

ADAMS

Advanced **D**ata mining **A**nd **M**achine learning **S**ystem

Module: adams-pdf



Peter Reutemann

June 23, 2015

©2012-2015



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato



Except where otherwise noted, this work is licensed under
<http://creativecommons.org/licenses/by-sa/3.0/>

Contents

| | | |
|----------|------------------------|-----------|
| 1 | Introduction | 7 |
| 2 | Flow | 9 |
| 3 | Tools | 13 |
| 4 | Troubleshooting | 15 |
| | Bibliography | 17 |

List of Figures

| | | |
|-----|--|----|
| 2.1 | Flow for creating a PDF file from various sources. | 10 |
| 2.2 | CSV files get added as tables. | 10 |
| 2.3 | Images can get inserted as well. | 11 |
| 2.4 | Plain text files get added as simple text. | 11 |
| 3.1 | Viewer for PDF files. | 13 |

Chapter 1

Introduction

The *pdf* module adds PDF authoring capabilities to ADAMS. This is possible thanks to the iText [2] and jPod Renderer [3] libraries for creating, manipulating and viewing of PDF files.

Chapter 2

Flow

For manipulating and viewing PDF files, you can use the following actors:

- *PDFCreate* – for creating PDFs, using text files, spreadsheets and images.¹
- *PDFExtract* – extracts a range of pages.
- *PDFExtractImages* – extracts the images from a PDF.²
- *PDFExtractText* – obtaining the plain text of the PDF.³
- *PDFMerge* – merges several PDFs into a single one.⁴
- *PDFPageCount* – determines the number of pages in a PDF.⁵
- *PDFStamp* – allows to add an overlay to a PDF.⁶
- *PDFViewer* – for viewing PDFs.⁷

Creating PDFs

Generating a PDF using the *PDFCreate* transformer is really easy. The transformer takes an array of file names as input, which will all get added to the specified output PDF file. Basic options, like page size and orientation, can be set as well. How and what files get added to the PDF, is determined by the “proclets”, i.e., little processor classes, that you specify and configure:

- *CsvPdfProclet* – for adding CSV files as tables.
- *HeadlinePdfProclet* – for adding a headline.
- *ImagePdfProclet* – for adding images (GIF, JPEG, PNG).
- *PlainTextPdfProclet* – for adding plain text files as paragraphs.

Figure 2.1 shows a flow⁸ that adds all files (i.e., three) found in a directory to a single PDF. Figures 2.2, 2.3 and 2.4 show the resulting pages of the PDF in the viewer.

¹adams-pdf-create_pdf.flow

²adams-pdf-extract_images.flow

³adams-pdf-extract_text.flow

⁴adams-pdf-page_count.flow

⁵adams-pdf-page_count.flow

⁶adams-pdf-page_overlay.flow

⁷adams-pdf-view_pdf.flow

⁸adams-pdf-create_pdf.flow

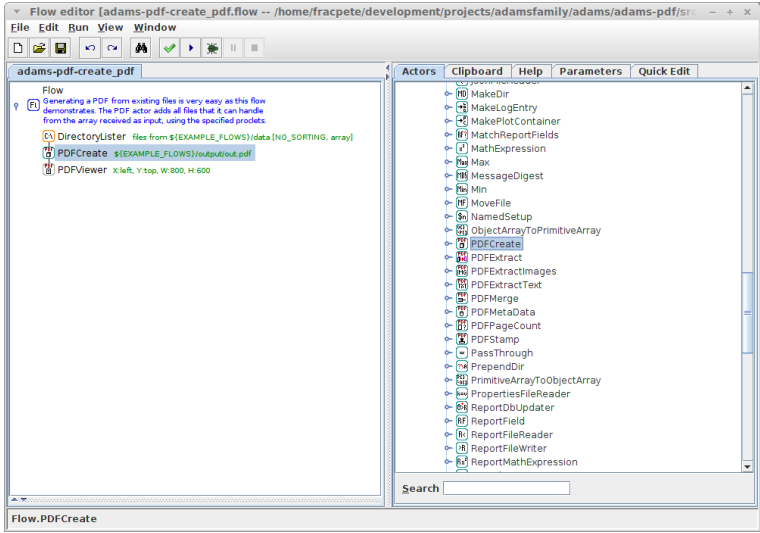


Figure 2.1: Flow for creating a PDF file from various sources.

