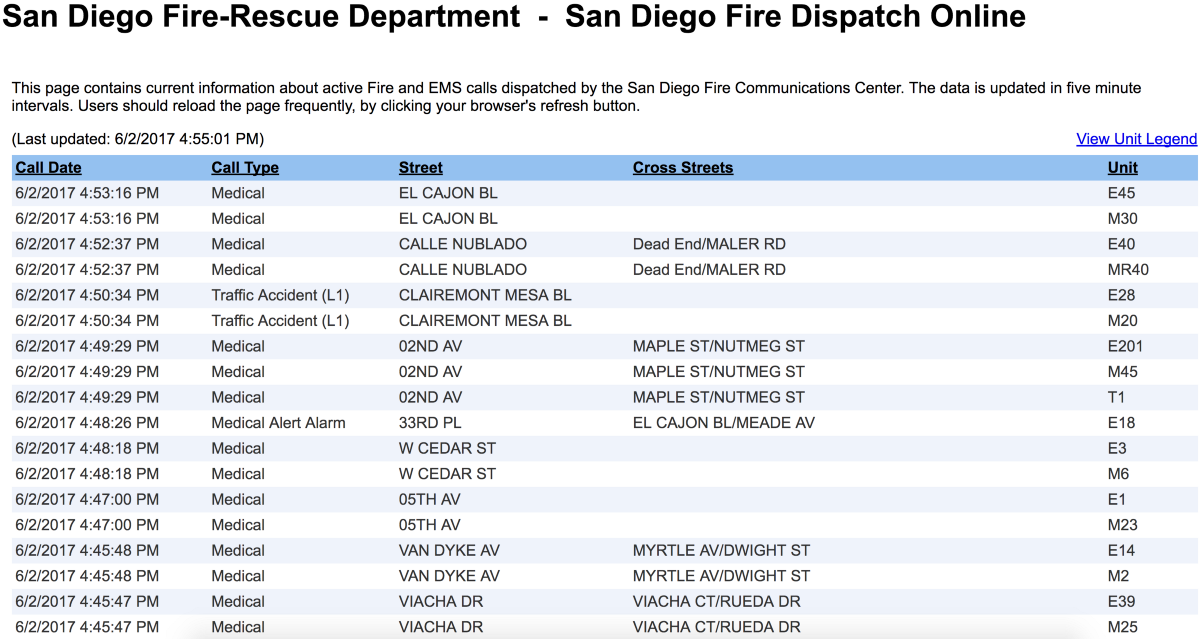
**Quick Tip: The easiest way to grab data out of a web page in Python**

**小贴士：在python中抓取网页数据最快的方式**

原文链接：https://medium.com/@ageitgey/quick-tip-the-easiest-way-to-grab-data-out-of-a-web-page-in-python-7153cecfca58

Let’s say you are searching the web for some raw data you need for a project and you stumble across a webpage like this:

让我们来做个假设，你需要查找一些源数据，建立一个项目然后偶然发现一个这样的网页



消防站网上处理记录dispatch

You found exactly what you need — an up-to-date page with exactly the data you need!

你发现的正是你所需要的-最新的页面中的数据

But the bad news is that the data lives inside a web page and there’s no API that you can use to grab the raw data. So now you have to waste 30 minutes throwing together a crappy script to scrape the data. It’s not hard, but it’s a waste of time that you could spend on something useful. And somehow 30 minutes always ends up being 2 hours.

For me, this kind of thing happens all the time.

但是坏消息是这些数据保存在网页中，并且没有可以用来获取源数据的API接口。因此你不得不浪费30分钟写一些脚本来爬取这些数据。这不是难事，但是这30分钟的时间本来可以做一些更有意义的事情。并且，有时候这30分钟的时间可能会以2个小时枯燥工作而收场。对我而言，这种情况总在发生。



Luckily, there’s a super simple answer. The [Pandas](https://pandas.pydata.org/) library has a built-in method to scrape tabular data from html pages called read\_html():

幸运的是，现在有一个超级简单的方式。Pandas类库有一个建立好的read\_html方法来爬取网页中表格类数据。

It’s that simple! Pandas will find any significant html tables on the page and return each one as a new [DataFrame](https://pandas.pydata.org/pandas-docs/stable/dsintro.html" \l "dataframe" \t "_blank) object.

