前言

好久不见，Python基础系列完结也有段时间了，希望帮到了大家。  
从今天开始我将开始更新一个新的系列：Python爬虫学习笔记。  
如你所见，本系列并不是复杂完备的教程，主要还是和大家一起分享我在学习Python爬虫的一些想法以及知识总结。  
如果你已经看完了我之前的Python基础系列文章，对Python的基本语法有了一定的概念，那么可能我接下来要开始写的东西可能会对你熟练运用Python有所帮助。

Python爬虫基础知识

1. 爬虫的基本概念

爬虫是一类用于信息搜集的程序，主要用于在一个或多个网页中爬取数据并进行保存、分类、分析等操作，目前最大的爬虫应该是各类搜索引擎，搜索引擎的实现原理简单来说就是他们部署了很多24小时不停扫描公网网站信息的大型爬虫程序，这些程序将爬取到的数据分类整理存储到数据数据库中，然后通过网站前端页面显示出网站标题、简介之类的信息，并提供了这些网站的网址让用户可以通过点击直接访问某一个网站。



image.png

个人其实基本没有那种需要爬全网之类的需求，所以爬虫的体量一般也比较小，大多是实现一些自动抢票、自动发帖、自动获取信息之类的功能。  
比如我现在需要一张北京到郑州的火车票，一直买不到，没得办法只能时不时刷新一下网页看有没有余票，但是人工刷新很难有那么好的运气能刷，所以这个时候就需要用到爬虫模拟人工，几秒钟刷新一次然后监控余票数值是否发生了变化。

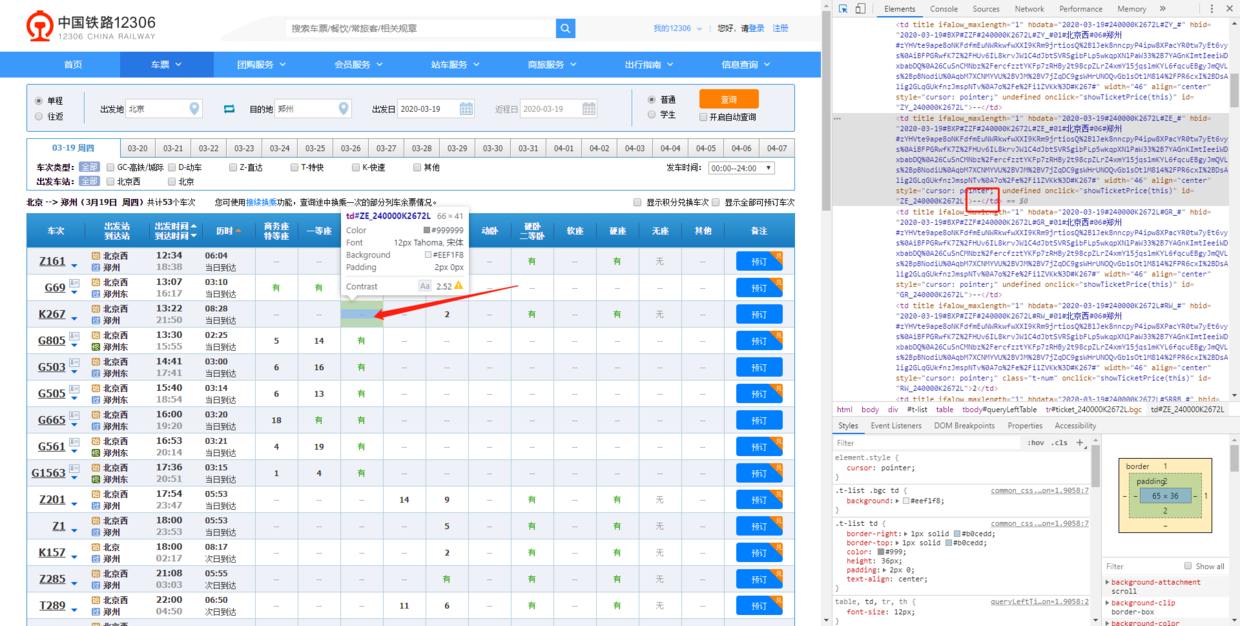


image.png

把爬虫放到服务器上24小时不间断运行，有票时这个字符会变为数值或“有”，这样就能第一时间通知你去抢票了。很多抢票软件的原理其实就是这样，没什么黑科技，就是让爬虫时刻检测是否有票。

