Project 3: PDF scraping in Python + REGEX

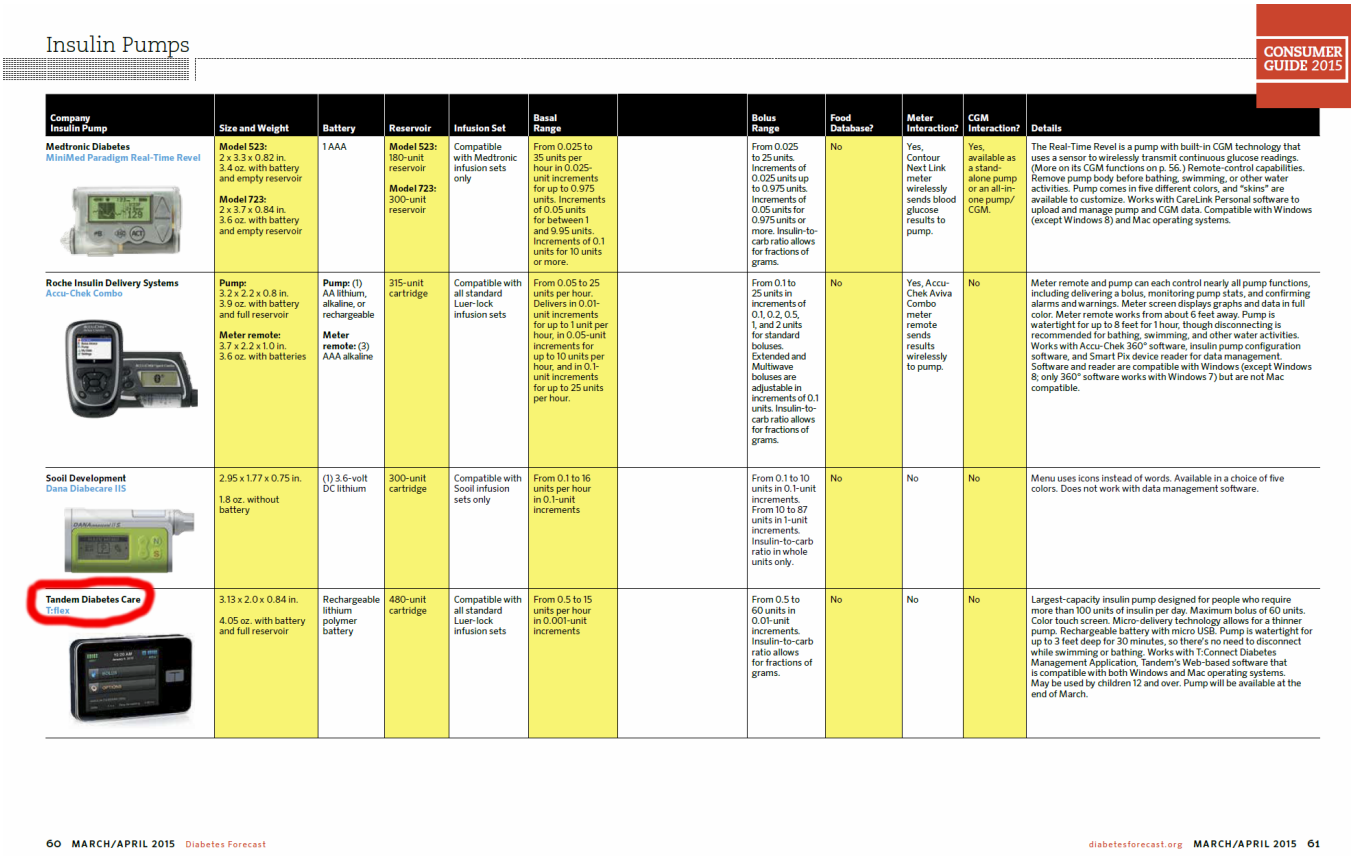
In this project we use regex to extract a list of items from a pdf file.

WE'LL COVER THE FOLLOWING

- PDF scraping example:
- Input file
- Solution

PDF scraping example:

In this project we will use a pdf file (see the screenshot below) from the diabetes.org website. Our goal is to list all the equipment models developed by the manufacturers names containing the word `tandem` (case insensitive).



