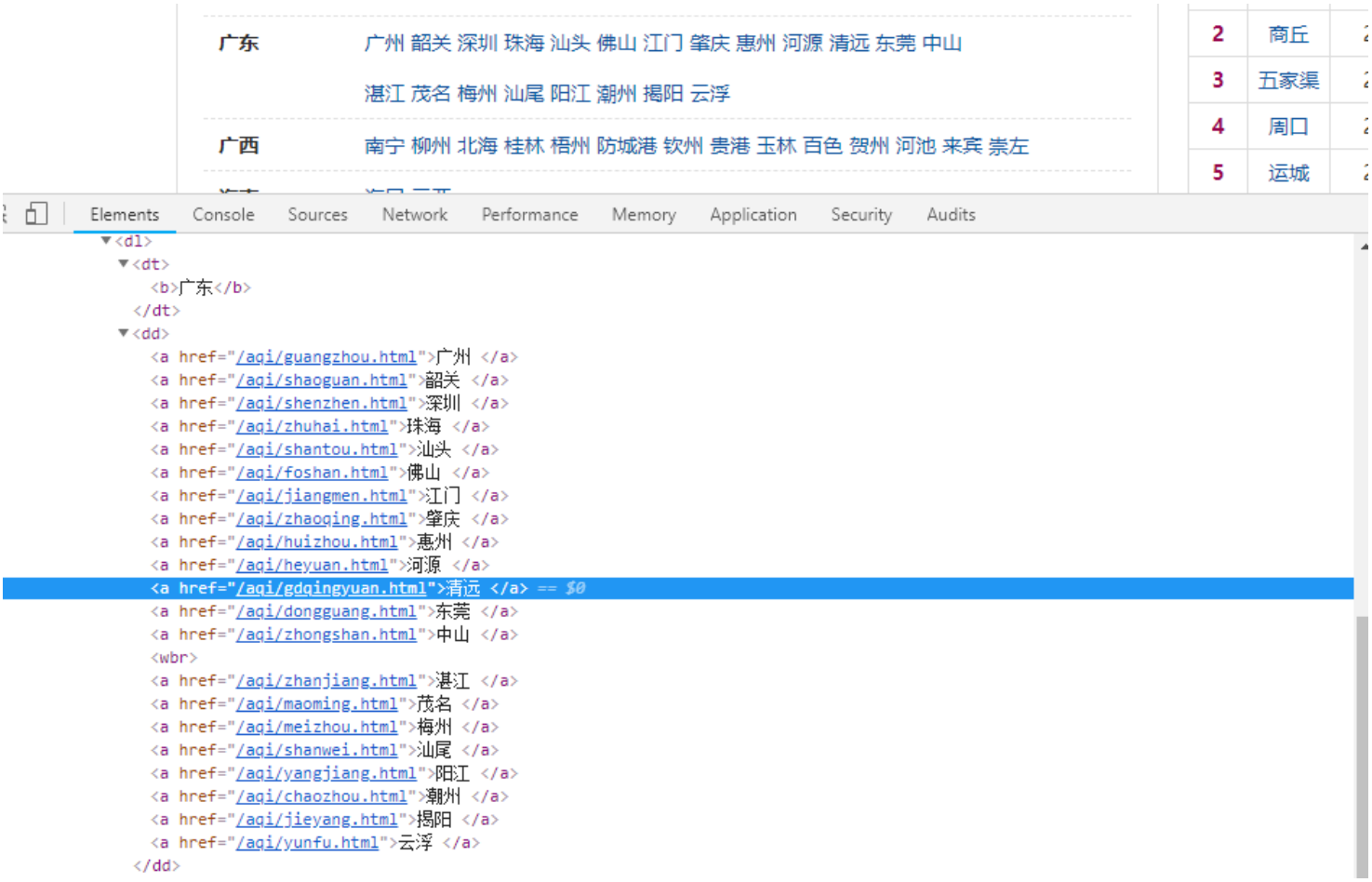
（十）空气质量分析

笔记本: Markdown笔记
创建时间: 2018/11/24 16:33 **更新时间:** 2018/12/1 22:27
作者: FWJ
URL: http://business.sohu.com/20140606/n400501809.shtml