2. 网页基础

我们在浏览器访问的每一个页面背后其实都是成千上万行的代码所组成的，而想要一个动态网站跑起来需要涉及到很多技术，有负责页面展示布局的html,css和动态交互的JavaScript、负责动态处理用户请求的后端开发语言（PHP/Java/python/golang)、数据库技术、web服务器软件等等。  
在学习爬虫时，我们经常接触的主要还是html和JavaScript。不过我还是建议你能够花一点时间去了解一下动态网页搭建的基础知识，这对于之后的爬虫学习会有很大帮助。  
网站的布局代码是开放的，也就是每个用户通过一些操作都可以看到一个网站的布局代码。基本上所有的浏览器都会提供查看网页源代码功能，一般情况下快捷键为crtrl+u按下之后可以看到网页的源代码了

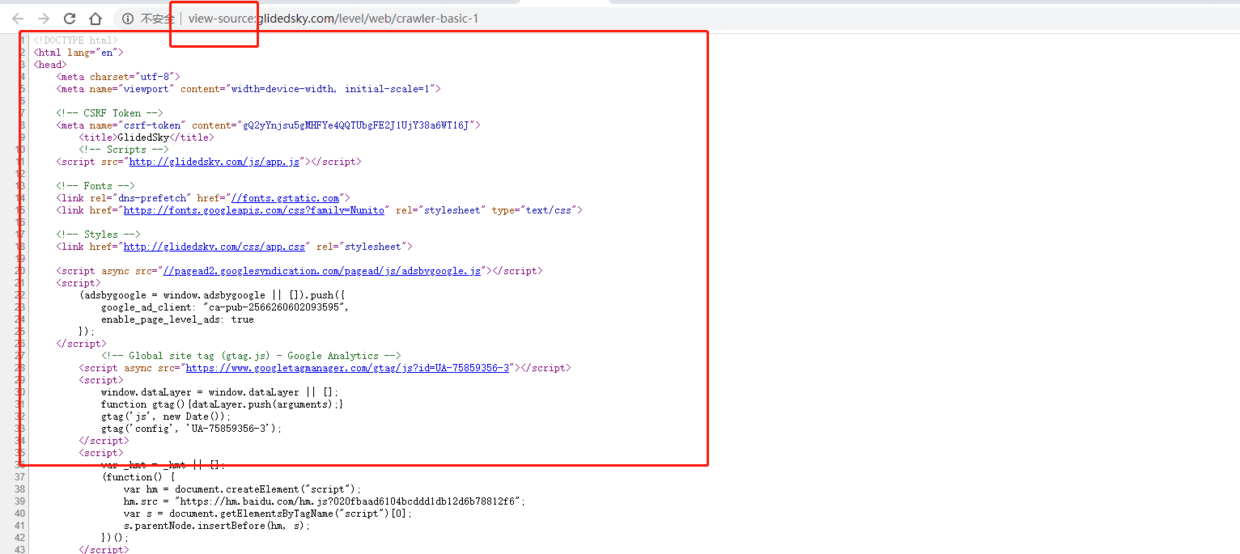


image.png

由于太长了所以只截了局部，这些代码最终形成的效果是这样的

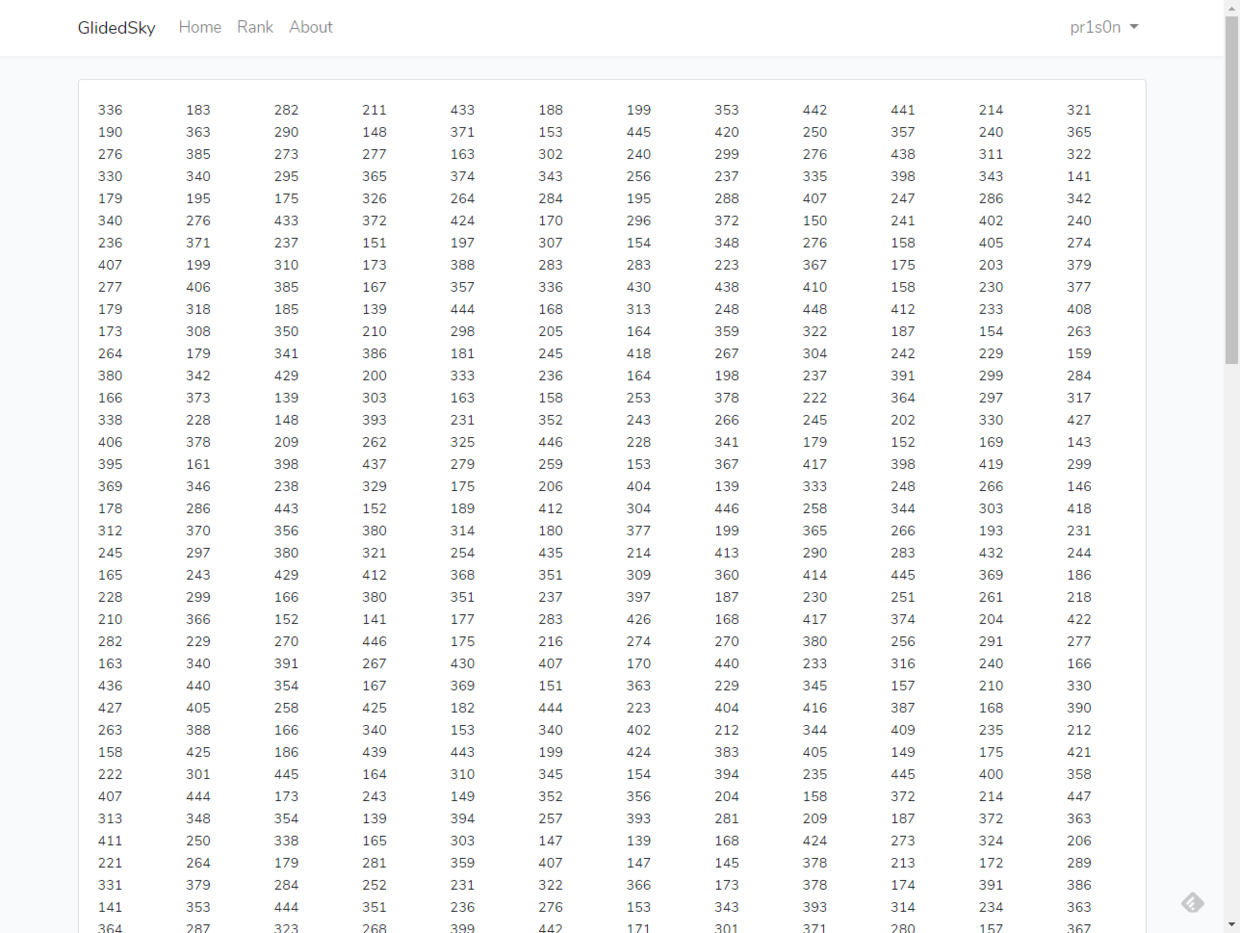


image.png

浏览器的作用就是将这些代码解析为相应的样式，前端开发者在实际开发过程中往往是写了一个样式之后就需要通过浏览器实时预览效果，并通过工具进行调整，在浏览器中按下f12，即可调出开发者工具。

