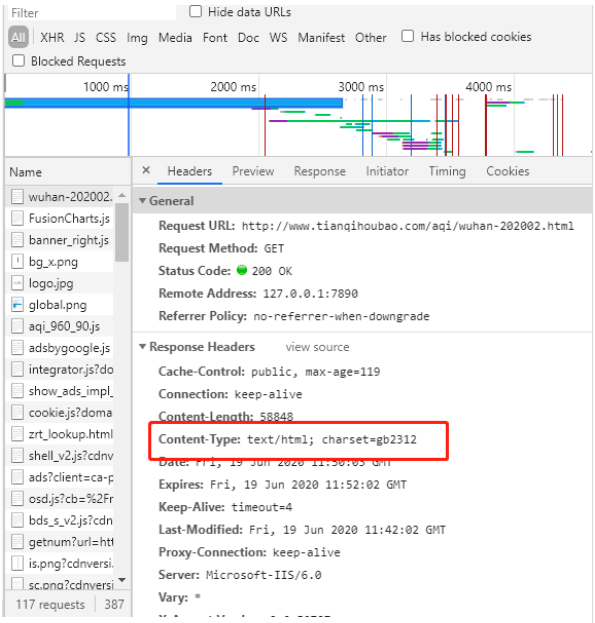
代码说明

网页分析

首先选取武汉2020年5月空气质量网页

http://www.tianqihoubao.com/aqi/wuhan-202005.html



2020年2月武汉空气质量指数AQI_PM2.5历史数据

分享到:

武汉02月份空气质量指数(AQI)数据： 数值单位：ua/m^3(CO为ma/m^3)

日期	质量等级	AQI指数	当天AQI排名	PM2.5	PM10	So2	No2	Co	O3
2020-02-01	良	89	248	65	73	10	34	1.03	67
2020-02-02	良	87	288	64	69	7	34	0.85	49
2020-02-03	良	98	324	73	79	8	29	1.03	62
2020-02-04	良	87	299	64	73	9	35	0.92	61
2020-02-05	轻度污染	125	344	94	100	9	33	1.21	60
2020-02-06	良	54	248	40	52	5	14	0.78	49
2020-02-07	优	26	73	18	19	5	11	0.56	43
2020-02-08	优	33	98	23	25	5	16	0.70	43
2020-02-09	优	44	133	30	34	5	19	0.76	53
2020-02-10	良	59	208	42	44	7	27	1.00	45
2020-02-11	优	40	91	28	32	6	16	0.96	42
2020-02-12	优	38	85	26	31	7	23	0.98	31
2020-02-13	优	35	82	24	31	9	23	1.23	32
2020-02-14	优	39	174	27	34	7	17	1.19	34
2020-02-15	优	28	164	15	26	4	10	0.80	52
2020-02-16	优	23	78	10	21	5	11	0.60	59
2020-02-17	优	26	91	14	22	5	17	0.58	62
2020-02-18	优	33	95	22	25	8	22	0.71	65
2020-02-19	优	42	98	29	33	11	21	0.84	70

在上图发现该网站支持的编码格式为gb2312

获取网页

首先在这里必须先导入requests库，然后构造url向服务器发送请求，获取响应。拿到网页源代码

代码如下：

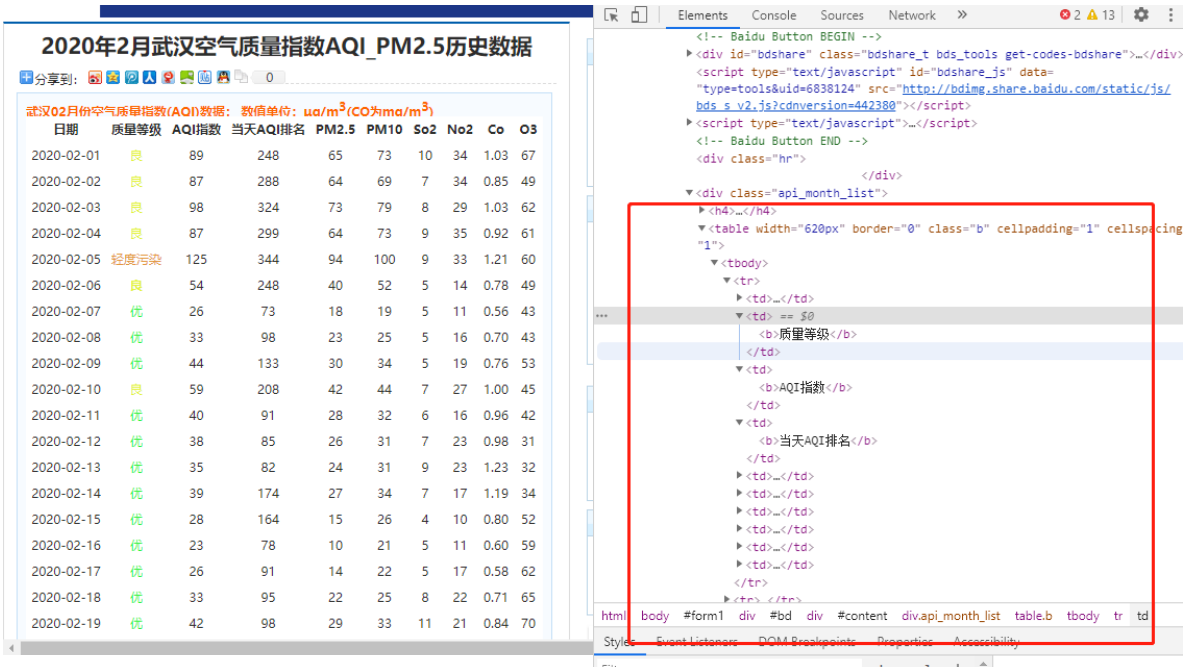
```
# 构造url
for city in ('wuhan', 'chengdu'):
        for report_time in range(202001,202006):
            url = 'http://www.tianqihoubao.com/aqi/%s-%d.html' % (city,
report_time)

# 获取网页源代码
def get_html(url):
    response = requests.get(url, headers)
    print(response.status_code)
    return response.text
```