分析网页

广东所有城市

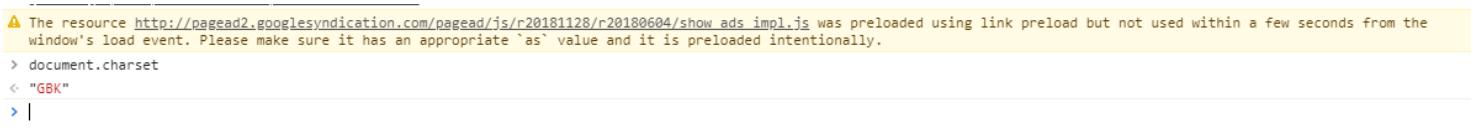


以佛山为例：
11月份空气质量

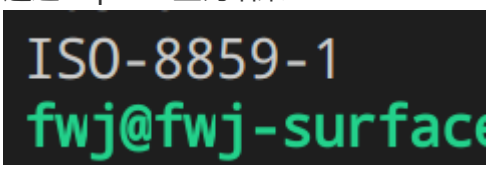
http://www.tianqihoubao.com/aqi/foshan-201811.html

网页编码问题

网页编码



通过requests查询结果



官方解释

> When you make a request, Requests makes educated guesses about the encoding of the response based on the HTTP headers. The text encoding guessed by Requests is used when you access r.text. You can find out what encoding Requests is using, and change it, using the r.encoding property.Beautiful Soup uses a sub-library called Unicode, Dammit to detect a document’ s encoding and convert it to

> Unicode. The autodetected encoding is available as the .original_encoding attribute of the BeautifulSoup object.Unicode, Dammit guesses correctly most of the time, but sometimes it makes mistakes. Sometimes it guesses correctly, but only after a byte-by-byte search of the document that takes a very long time. If you happen to know a document’ s encoding ahead of time, you can avoid mistakes and delays by passing it to the BeautifulSoup constructor as from_encoding.

处理方法:

```
aqiurl = 'http://www.tianqihoubao.com/aqi/'
response = requests.get(url=aqiurl, headers=headers)
response.encoding = 'gbk'
```

参考资料:
BeautifulSoup与requests爬取网页中文转码问题
https://blog.csdn.net/he_and/article/details/78629675

什么是BOM头
https://zhuanlan.zhihu.com/p/28986236

彻底搞懂编码 GBK 和 UTF8
https://www.cnblogs.com/batsing/p/charset.html

查看网页编码方式的通用方法
https://blog.csdn.net/qq512028505/article/details/78035172

beautifulsoup库

> name参数

```
name  参数可以查找所有名字为  name  的tag,字符串对象会被自动忽略掉.
```

> 按CSS搜索

```
按照CSS类名搜索tag的功能非常实用,但标识CSS类名的关键字  class  在Python中是保留字,使用  class  做参数会导致语法错误.从Beautiful Soup的4.1.1版本开始,可以通过  class_  参数搜索有指定CSS类名的tag
```

pandas库

> read.csv()

读取csv文件内容

部分参数说明:

```
header  : int or list of ints, default ‘infer’
Row number(s) to use as the column names, and the start of the data. Default behavior is to infer the column names: if no names are passed the behavior is identical to  header=0  and column names are inferred from the first line of the file, if column names are passed explicitly then the behavior is identical to  header=None. Explicitly pass  header=0  to be able to replace existing names. The header can be a list of integers that specify row locations for a multi-index on the columns e.g. [0,1,3]. Intervening rows that are not specified will be skipped (e.g. 2 in this example is skipped). Note that this parameter ignores commented lines and empty lines ifskip_blank_lines=True, so header=0 denotes the first line of data rather than the first line of the file.

names  : array-like, default None
List of column names to use. If file contains no header row, then you should explicitly pass header=None. Duplicates in this list will cause a  UserWarning  to be issued.
```