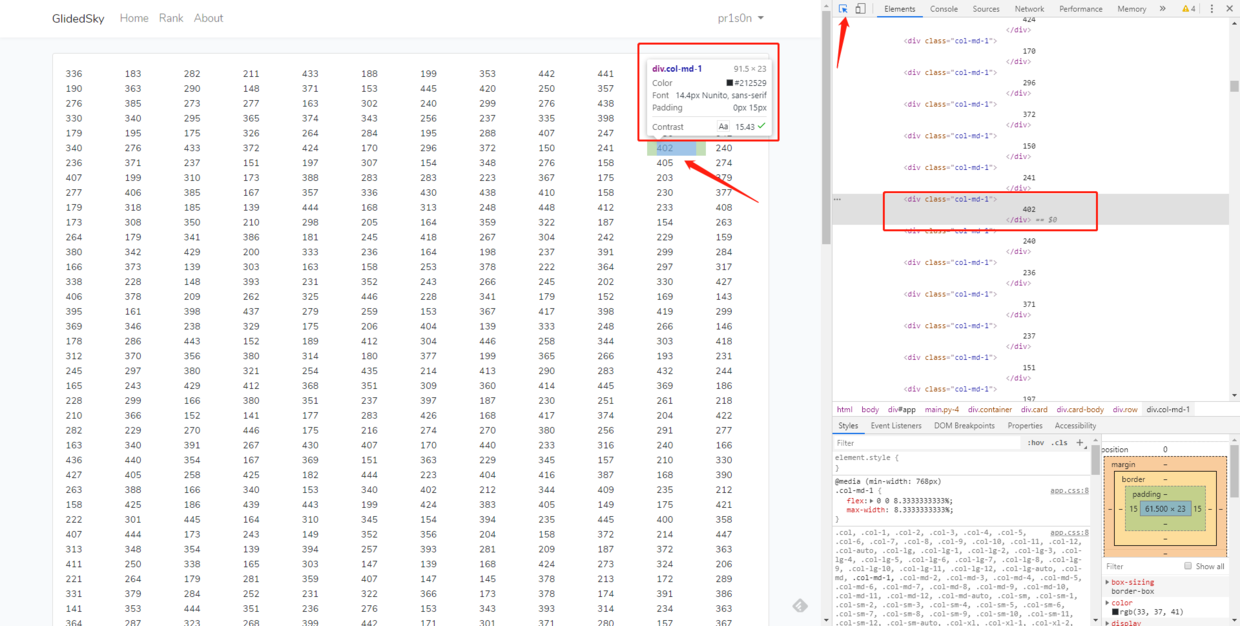


image.png

点击这个小箭头让它处于激活状态后在原网页中用鼠标选择一个元素，即可快速定位到这个元素在源代码中的位置以及显示CSS样式信息。  
如果暂时实在看不懂这些代码也没有关系，如果将一个网页看作是一个机器人的话



