# 收集气象数据，用硒促进数据科学模型

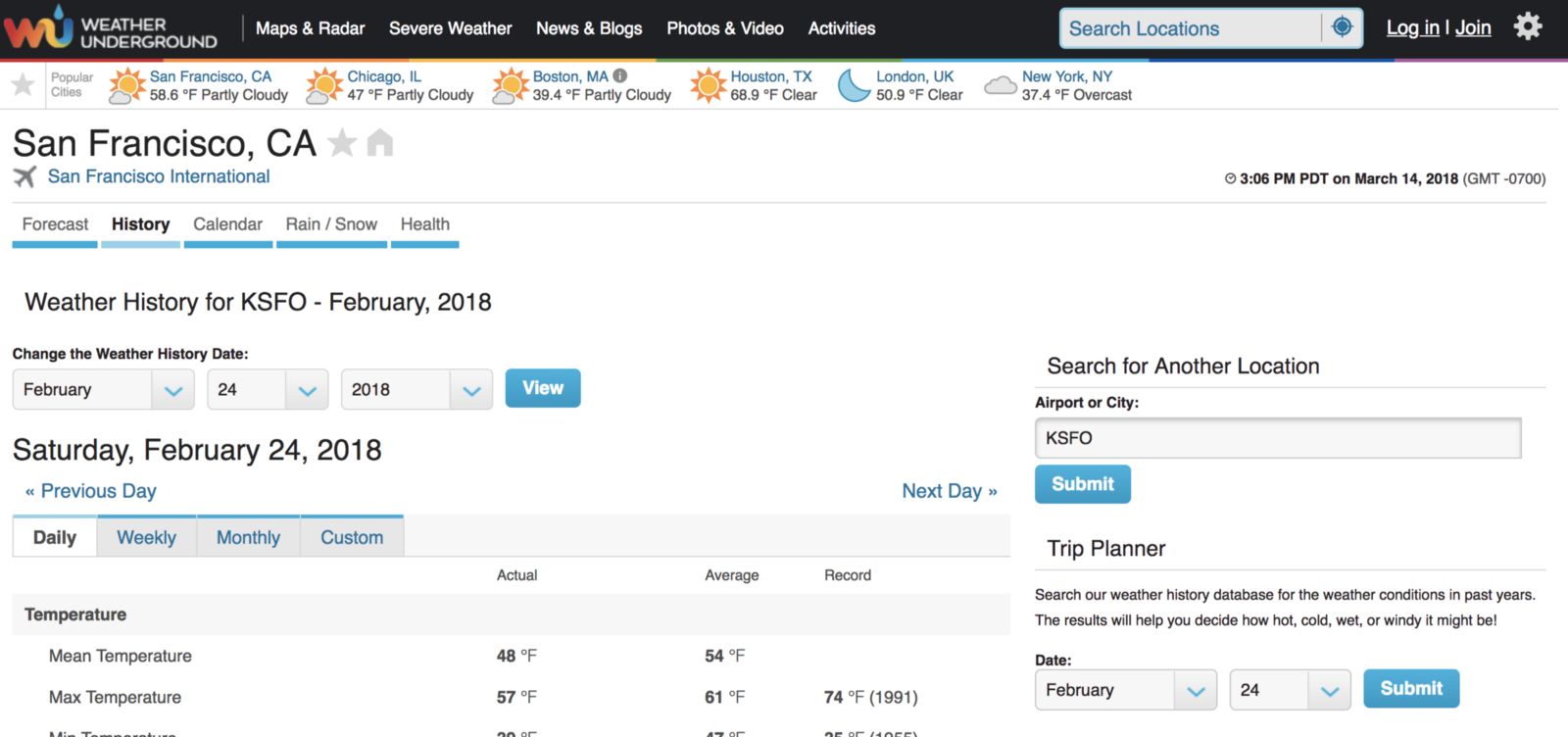


你有没有想过天气数据会对你的数据集产生什么影响，或者想了解你所在城市的历史天气模式？你觉得很难找到适合你特定需求的天气数据吗？是的。

我正在进行一个涉及地理空间数据的Kaggle竞赛，我想要与我的数据集兼容的更相关的数据。不幸的是，我什么也找不到。我花了好几天的时间寻找，我发现的只是一个网站，它只提供天气状况，比如某一天的天空是否多云或晴朗。这对我来说还不够。