就是这么简单。Pandas库会发现页面中所有有意义的表格并且把每一个表格作为一个新的DataFrame数据框架对象。

To upgrade our program from *toy* to *real*, let’s tell Pandas that row 0 of the table has column headers and ask it to convert text-based dates into time objects:

鸟枪换跑的感觉，有木有？让我们来给Pandas传递表格的列头并且转换为文本类型的时间对象格式

Which gives you this beautiful output:

展现出如此漂亮的输出：

And how that the data lives in a DataFrame, the world is yours. Wish the data was available as json records? That’s just one more line of code!

既然数据在DataFrame里了，那么这个世界就属于你了。想要json格式的数据？简单，就是一行代码的事情。

If you run that, you’ll get this beautiful json output (even with proper ISO 8601 date formatting!):

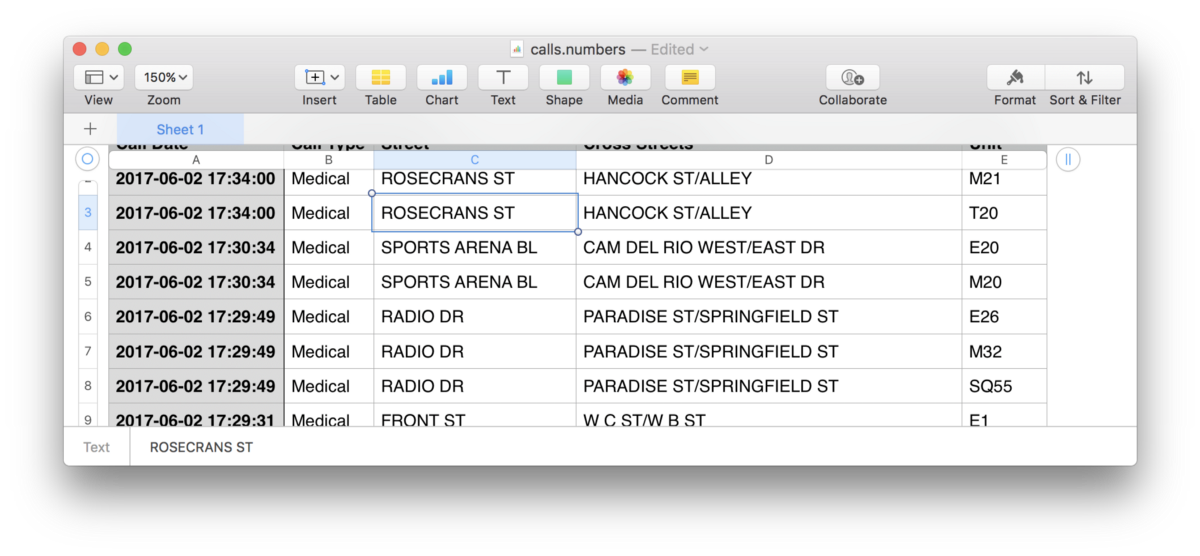
如果你运行这个，会得到优雅的json输出（甚至是ISO8601格式的日期类型）

You can even save the data right to a CSV or XLS file:

你甚至可以把数据保存成csv或者xls格式文件。

Run that and double-click on *calls.csv* to open it up in your spreadsheet app:

运行代码并且双击文件用表格程序去打开call.csv



And of course Pandas makes it simple to filter, sort or process the data further:

当然pandas让过滤、排序更简单，并且对数据的处理更深入。

None of this is rocket science or anything, but I use it so often that I thought it was worth sharing. Have fun!

这不是什么火箭科学或者其他高大上东西，但是我在工作中经常用到，因此我认为值得分享。祝旅途愉快！

Thanks for reading! If you are interested in machine learning (or just want to understand what it is), check out my [*Machine Learning is Fun!*](https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471) series too.