那么html（超文本编辑语言）代码即为机器人的零部件

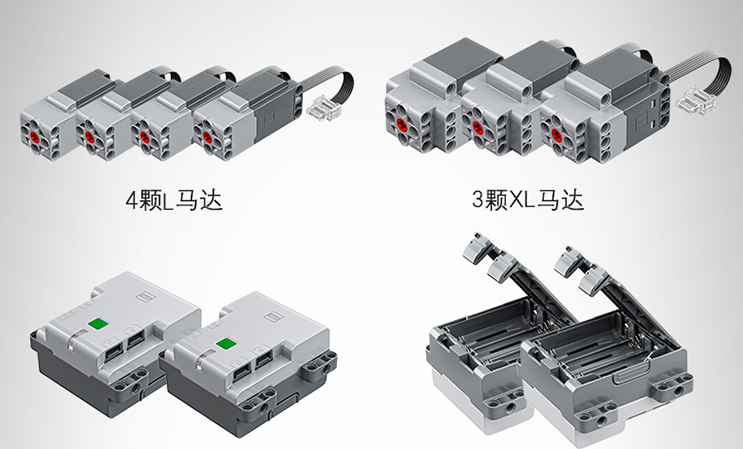


image.png

而CSS（层叠样式表）则为组装图纸

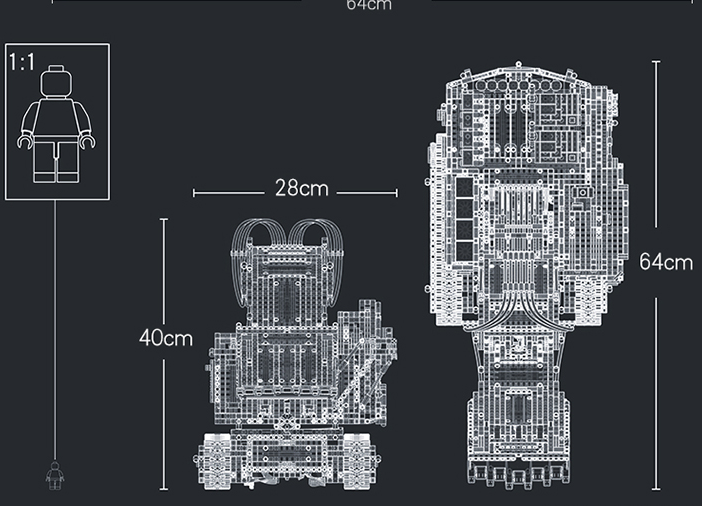
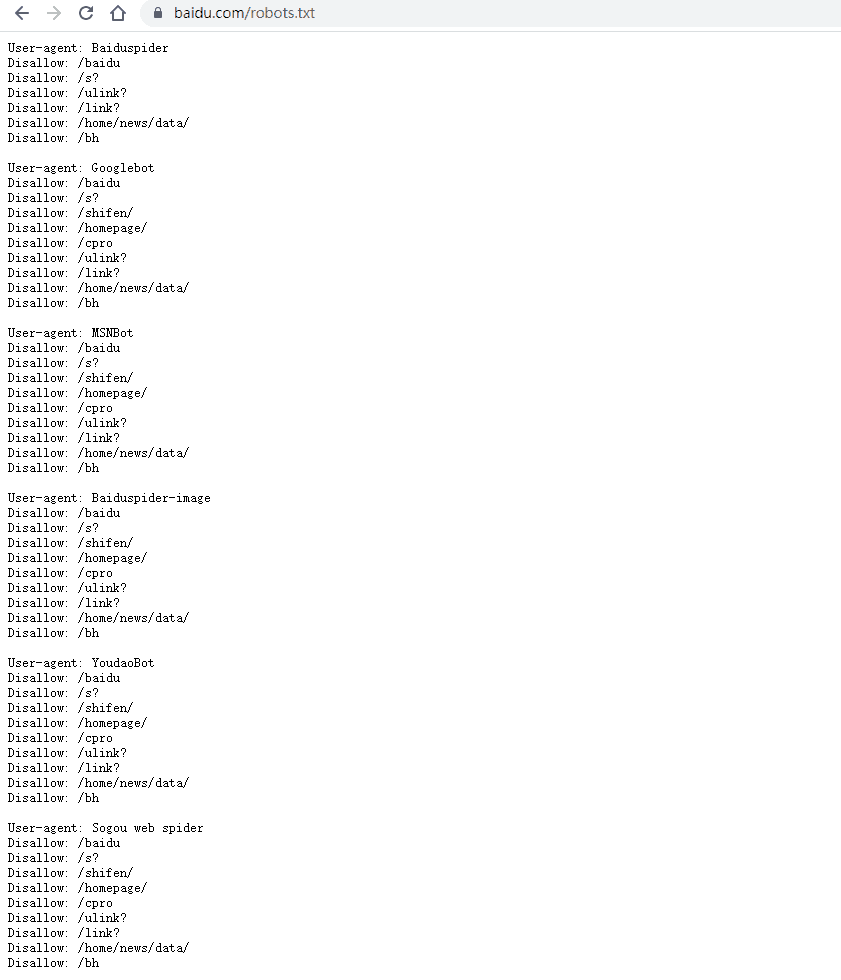


image.png

JavaScript让这个机器人可以自动摆臂或者旋转。  
（如果看到这还有女孩子，建议将机器人换为闪耀暖暖里的娃娃，零部件就是衣服，javaScript负责改变哇娃娃的动作和表情）

爬虫目标

一般来说写一个爬虫之前需要先明确：  
1. 目标网站  
2. 数据内容  
3. 是否合法  
首先需要确定自己想要爬取哪个网站的什么数据，然后很重要的一点大部分的网站并不欢迎爬虫，有一些较为隐私的目录、数据会明确告诉告诉你不能爬取。  
关于这个爬取范围，首要根据就是robots协议，这个协议简单来说就是在网页的根目录下定义一个robots.txt文件，里面定义了哪些属于可爬取的公开数据，哪些是网站禁止爬取的目录或文件



上图为百度根目录下robots.txt文件内容，其中还根据不同的搜索引擎进行了限制，凡是disallow后的目录均为非公开页面。  