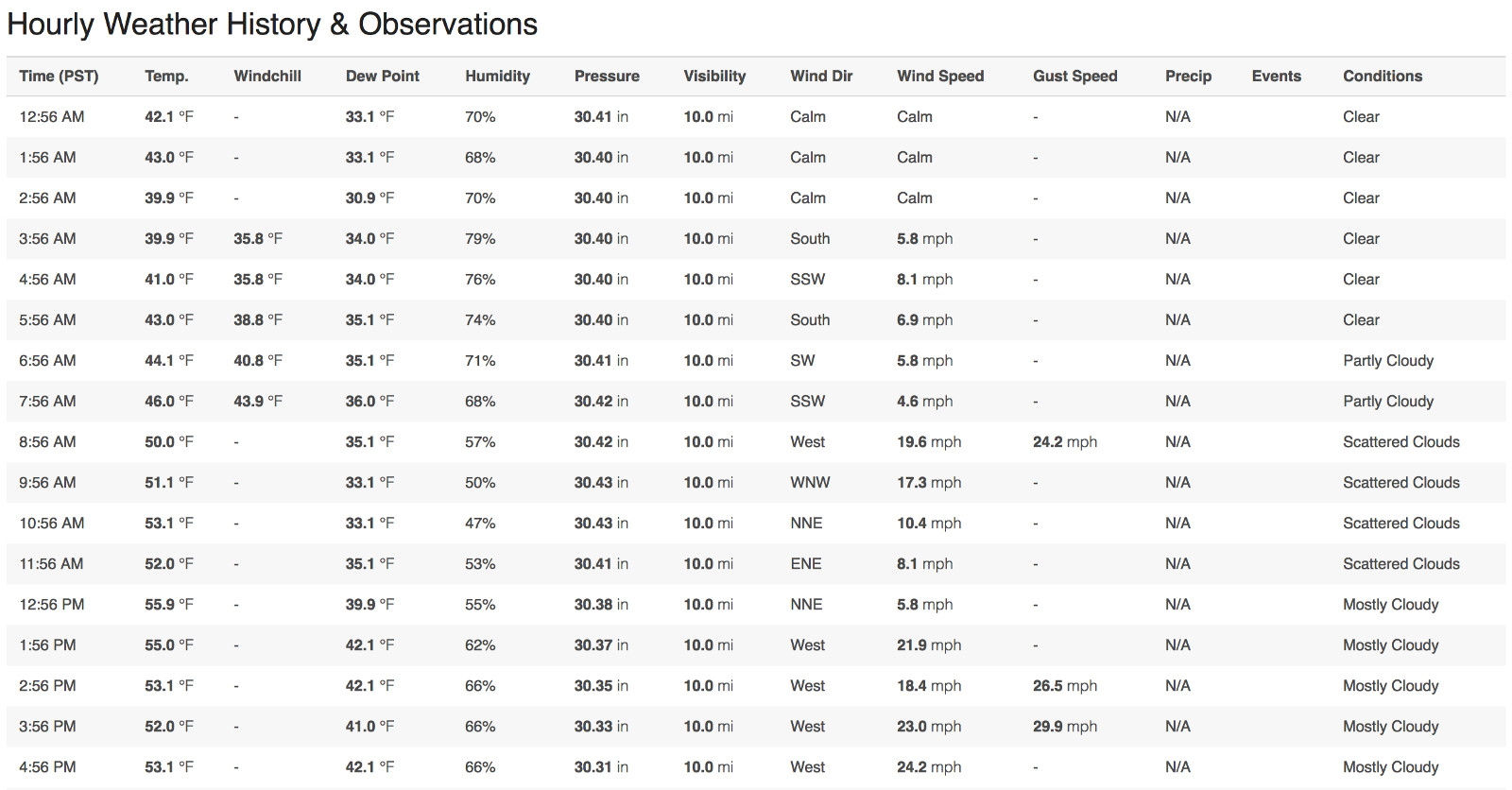
TLDR：因此，我建立了自己的web刮板，收集了自己的数据，并将其应用于我的数据集。这些结果产生了一个模型，该模型对新的数据集更加通用。现在你也可以通过这个教程来提高你的成绩。

