

实习一 Python 爬虫和数据处理

钱一鸣 1801221564

要求：使用 Python 语言，爬取感兴趣的网站信息，保存成 csv 文件，导入和处理数据，最后进行数据分析和可视化，最后提交 pdf 格式的 report 以及代码压缩包，报告中应有关键代码、遇到的问题以及怎么解决的、运行结果截图，pdf 文件名和 report 注明学号和姓名。

比如豆瓣图书/电影信息，股票信息，微博知乎各大网站的粉丝关系、用户画像，等等。

1. 任务介绍

使用 python 爬取 aqistudy.cn 上的空气质量在线数据，地点设置为北京地区，结果以 CSV 格式保存在文本文件中。

2. 前期准备

环境：Windows 10, Python 3.7, Google Chrome(80.0.3987.149 (正式版本)), ChromeDriver(80.0.3987.106/)

3. 爬取过程

1) 爬取链接的选择：

对比了几个可以查历史数据的网站后，选取的雾霾数据网站：

<https://www.aqistudy.cn/historydata/>.

此网站的数据易于 Python 抓取且来源于环保部官方数据，支持查询时间从最早 2013 年 12 月到现在的记录，全国各主要城市的历史数据页面如下图。

空气质量历史数据查询

热门城市：

北京 上海 广州 深圳 杭州 天津 成都 南京 西安 武汉

全部城市：

A. 阿坝州 安康 阿克苏地区 阿里地区 阿拉善盟 阿勒泰地区 安庆 安顺 鞍山 克孜勒苏州

B. 蚌埠 白城 保定 北海 宝鸡 北京 毕节 博州 白山 百色

C. 长春 昌都 常德 成都 承德 赤峰 昌吉州 五家渠 昌江 澄迈


D. 池州 长沙 崇左 楚雄州 朝阳 沧州 长治 达州 大理州 大庆 大同 定西

E. 定安 大兴安岭地区 丹东 东方 东莞 德宏州 德州 德州 德州 德州

全国主要城市历史数据

北京	上海	天津
重庆	杭州	哈尔滨
长春	沈阳	石家庄
太原	西安	济南
乌鲁木齐	拉萨	西宁
兰州	银川	郑州
南京	武汉	合肥
福州	南昌	长沙
贵阳	成都	广州
昆明	南宁	深圳

微信公众平台



2) 爬取数据的选择：

本次需要爬取的是北京空气质量指数历史数据，打开 aqistudy.cn 空气质量网站，包括 PM2.5, PM2.5, SO2 等，即是从下列网页内置的表格中爬取，时间范围从 2013-12-02 到 2020-10-17，存储为 csv 文件格式。



北京空气质量指数月统计历史数据



3) 爬取数据的过程:

a) 初始化: `driver = webdriver.PhantomJS(executable_path=r'phantomjs-2.1.1-windows\bin\phantomjs.exe')`

b) 为了防止网站针对爬虫的限制, 把爬虫伪装成浏览器: `headers = { 'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/55.0.2883.87 Safari/537.36' }`

c) 定义需要爬取的时间范围函数, 从 2013-12-02 到 2019-12-31:

```
def get_month_set():
    month_set = list()
    for i in range(12, 13):
        month_set.append(('2013-%s' % i))
    for i in range(1, 10):
        month_set.append(('2014-0%s' % i))
    .....
    for i in range(10, 13):
        month_set.append(('2019-%s' % i))
    return month_set
```

d) 设定需要爬取的城市, 并保存到“城市”+.csv 文件中: `city = '北京':`

`file_name = city + '.csv'`

f) 定义需要爬取的字段范围, 需要分别获取的信息如下:

属性	含义	单位
时间	2013.12~2019.12	天
AQI	空气质量指数	0~600

Grade	空气质量等级	优、良、轻度污染、中度污染、重度污染、严重污染
Pm25	直径小于等于 2.5 微米的颗粒物	μg/m3
Pm10	10 微米以下的颗粒物	μg/m3
SO2	二氧化硫污染物浓度	μg/m3
CO	一氧化碳污染物浓度	mg/m3
NO2	二氧化氮污染物浓度	μg/m3
O3	臭氧污染物浓度	μg/m3

代码如下所示：

```

for i in range(len(month_set)):
    str_month = month_set[i]
    weburl = ('%s%s&month=%s' % (base_url, parse.quote(city),
str_month))

    driver.get(weburl)
    dfs = pd.read_html(driver.page_source,header=0)[0]
    time.sleep(1)#防止页面一带而过，爬不到内容
    for j in range(0,len(dfs)): date =
        dfs.iloc[j,0] aqi =
        dfs.iloc[j,1] grade =
        dfs.iloc[j,2] pm25 =
        dfs.iloc[j,3] pm10 =
        dfs.iloc[j,4] so2 =
        dfs.iloc[j,5] co =
        dfs.iloc[j,6] no2 =
        dfs.iloc[j,7] o3 =
        dfs.iloc[j,8] print(date)
        print(aqi)
        fp.write((' %s,%s,%s,%s,%s,%s,%s,%s,%s\n' %
(date,aqi,grade,pm25,pm10,so2,co,no2,o3)))
        print('%d---%s,%s---DONE' % (city.index(city), city,
str_month))