一旦爬虫违反了这个协议，那么网站就有充分理由提出司法诉讼。  
就像老师提前告诉你交白卷会挂科，你就迎着头皮交白卷，老师肯定饶不了你。



另外，即使没有违背robots协议，如果你的爬虫运行时对网站造成了恶劣影响，比如访问太过于频繁导致网页崩了或者影响了其他正常用户的使用，又或者对网站造成了经济损失（比如将数据卖给竞争对手），网站都是可以申请送你一堆紫金手镯的。



image.png

到时候可就是Python爬虫入门失败，入狱成功了。

爬虫的简单工作流程

从上面的抢票例子也可以看得出来，爬虫的工作流程其实和人浏览网页基本是一样的，都是打开网页，找到对应元素，判断元素是否发生了变化或者是保存某些数据。  
不同的人在看的时候存储数据用的是小脑瓜，程序用的是硬盘。  
还有一个不同点在于程序并不在意这个网页长什么样，很明显他们都是直男，只会在网页的源代码中找自己需要的东西。  
所以爬虫的简单工作流程是这样的：

访问网页获取源代码

分析源代码获取指定数据

操作数据或执行其他命令

你看，爬虫做的最多的其实是对获取到的源代码进行分析，只不过它并不能理解代码，只是把这些源代码当作是一个比较长的字符串。  
对字符串进行操作就是我们之前在Python基础里面讲过，所以不用害怕，爬虫也可以很简单。