第一个爬虫

首先,我们还是需要回顾一下爬虫的概念. 爬虫就是我们通过我们写的程序去抓取互联网上的数据资源. 比如, 此时我需要百度的资源. 在不考虑爬虫的情况下, 我们肯定是打开浏览器, 然后输入百度的网址, 紧接着, 我们就能在浏览器上看到百度的内容了. 那换成爬虫呢? 其实道理是一样的. 只不过, 我们需要用代码来模拟一个浏览器, 然后同样的输入百度的网址. 那么我们的程序应该也能拿到百度的内容. 对吧~

在python中, 我们可以直接用urllib模块来完成对浏览器的模拟工作~, 直接上代码

```
1 from urllib.request import urlopen
2
3 resp = urlopen("http://www.baidu.com")  # 打开 百度
4 print(resp.read().decode("utf-8"))  # 打印 抓取到的
  内容
```

是不是很简单呢?

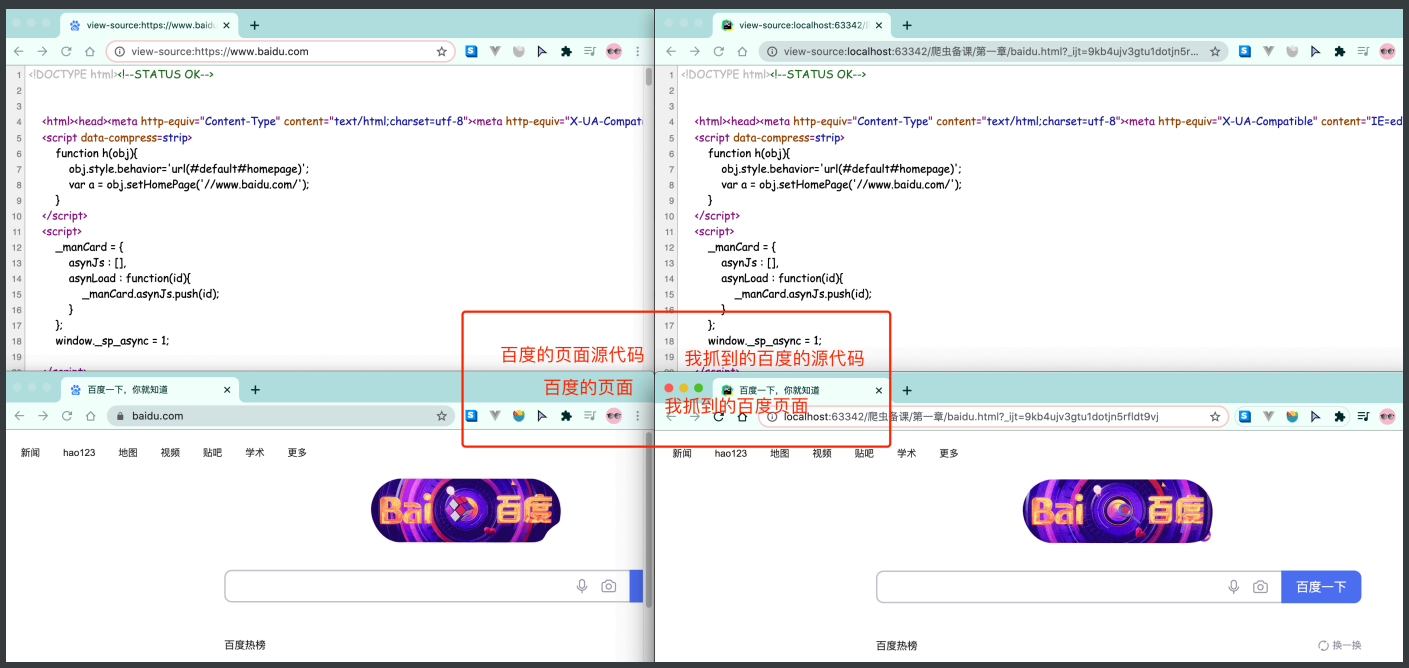
我们可以把抓取到的html内容全部写入到文件中, 然后和原版的百度进行对比, 看看是否一致

```python
from urllib.request import urlopen

resp = urlopen("http://www.baidu.com")  # 打开 百度

# print(resp.read().decode("utf-8"))  # 打印 抓取到的内容

with open("baidu.html",mode="w", encoding="utf-8") as f:  # 创建文件
    f.write(resp.read().decode("utf-8"))  # 保存在文件中
```



OK ~ 我们成功的从百度上爬取到了一个页面的源代码. 就是这么简单, 就是这么炫酷.

你也试一下吧~