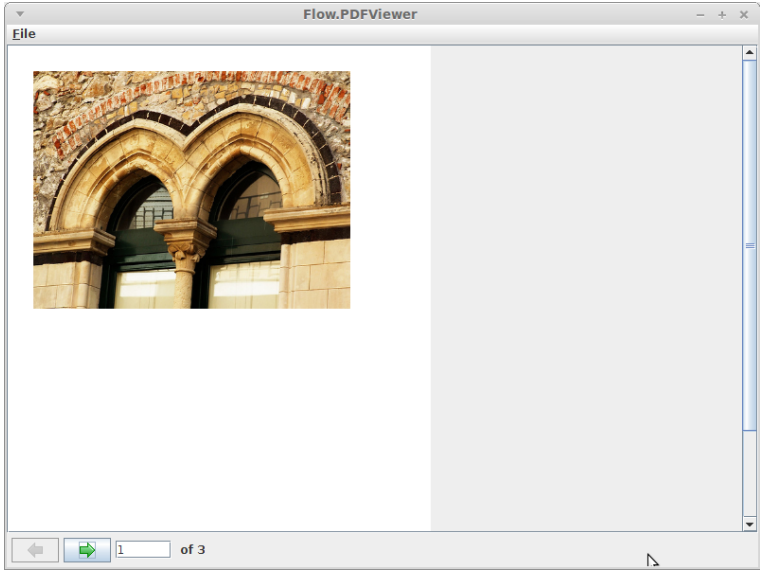
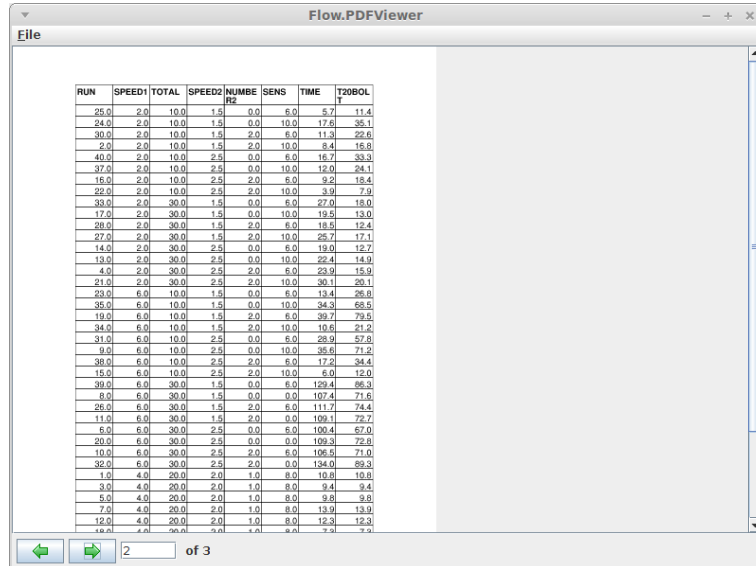


Figure 2.2: CSV files get added as tables.