```

4) 爬取的结果：

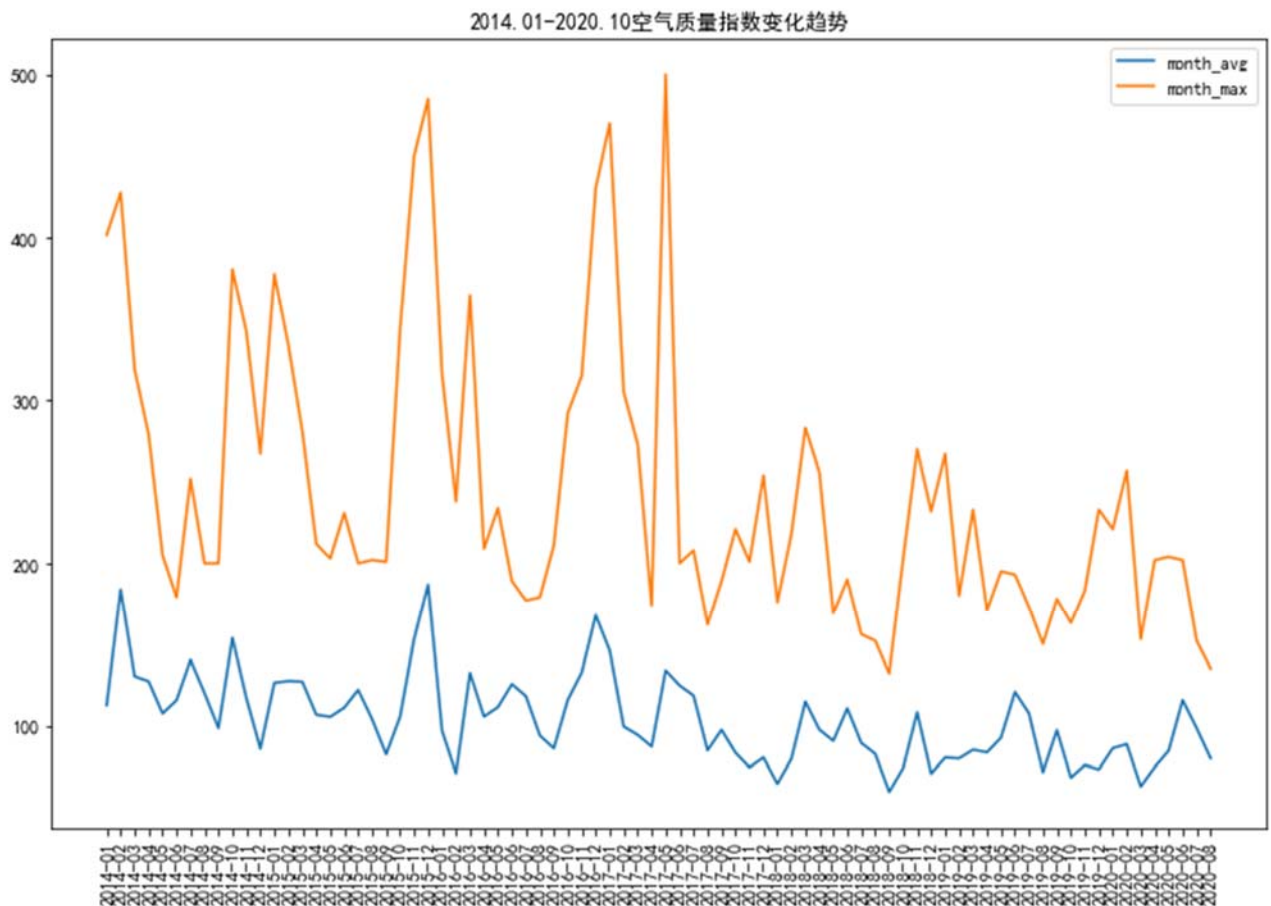
本次需要爬取的是北京空气质量指数历史数据，包括 PM2.5，PM2.5，SO2 等共计 2189 条数据，最后将爬取结果存储为 csv 文件格式。爬取的结果如下所示：

date	AQI	grade	PM25	PM10	SO2	CO	NO2	O3_8h
2013/12/2	142	轻度污染	109	138	61	2.6	88	11
2013/12/3	86	良	64	86	38	1.6	54	45
2013/12/4	109	轻度污染	82	101	42	2	62	23
2013/12/5	56	良	39	56	30	1.2	38	52
2013/12/6	169	中度污染	128	162	48	2.5	78	15
2013/12/7	291	重度污染	241	285	64	4.2	98	6
2013/12/8	223	重度污染	173	189	47	2.9	60	41
2013/12/9	26	优	11	16	10	0.6	22	51
2013/12/10	45	优	21	45	14	1	29	52
2013/12/11	30	优	19	30	15	0.7	30	45
2013/12/12	29	优	16	29	11	0.8	25	56
2013/12/13	66	良	48	63	29	1.3	45	29
2013/12/14	56	良	40	48	29	1.2	41	46
2013/12/15	64	良	46	55	31	1.5	49	31
2013/12/16	134	轻度污染	102	126	59	2.5	70	10
2013/12/17	80	良	59	41	35	1.4	39	42
2013/12/18	45	优	29	45	22	0.9	32	43
2013/12/19	63	良	45	60	30	1.2	50	35

