WRANGLING DATA OUT OF PDF FILES

Years ago, it was how they got you. Some pesky government agency would comply with your request for data by sending you a PDF file. Then you’d have to negotiate/cajole/badger them incessantly for the underlying data.

Those days, thankfully, are mostly over. Not only are more agencies opening up their public data sets, but the number of tools available for reporters to deal with PDF files is as good as ever.

This tutorial will walk you through a bunch of possible solutions. Why a bunch? Because experience shows that one tool doesn’t handle every situation you’ll encounter, and having a working knowledge of several options will come in handy.

I will divide this discussion into three categories of PDF tools:

**Free software**: Programs that cost $0 but generally have the most limited feature sets.

**Commercial software:** Programs that have powerful additional features that you might need to get a project done.

**Programming techniques:** I will show quick demonstrations of how you can open and manipulate PDF files in Ruby, Python and R, three programming languages that are increasingly common in newsrooms.

But before I get to all of the cool solutions, let’s first stop and talk about a few basics. The solution you choose will depend a lot on what you want to accomplish and what shape your source PDF files are in.

For example, sometimes there’s a table embedded in a document that you want to extract into a spreadsheet. If you try cutting and pasting from your PDF reader, you lose all of the columns and rows and are left with an un-analyzable mess.

Another document might also have the table, but when you try to copy and paste it, you get a discouraging message announcing that the PDF you’re reading isn’t text, but an image of text. This means you’ll have an extra step of applying “Optimal Character Recognition” to the document (otherwise known as OCRing) before you can pull out the data.

Sometimes the table you’re going after is not in a simple rows-and-columns format. There might be tables nested inside tables. Or columns filled with text that wraps over more than one line.

These are three different problems with three different solutions. So as I go through your options, I will make note of what situation each solution is best suited for.

tabula

able2extract

cogniview

comet

acrobat

ruby

python

R

document cloud

monarch

abby

Levine at the terminal