我用Selenium用Python构建了一个web scraper，从中收集温度、湿度、风速和天气状况等气象信息。



使这个网站不同于你的典型网站的是，信息是基于附近机场的天气状况。虽然这听起来有悖常理，但事实上，对于靠近这些机场的城市和比典型地点信息量更大的城市来说，这是可以概括的。这使得使用邮政编码搜索最近的机场变得很容易。

如果你一直向下滚动到底部，你会看到一个气象数据的金矿。很漂亮吧？你可以在特定的日子或一系列的日子里挖掘这个。



下面是我如何从这个表中收集一系列日期的数据的步骤。如果要查看完整的代码片段，可以跳到页面底部。

步骤1：加载所需的包

Selenium是一个工具，用于通过Python或Java等编程语言通过浏览器自动执行任务和操作。您可以将其视为使用提供的命令来控制web浏览器。我们将使用它从HTML网页收集数据。

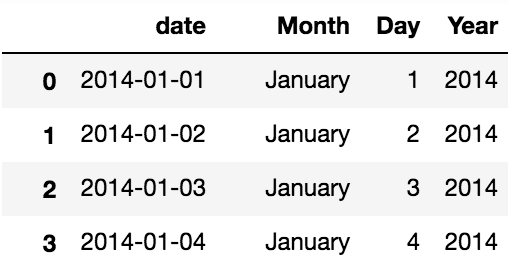
要初始化Selenium，需要在给定的路径加载webdriver。（以下是使用Python的Selenium安装指南：）

Chrome将加载在Chrome上运行Selenium所需的软件。driver.get（）将打开它可以控制的新web浏览器。

步骤2：获取要搜索的日期序列

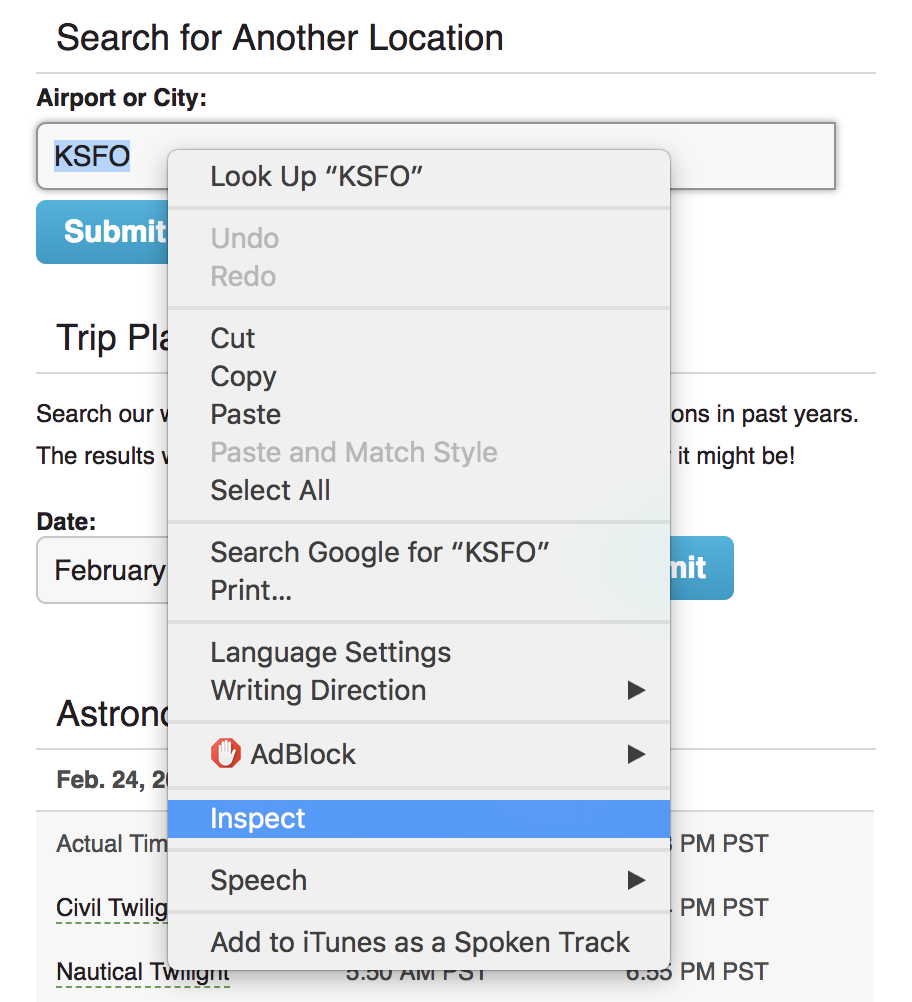
我编写了list\_dates（）函数来生成从开始日期到结束日期的一系列日期。这很有用，因为您不需要手动输入日期。