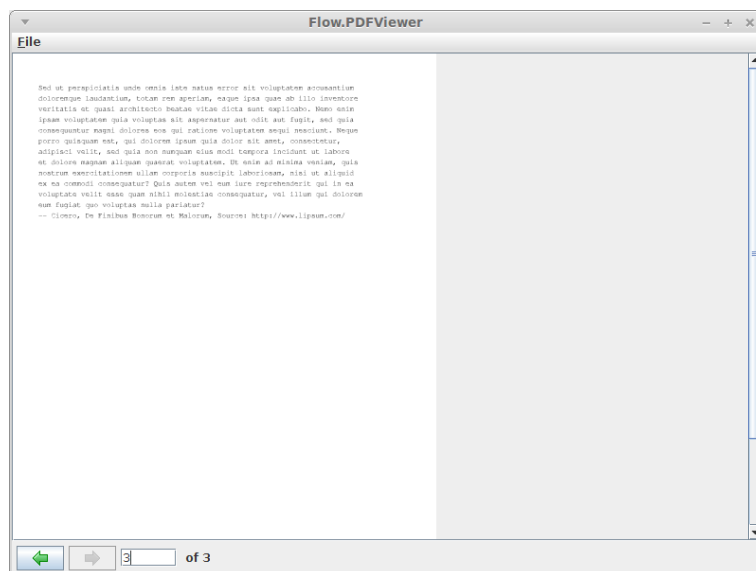
Flow.PDFViewer

File

| RUN | SPEED1 | TOTAL | SPEED2 | NUMBER | SENS | TIME | TZOBOL |
|------|--------|-------|--------|--------|------|-------|--------|
| | | | RS | | | | T |
| 25.0 | 2.0 | 10.0 | 1.5 | 0.0 | 5.0 | 5.7 | 11.4 |
| 24.0 | 2.0 | 10.0 | 1.5 | 0.0 | 10.0 | 17.6 | 25.1 |
| 30.0 | 2.0 | 10.0 | 1.5 | 2.0 | 5.0 | 11.3 | 22.0 |
| 2.0 | 2.0 | 10.0 | 1.5 | 2.0 | 10.0 | 8.4 | 16.8 |
| 40.0 | 2.0 | 10.0 | 2.5 | 0.0 | 5.0 | 16.7 | 33.3 |
| 37.0 | 2.0 | 10.0 | 2.5 | 0.0 | 10.0 | 12.0 | 24.1 |
| 16.0 | 2.0 | 10.0 | 2.5 | 2.0 | 5.0 | 9.2 | 18.4 |
| 22.0 | 2.0 | 10.0 | 2.5 | 2.0 | 10.0 | 3.9 | 7.9 |
| 33.0 | 2.0 | 30.0 | 1.5 | 0.0 | 5.0 | 27.0 | 15.0 |
| 17.0 | 2.0 | 30.0 | 1.5 | 0.0 | 10.0 | 19.5 | 13.0 |
| 28.0 | 2.0 | 30.0 | 1.5 | 2.0 | 5.0 | 18.5 | 12.4 |
| 27.0 | 2.0 | 30.0 | 1.5 | 2.0 | 10.0 | 25.7 | 17.1 |
| 14.0 | 2.0 | 30.0 | 2.5 | 0.0 | 5.0 | 19.0 | 12.7 |
| 13.0 | 2.0 | 30.0 | 2.5 | 0.0 | 10.0 | 22.4 | 14.9 |
| 4.0 | 2.0 | 30.0 | 2.5 | 2.0 | 5.0 | 23.9 | 15.9 |
| 21.0 | 2.0 | 30.0 | 2.5 | 2.0 | 10.0 | 30.1 | 25.1 |
| 23.0 | 6.0 | 10.0 | 1.5 | 0.0 | 5.0 | 13.4 | 26.8 |
| 35.0 | 6.0 | 10.0 | 1.5 | 0.0 | 10.0 | 34.3 | 66.5 |
| 19.0 | 6.0 | 10.0 | 1.5 | 2.0 | 5.0 | 29.7 | 73.5 |
| 34.0 | 6.0 | 10.0 | 1.5 | 2.0 | 10.0 | 10.6 | 21.2 |
| 31.0 | 6.0 | 10.0 | 2.5 | 0.0 | 5.0 | 28.9 | 57.8 |
| 9.0 | 6.0 | 10.0 | 2.5 | 0.0 | 10.0 | 35.8 | 71.2 |
| 38.0 | 6.0 | 10.0 | 2.5 | 2.0 | 5.0 | 17.2 | 34.4 |
| 15.0 | 6.0 | 10.0 | 2.5 | 2.0 | 10.0 | 6.0 | 12.0 |
| 39.0 | 6.0 | 30.0 | 1.5 | 0.0 | 5.0 | 129.4 | 86.3 |
| 8.0 | 6.0 | 30.0 | 1.5 | 0.0 | 0.0 | 107.4 | 71.6 |
| 26.0 | 6.0 | 30.0 | 1.5 | 2.0 | 5.0 | 111.7 | 74.4 |
| 11.0 | 6.0 | 30.0 | 1.5 | 2.0 | 0.0 | 109.1 | 72.7 |
| 5.0 | 6.0 | 30.0 | 2.5 | 0.0 | 5.0 | 100.4 | 67.0 |
| 30.0 | 6.0 | 30.0 | 2.5 | 0.0 | 0.0 | 106.3 | 72.8 |
| 10.0 | 6.0 | 30.0 | 2.5 | 2.0 | 5.0 | 106.5 | 71.0 |
| 32.0 | 6.0 | 30.0 | 2.5 | 2.0 | 0.0 | 134.0 | 89.3 |
| 1.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 19.8 | 10.8 |
| 3.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 9.4 | 9.4 |
| 5.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 9.8 | 9.8 |
| 7.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 13.9 | 13.9 |
| 12.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 12.3 | 12.3 |
| 16.0 | 4.0 | 20.0 | 2.0 | 1.0 | 5.0 | 3.3 | 3.3 |

2 of 3

Figure 2.3: Images can get inserted as well.