解析数据

在这里由于我们需要抓取的数据都保存在table内，因此我们通过pandas的read_html可以很快的得到数据。

需要抓取的数据内容如下图：



2020年2月武汉空气质量指数AQI_PM2.5历史数据

分享到：

武汉02月份空气质量指数(AQI)数据： 数值单位：ua/m3(CO为mg/m3)

日期	质量等级	AQI指数	当天AQI排名	PM2.5	PM10	So2	No2	Co	O3
2020-02-01	良	89	248	65	73	10	34	1.03	67
2020-02-02	良	87	288	64	69	7	34	0.85	49
2020-02-03	良	98	324	73	79	8	29	1.03	62
2020-02-04	良	87	299	64	73	9	35	0.92	61
2020-02-05	轻度污染	125	344	94	100	9	33	1.21	60
2020-02-06	良	54	248	40	52	5	14	0.78	49
2020-02-07	优	26	73	18	19	5	11	0.56	43
2020-02-08	优	33	98	23	25	5	16	0.70	43
2020-02-09	优	44	133	30	34	5	19	0.76	53
2020-02-10	良	59	208	42	44	7	27	1.00	45
2020-02-11	优	40	91	28	32	6	16	0.96	42
2020-02-12	优	38	85	26	31	7	23	0.98	31
2020-02-13	优	35	82	24	31	9	23	1.23	32
2020-02-14	优	39	174	27	34	7	17	1.19	34
2020-02-15	优	28	164	15	26	4	10	0.80	52
2020-02-16	优	23	78	10	21	5	11	0.60	59
2020-02-17	优	26	91	14	22	5	17	0.58	62
2020-02-18	优	33	95	22	25	8	22	0.71	65
2020-02-19	优	42	98	29	33	11	21	0.84	70

代码如下：

```
def parse_html(html):
    # header 标题行，为0表示取消标题行
    data = pd.read_html(html, header=0,encoding='utf-8')[0]
    return data
```

保存数据

由于通过read_html获得的数据为DataFrame对象，因此在这里我们直接通过DataFrame对象方法to_excel将数据保存到excel

代码如下：

```
def data_save(data,filename):
    data.to_excel('%s.xlsx' % filename,sheet_name='空气质量指数')
    print(filename,'完成！')
```

可视化输出

在这里，我们通过pandas读取我们需要的excel为DataFrame对象，然后通过索引的方式获取到做图数据'日期'和'AQI指数'，最后通过matplotlib输出结果图。

代码如下：

```
# 利用matplotlib可视化输出
mpl.rcParams['font.sans-serif'] = ['SimHei']


for city in ('wuhan', 'chengdu'):
        for report_time in range(202001,202006):
            filename = '%s-%d' % (city, report_time)
            df = pd.read_excel('%s.xlsx' % filename, header=0)
            x = df['日期']
            y = df['AQI指数']
            if city == 'wuhan':
                city_name = '武汉'
            elif city == 'chengdu':
                city_name = '成都'
            plt.suptitle('%s%s月空气质量指数(AQI)' % (city_name, str(report_time)
[-1]))
            plt.bar(range(len(x)), y, color='lightsteelblue')
            plt.plot(range(len(x)), y, marker='o', color= 'coral')
            plt.xticks(range(len(x)),x)
            plt.xticks(rotation=90)
            plt.legend(['AQI'])
            plt.show()
```