3. 数据分析与可视化

1) 空气质量指数变化：

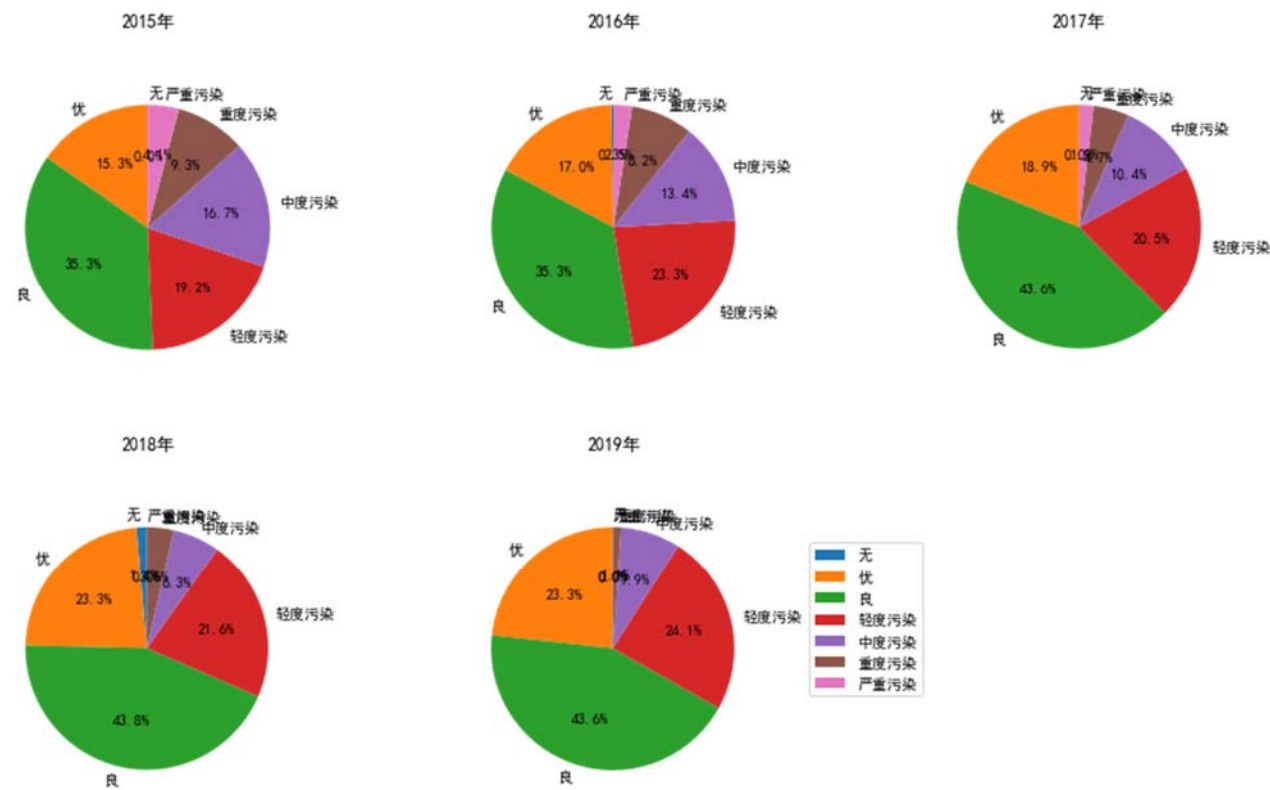
画出每个月平均、以及最大值的 aqi 随时间的变化趋势，如下图所示：



从上图可以看出，空气质量指数整体呈下降趋势。对于每一年来说，每年的秋冬季（集中在 10 月-次年 2 月）AQI 值较高，空气质量较差。

2) 逐年污染等级的占比饼图:

按['严重污染' '中度污染' '优' '无' '良' '轻度污染' '重度污染']不同等级对从 2014 年 1 月至 2019 年 12 月的污染等级绘制占比饼图, 如下图所示:



从上图可以看出, 空气质量状况逐年好转, 污染天数占比逐年减少, 严重污染天数到 2018 年几乎为 0。

4.参考资料

https://blog.csdn.net/weixin_40651515/article/details/84592530, python 爬虫爬取（中国空气质量在线监测分析平台）北京 PM2.5，2013 年至 2018 年的数据。

https://blog.csdn.net/bzd_111/article/details/50496500 , python 初学 selenium+phantomjs 遇到的问题。

<https://www.jianshu.com/p/4b2205ffefe5>, Python3 爬取城市历史 PM2.5 数据。