PDF Input

Description

The PDF Input tool is designed to allow users to input the content of PDFs into Alteryx for further analysis.

Prerequisites

The macro leverages the R tool and as such users must ensure they have this installed using the separate ‘R installer’.

The script that runs to read in the PDFs uses the package ‘pdftools’; this package is not installed by the ‘R Installer’ and we recommend the following steps for installing the package.

1. Go to C:\Program Files\Alteryx\R-3.4.4\bin (the version may differ)
2. Right click on the ‘R.exe’ file and select ‘Run as administrator’
3. Enter the following command

   ```
   install.packages(“pdftools”)
   ```

4. You will be prompted to select a ‘cran mirror’, select any (though note some cran mirrors will not be able to provide the ‘pdftools’ package and you may get an error message).

5. Upon completion you should receive a message in the command prompt saying ‘package pdftools successfully unpacked and MD5 sums checked’. If you do not receive this message use stackoverflow or the Alteryx community to begin debugging.

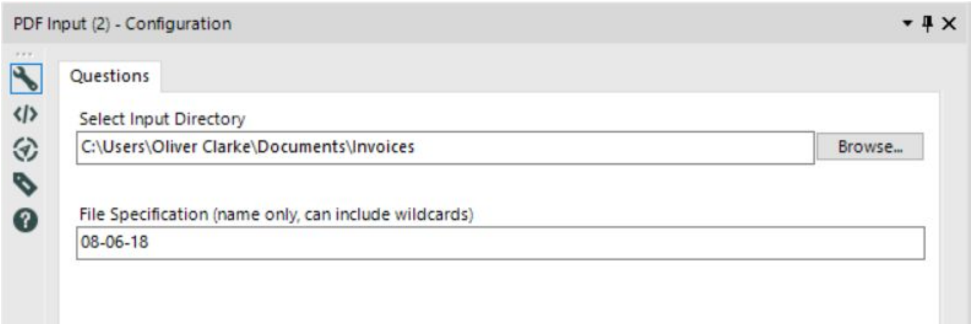
Alternatively there is an installer application available here in the Alteryx Gallery which can be used to install additional packages.

Inputs

This macro requires no input data stream.

Configuration

The configuration pane consists of a single tab, allowing users to specify the directory containing their PDF files. Users can also specify some conditions regarding the file names, built in in the same manner as the wildcard input available with the standard input tool.



Select Input Directory
C:\Users\Oliver Clarke\Documents\Invoices

File Specification (name only, can include wildcards)
08-06-18

Select Input Directory

Users should either type in, or use the browse folder functionality to enter the full path of the folder containing their PDFs.

For example “*C:\Users\Oliver Clarke\Documents\Invoices*”

File Specification

Users can choose specific files, or utilise a wildcard input to select all or a subset of PDFs in their chosen folder.

Specific file example: “*06-08-18*”
Subset example: “*-18*”
All example: “*”

Outputs

The output of this macro contains two fields. The first field contains the data from the PDF, this is already parsed by line break in the original file. The second field contains the file from which the data comes.

Given that PDFs come in many different forms, it is impossible to create a method of automatically parsing a file into a tabular format. Users will be required to create their own parsing method, though we have provided an example of the methods that can be used to perform such a transformation.

Data	File
Invoice Number - 00100	*C:\Users\Oliver Clarke\Documents\Invoices*\06-08-18.pdf
Amount - £2.50	*C:\Users\Oliver Clarke\Documents\Invoices*\06-08-18.pdf
Invoice Number - 00101	*C:\Users\Oliver Clarke\Documents\Invoices*\07-08-18.pdf
Amount - £4.20	*C:\Users\Oliver Clarke\Documents\Invoices*\07-08-18.pdf

Example Workflow

Users can find an example workflow containing the 'PDF Input' macro here. This macro also showcases example parsing techniques that can be applied to bring the resulting single cell string into a data table.

Known Issues

None.

References

This macro, and especially the R code within it, is based on this blogpost by Oliver Power.