Insulin Pumps

CONSUMER GUIDE 2015

Company Insulin Pump	Size and Weight	Battery	Reservoir	Infusion Set	Basal Range		Bolus Range	Food Database?	Meter Interaction?	CGM Interaction?	Details
Medtronic Diabetes MiniMed Paradigm Real-Time Revel	**Model 523:** 2 x 3.3 x 0.82 in. 3.4 oz. with battery and empty reservoir **Model 723:** 2 x 3.7 x 0.84 in. 3.6 oz. with battery and empty reservoir	1 AAA	**Model 523:** 180-unit reservoir **Model 723:** 300-unit reservoir	Compatible with Medtronic infusion sets only	From 0.025 to 35 units per hour in 0.025-unit increments for up to 0.975 units. Increments of 0.05 units for between 1 and 9.95 units. Increments of 0.1 units for 10 units or more.		From 0.025 to 25 units. Increments of 0.025 units up to 0.975 units. Increments of 0.05 units for 0.975 units or more. Insulin-to-carb ratio allows for fractions of grams.	No	Yes, Contour Next Link meter wirelessly sends blood glucose results to pump.	Yes, available as a stand-alone pump or an all-in-one pump/CGM.	The Real-Time Revel is a pump with built-in CGM technology that uses a sensor to wirelessly transmit continuous glucose readings. (More on its CGM functions on p. 56.) Remote-control capabilities. Remove pump body before bathing, swimming, or other water activities. Pump comes in five different colors, and "skins" are available to customize. Works with CareLink Personal software to upload and manage pump and CGM data. Compatible with Windows (except Windows 8) and Mac operating systems.
Roche Insulin Delivery Systems Accu-Chek Combo	**Pump:** 3.2 x 2.2 x 0.8 in. 3.9 oz. with battery and full reservoir **Meter remote:** 3.7 x 2.2 x 1.0 in. 3.6 oz. with batteries	**Pump:** (1) AA lithium, alkaline, or rechargeable **Meter remote:** (3) AAA alkaline	315-unit cartridge	Compatible with all standard Luer-lock infusion sets	From 0.05 to 25 units per hour. Delivers in 0.01-unit increments for up to 1 unit per hour, in 0.05-unit increments for up to 10 units per hour, and in 0.1-unit increments for up to 25 units per hour.		From 0.1 to 25 units in increments of 0.1, 0.2, 0.5, 1, and 2 units for standard boluses. Extended and Multiwave boluses are adjustable in increments of 0.1 units. Insulin-to-carb ratio allows for fractions of grams.	No	Yes, Accu-Chek Aviva Combo meter remote sends results wirelessly to pump.	No	Meter remote and pump can each control nearly all pump functions, including delivering a bolus, monitoring pump stats, and confirming alarms and warnings. Meter screen displays graphs and data in full color. Meter remote works from about 6 feet away. Pump is watertight for up to 8 feet for 1 hour, though disconnecting is recommended for bathing, swimming, and other water activities. Works with Accu-Chek 360° software, insulin pump configuration software, and Smart Pix device reader for data management. Software and reader are compatible with Windows (except Windows 8; only 360° software works with Windows 7) but are not Mac compatible.
Sooil Development Dana Diabecare IIS	2.95 x 1.77 x 0.75 in. 1.8 oz. without battery	(1) 3.6-volt DC lithium	300-unit cartridge	Compatible with Sooil infusion sets only	From 0.1 to 16 units per hour in 0.1-unit increments		From 0.1 to 10 units in 0.1-unit increments. From 10 to 87 units in 1-unit increments. Insulin-to-carb ratio in whole units only.	No	No	No	Menu uses icons instead of words. Available in a choice of five colors. Does not work with data management software.
Tandem Diabetes Care T:flex	3.13 x 2.0 x 0.84 in. 4.05 oz. with battery and full reservoir	Rechargeable lithium polymer battery	480-unit cartridge	Compatible with all standard Luer-lock infusion sets	From 0.5 to 15 units per hour in 0.001-unit increments		From 0.5 to 60 units in 0.01-unit increments. Insulin-to-carb ratio allows for fractions of grams.	No	No	No	Largest-capacity insulin pump designed for people who require more than 100 units of insulin per day. Maximum bolus of 60 units. Color touch screen. Micro-delivery technology allows for a thinner pump. Rechargeable battery with micro USB. Pump is watertight for up to 3 feet deep for 30 minutes, so there's no need to disconnect while swimming or bathing. Works with T:Connect Diabetes Management Application, Tandem's Web-based software that is compatible with both Windows and Mac operating systems. May be used by children 12 and over. Pump will be available at the end of March.

60 MARCH/APRIL 2015 Diabetes Forecast

diabetesforecast.org MARCH/APRIL 2015 61

Find all the product models by the manufacturer called `Tandem`

Input file

You can download the input file from here: data.pdf.

Solution

A complete explanation of the Python code is **out of the scope** for this course (hint: learn the Python module `pdfquery`). It should be easy enough for you to understand how we capture the `porduct_name` from the pdf file using bounding box function `LTTextLineHorizontal:in_bbox("40, 48, 181, 633")` and then iterate over the products and search using regex and then only print the `Tandem` Manufacurers.

```
import re

import pdfquery
from lxml import etree


PDF_FILE = 'data.pdf'

pdf = pdfquery.PDFQuery(PDF_FILE)
pdf.load()

product_info = []
page_count = len(pdf._pages)
for pg in range(page_count):
    data = pdf.extract([
        ('with_parent', 'LTPage[pageid="{}"]'.format(pg+1)),
        ('with_formatter', None),
        ('product_name', 'LTTextLineHorizontal:in_bbox("40, 48, 181, 633")'),
    ])

    for ix, pn in enumerate(sorted([d for d in data['product_name'] if d.text.strip()], key=l
        if ix % 2 == 0:
            product_info.append({'Manufacturer': pn.text.strip(), 'page': pg, 'y_start': floa
            if ix > 0:
                product_info[-2]['y_end'] = float(pn.get('y0'))+10.0
        else:
            product_info[-1]['Model'] = pn.text.strip()

pdf.file.close()


for p in product_info:
    s = p['Manufacturer']
    m = re.search(r"Tandem",s,re.I)
    if m:
        print('Manufacturer: {}[Model {}]\n'.format(p['Manufacturer'],p['Model']))
```

We have preloaded the data onto educative.io's server and you should be able to run the code straight ahead and get the output as follows:

```
Manufacturer: Tandem Diabetes Care[Model T:flex]
Manufacturer: Tandem Diabetes Care[Model T:slim]
```

From this result we can see that there are two models `T:fles` and `T:slim` supplied by the manufacturer called 'Tandem Diabetes Care'. The problem solution has been adopted and simplified from the reddit user insainodwayno.