函数的作用是：将list\_date（）生成的日期分解为月、日和年，这样就可以很容易地将其发送到wunderground以搜索要从中收集数据的日期。例如，这两个函数都将生成以下pandas数据帧：



步骤3：按位置搜索

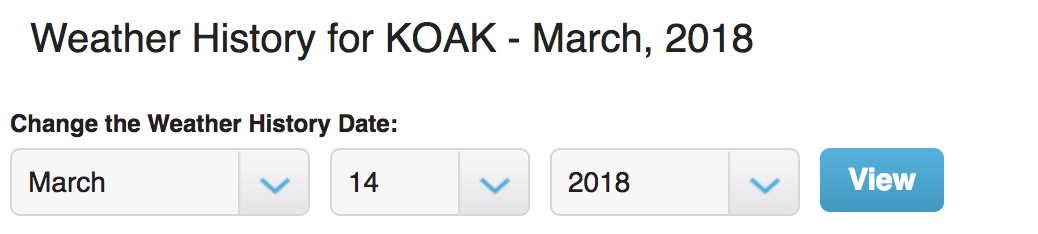
现在我们有了日期字典，你需要提供一个邮政编码，这样地下室就可以提取天气信息。您可以使用Selenium自动化此任务。如果您转到搜索栏，右键单击它，然后单击inspect，您将获得有关该搜索栏的HTML标记信息。这对网页上的任何HTML-JavaScript-CSS功能都是通用的。例如，在这里可以看到搜索栏的唯一“id”：

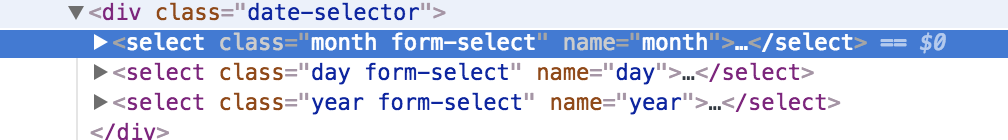




若要将邮政编码发送到搜索栏，以下代码将按具有“find\_element\_by\_id”的特定“id”查找搜索栏，清除搜索栏，发送邮政编码，最后将其提交到wunderground以搜索城市。

步骤4：搜索日期





您也可以类似地搜索步骤3这样的日期。通过查找与“天气历史日期”下拉菜单相关联的HTML标记，您将看到它包含诸如月、日和年之类的类属性。我们可以用它来发送信息来搜索下拉菜单。您可以遍历步骤2生成的pandas数据帧，以搜索给定日期的天气。

步骤5：收集天气数据



我们现在准备好收集天气数据了。我们可以看到天气表包含在这个div标签中，惟一的id是“observations\u details”。我们可以使用这些信息，用一行简单的代码来擦掉整个天气表：

**Step 6**: Bring it all together

Finally, we can combine all the previous steps into a compact script. I have also including some preprocessing steps and a function in runs everything with one function:

Congratulations! Now you can collect your weather data with Selenium and enhance all your Kaggle / work / school / side projects without compromising validation scores! You can also use what you’ve from this and collect data from other sources too!

References:

1. [Weather Underground](https://www.wunderground.com/)
2. [Selenium Download Page](https://www.seleniumhq.org/download/)
3. [Personal GitHub Repository](https://github.com/davidkes/Weather-Data-Scaper)