Flow.PDFViewer

File

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium
 doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore
 veritatis et quasi architecto nuncupata vitae sunt explicatione. Nam esse
 ipsam voluptatem quia voluptas est aspernatur aut odit aut fugit, et quia
 consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque
 porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur,
 adipisci velit, et quia non numquam eius modi tempora incidunt ut labore
 et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis
 nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid
 ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea
 voluptate velit esse quam nihil molestias consequatur, vel illum qui dolorem
 eum fugiat quo voluptas nulla pariatur?
 -- Lorem, De Finibus Bonorum et Malorum, Source: <http://www.lipsum.com/>

3 of 3

Figure 2.4: Plain text files get added as simple text.

Chapter 3

Tools

The *PDF viewer* can be used to load and browse PDF files. Figure 3.1 shows a screenshot of the viewer with a PDF file loaded.

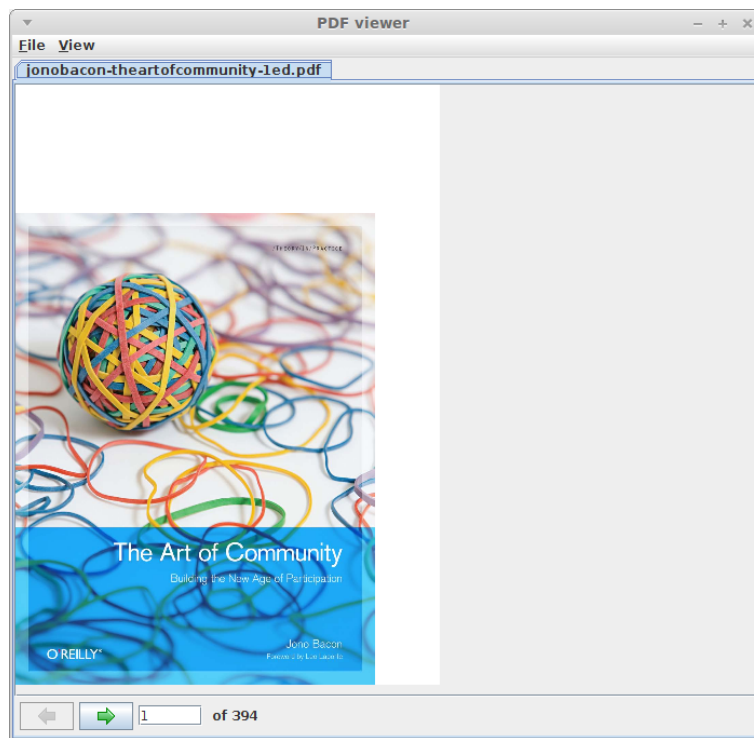


Figure 3.1: Viewer for PDF files.

Chapter 4

Troubleshooting

- **Problem:** 64bit Linux shows error message when viewing PDF that 'libfreetype.so' library was not found ("UnsatisfiedLinkError").

Solution: Make sure that /usr/lib/libfreetype.so exists. If not, add a symbolic link to the 64bit library, using a similar command as follows (for Kubuntu 11.10 or LinuxMint 11, 13):

```
sudo ln -s /usr/lib/x86_64-linux-gnu/libfreetype.so.6 /usr/lib/libfreetype.so
```

And restart the application.

Bibliography

- [1] *ADAMS* – Advanced Data mining and Machine learning System
<https://adams.cms.waikato.ac.nz/>
- [2] *iText* – A library that allows you to create and manipulate PDF documents
<http://itextpdf.com/>
- [3] *jPod Renderer* – A PDF rendering implementation for AWT and SWT.
<http://opensource.intarsys.de/home/en/index.php?n=JPodRenderer.HomePage>