	0
3	0.09072	0.190549	0.658026	0.672066	Date of Birth	0	0
4	0.21804	0.302544	0.669313	0.679873	March 15, 1992	0	0
5	0.3738	0.417488	0.669313	0.679873	Gender	0	0
6	0.4446	0.484901	0.669313	0.679873	Female	0	0
7	0.09072	0.154912	0.600786	0.614826	Address	0	0
8	0.21804	0.438483	0.612073	0.622633	742 Maplewood Lane, Portland, OR 97205	0	0
9	0.1242	0.156502	0.433212	0.445212	Year	0	0
10	0.31464	0.390776	0.433212	0.445212	Description	0	0
11	0.09072	0.151055	0.406753	0.417313	2019-2022	0	0
12	0.196792	0.428595	0.406753	0.417313	Marketing Coordinator GreenLeaf Solutions	0	0
13	0.09072	0.165312	0.388273	0.398833	2022–Present	0	0
14	0.196793	0.405005	0.388273	0.398833	Marketing Manager BrightWave Digital	0	0
15	0.09072	0.20629	0.543666	0.557706	Phone Number	0	0
16	0.21804	0.304354	0.554953	0.565513	(503) 555-8372	0	0
17	0.09306	0.186003	0.467184	0.483145	Job History	0	0
18	0.08508	0.093156	0.467184	0.508273	J	0	0
19	0.09072	0.13862	0.715266	0.729306	Name	0	1

Creating Spatial Objects from Coordinate Data

Here is a more advanced example. Although we were able to extract text data as described earlier, it's still difficult to determine which records are the target data we need. To address this, spatial analysis can be utilized. This approach is similar to the concept behind the **Image Template Tool**. In this method, you need to pre-mark the areas containing the data you want to extract on a template PDF. Then, by matching this template with the target PDF, you can extract the target data.

Creating the Template

First, create a template by marking the areas that contain the data you want to extract. Assuming that the same format repeats across multiple pages in the same PDF file, let's use the first page as the template:



- Community | Alteryx IO | MyAlteryx | [Alteryx](#)
1. Place annotation text boxes to cover the areas where you want to extract data on the first page.
 2. Enter the data labels ('Name', 'DOB', 'Gender', etc.) inside each text box

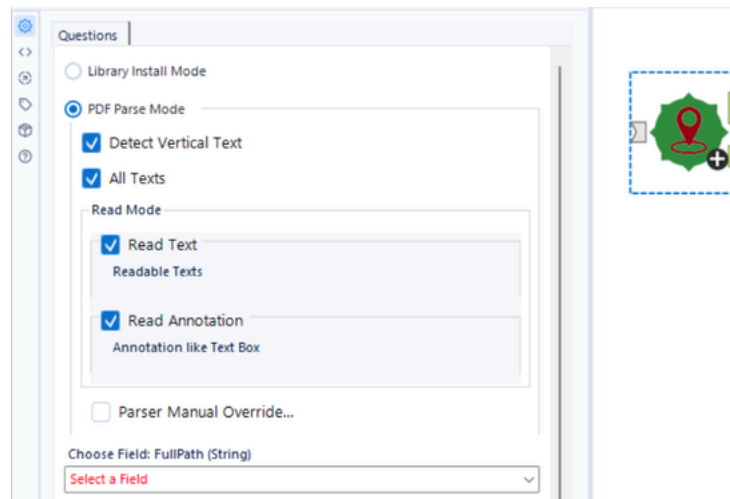
Name	Emily Carter		
Date of Birth	March 15, 1992	Gender	Female
Address	742 Maplewood Lane, Portland, OR 97205		
Phone Number	(503) 555-8371		

