Extract data from a PDF

This notebook demonstrates how to extract data from a PDF using the pdfplumber library, which was developed primarily by Jeremy Singer-Vine, currently at BuzzFeed News.

⚠ Note: This library works only on native PDFs (e.g. the text is selectable), not on scanned images. If you have a scanned image saved as a PDF, you'd first need to extract the text using some kind of Optical Character Recognition software (I'd suggest using DocumentCloud, which OCRs image PDFs on import, or a tool like OCRMyPDF.)

This example extracts the data tables from a PDF that began life as a spreadsheet -- a list of WARN notices in California -- but please note that this library is powerful enough to target specific pieces of data from pages using a number of highly customizable extraction strategies.

In [ ]:
```
# import Python's csv module to write the data to file as delimited text
import csv

# import the library AS pp
import pdfplumber as pp
```

In [ ]:
```
# open the file using pdfplumber's open() method AS a variable called pdf
with pp.open('../data/warnreport.pdf') as pdf:
    # loop over list of pages
    for page in pdf.pages:

        # extract the first table it finds on each page
        table = page.extract_table()

        # loop over the lines in each table
        for line in table:
            # and print each line to see what's up
            print(line)
```

Save to file

Now that we've freed the data, we could do any number of things with it. Typically the best move is to save the data to a delimited text file, then decide what to do with it from there.

So to save the data as a CSV, we can use the `csv` module, which is part of Python's standard library.

Putting it all together (normally this would all happen in one block of code, but I've broken it out here to demonstrate the iterative process of developing a script):

In [ ]:
```
# open the file using pdfplumber's open() method AS a variable called pdf
# also, use Python's built-in open() method to create a new file to write to
with pp.open('../data/warnreport.pdf') as pdf, open('ca-warn-data.csv', 'w') as outfile:

    # create a CSV writer object and attach it to the file just opened
    writer = csv.writer(outfile)

    # loop over list of pages
    for page in pdf.pages:
```

```
        # extract the first table it finds on each page
        table = page.extract_table()

        # ... and write to file
        writer.writerows(table)
```