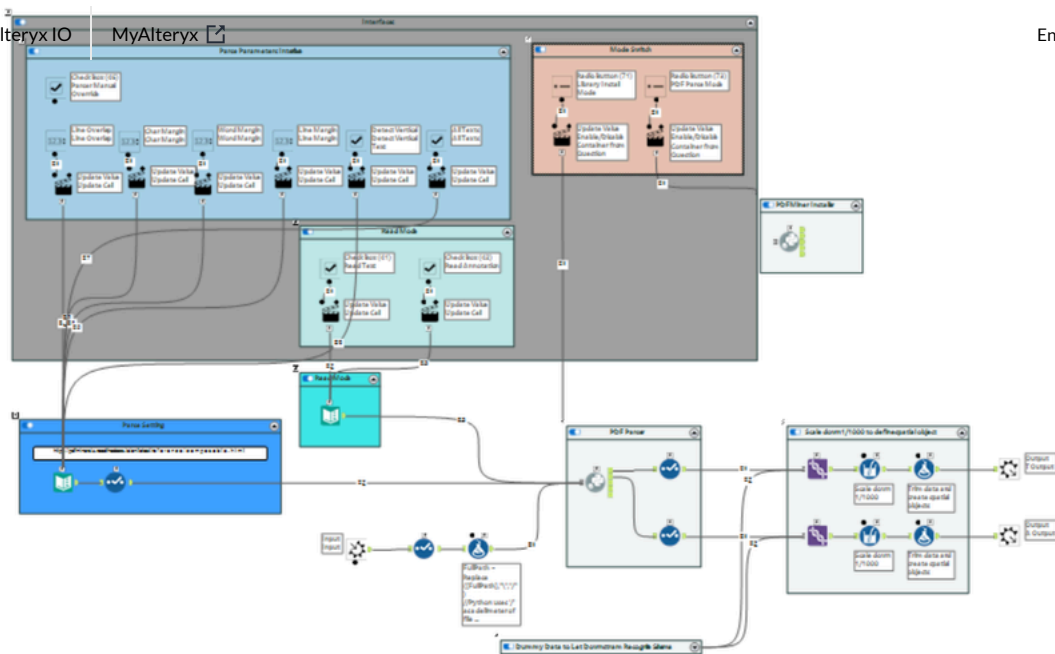
Job History

Year	Description
2019-2022	Marketing Coordinator GreenLeaf Solutions
2022-Present	Marketing Manager BrightWave Digital

Text and Annotation Objects in PDFs

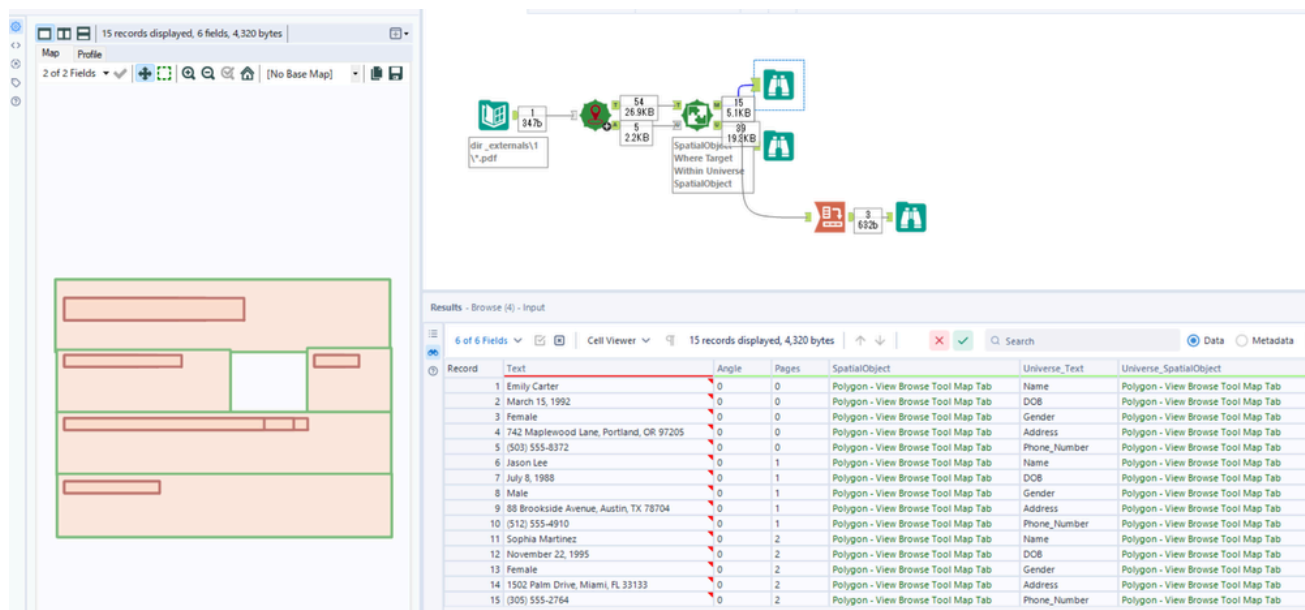
In a PDF, the text objects and the annotation objects (like the ones added in the previous step) are recognized as different types of objects. Here is the macro to automate the process of extracting these objects separately and retrieving their text and coordinate information.





- **T Anchor:** Text objects → text data and rectangular spatial objects in the all pages
- **A Anchor:** Annotation objects → annotation text and rectangular spatial objects in the template

Next, the spatial objects from the **T Anchor** and **A Anchor** are matched using **Spatial Match tool**. This enables you to extract the match text that exists within the area of the annotation objects on the template. See the configuration window to visually know how the target texts are captured.



For example, grouping by page and using the Crosstab tool can transform the data into a structured format like the table below.

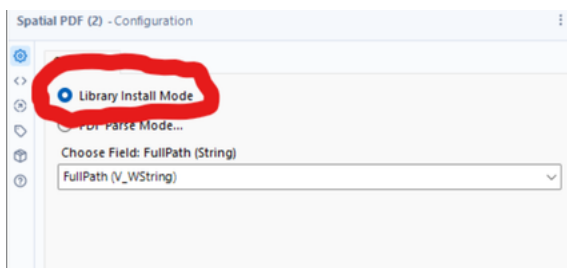
Record	Pages	Address	DOB	Gender	Name	Phone_Number
1	0	742 Maplewood Lane, Portland, OR 97205	March 15, 1992	Female	Emily Carter	(503) 555-8372
2	1	88 Brookside Avenue, Austin, TX 78704	July 8, 1988	Male	Jason Lee	(512) 555-4910
3	2	1502 Palm Drive, Miami, FL 33133	November 22, 1995	Female	Sophia Martinez	(305) 555-2764



Notes on Using the Spatial PDF Macro

steps:

1. **Launch Alteryx Designer with Administrator Privileges.**
 - Right-click the Alteryx icon → Select **Run as administrator**
2. **Extract the .yxzp and open the workflow (WF).**
 - In the **Spatial PDF** configuration screen, select '**Library Install Mode**', then run the workflow.



If no errors occur after execution, the library installation was successful.

If an error appears, check the following:

- Is the Python library installation command (**pip**) being blocked? Office networks may block it via proxy settings, so consult your IT administrator if this is the case.
- Be sure again that Alteryx Designer is running with administrator privileges.

Conclusion

In this blog, the method for parsing PDF data using Python tools and spatial tools is introduced. This approach is effective when extracting data that consistently repeats in the same location across all pages in a PDF file. Additionally, even without using spatial objects, coordinate information could be applied in various ways, such as reconstructing tables.

As for PDF parsing, it does not have a universal 'best' solution. I encourage you to explore a wide range of options, including the features of the [Intelligence Suite](#), to determine what could be the best practice for your needs.

Notes on the Workflow Shared in This Blog:

- The WF shared in this blog is aiming for technical demonstration purposes and its functionality is not guaranteed. Please use and modify it at your own risk.
- Unable to provide technical support for errors/issues, and not intended to update it even if any feature request is received.

[SampleWF_PDF_Spatial_Parse.yxzp](#)

Alteryx Designer Spatial Analysis

Subscribe to blog



Comments



ACE BS_THE_ANALYST
14 - Magnetar

01-22-2025 10:20 AM

...



Oh, very cool @gawa! What an interesting way to think about parsing pdfs. This was great read 🍌.

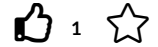


ACE clmc9601
13 - Pulsar

01-22-2025 11:31 AM

...



This is awesome! Thanks, [@gawa!](#)



You must be a registered user to add a comment. If you've already registered, sign in. Otherwise, register and sign in.

[Comment](#)

Recommendations

 [Upgrading from Python Engine SDK to Python SDK V2](#) 

 [Python tool: Installing New Packages correct steps](#) 

 [Python Tool notebook not loading](#)  

 [PDF mining tool](#) 

 [FileCopyMove Python Versioning](#) 

[Back to Blog](#) [< Newer Article](#) [Older Article >](#)



© 2025 Alteryx, Inc.

[Privacy Policy](#)

[Cookie Settings](#)

[Cookie Notice](#)

[Terms and Conditions](#)

[Version History](#)