> groupby()

```
        Group series using mapper (dict or key function, apply given function to group, return result as series) or by a series of columns.

        在分组group1、group2上应用size()、sum()、count()等统计函数，能分别统计分组数量、不同列的分组和、不同列的分组数量。

        agg()
        对于分组的某一列或者多个列，应用agg(func)可以对分组后的数据应用func函数。例如：用group1['data1'].agg('mean')对分组后的’data1’列求均值。当然也可以推广到同时作用于多个列和使用多个函数上。
```

参考资料:
https://blog.csdn.net/elecjack/article/details/50760736

http://pandas.pydata.org/pandas-docs/stable/generated/pandas.core.groupby.DataFrameGroupBy.agg.html

> reset_index()

```
    DataFrame.set_index(keys, drop=True, append=False, inplace=False, verify_integrity=False)  append添加新索引，drop为False，inplace为True时，索引将会还原为列
    For DataFrame with multi-level index, return new DataFrame with labeling information in the columns under the index names, defaulting to ‘level_0’, ‘level_1’, etc. if any are None. For a standard index, the index name will be used (if set), otherwise a default ‘index’ or ‘level_0’ (if ‘index’ is already taken) will be used.
```

参考资料：
https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.reset_index.html

> sort_index()

```
对 index 进行排序操作
```

参考资料：
https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.sort_index.html

可视化工具pycharts

示例：

> Line（折线/面积图）折线图是用折线将各个数据点标志连接起来的图表，用于展现数据的变化趋势。

Line.add() 方法签名

> add(name, x_axis, y_axis,
> is_symbol_show=True,
> is_smooth=False,
> is_stack=False,
> is_step=False,
> is_fill=False, **kwargs)

- name -> str
  图例名称
- x_axis -> list
  x 坐标轴数据
- y_axis -> list
  y 坐标轴数据
- is_symbol_show -> bool
  是否显示标记图形，默认为 True
- is_smooth -> bool
  是否平滑曲线显示，默认为 False
- is_stack -> bool
  数据堆叠，同个类目轴上系列配置相同的 stack 值可以堆叠放置。默认为 False
- is_step -> bool/str
  是否是阶梯线图。可以设置为 True 显示成阶梯线图。默认为 False
  也支持设置成'start', 'middle', 'end'分别配置在当前点，当前点与下个点的中间下个点拐弯。
- is_fill -> bool
  是否填充曲线所绘制面积，默认为 False

```
from pyecharts import Line
attr = ["衬衫", "羊毛衫", "雪纺衫", "裤子", "高跟鞋", "袜子"]
v1 = [5, 20, 36, 10, 10, 100]
v2 = [55, 60, 16, 20, 15, 80]
line = Line("折线图示例")
line.add("商家A", attr, v1, mark_point=["average"])
line.add("商家B", attr, v2, is_smooth=True, mark_line=["max", "average"])
line.render()
```

参考资料:
Python数据可视化工具pyecharts使用细则
https://www.jiqizhixin.com/articles/2018-08-16-6

pyecharts 官方文档
http://pyecharts.org/#/zh-cn/

汉字拼音转换工具（Python 版）

安装

```
$ pip install pypinyin
```

示例

```
>>> from pypinyin import pinyin, lazy_pinyin, Style
>>> pinyin('中心')
[['zhōng'], ['xīn']]
>>> pinyin('中心', heteronym=True)  # 启用多音字模式
[['zhōng', 'zhòng'], ['xīn']]
>>> pinyin('中心', style=Style.FIRST_LETTER)  # 设置拼音风格
[['z'], ['x']]
>>> pinyin('中心', style=Style.TONE2, heteronym=True)
[['zho1ng', 'zho4ng'], ['xi1n']]
>>> pinyin('中心', style=Style.BOPOMOFO)  # 注音风格
[['ㄓㄨㄥ'], ['ㄒㄧㄣ']]
>>> pinyin('中心', style=Style.CYRILLIC)  # 俄语字母风格
[['ч ж у н1'], ['с и н ь1']]
>>> lazy_pinyin('中心')  # 不考虑多音字的情况
['zhong', 'xin']
```

其他参考资料

python:pandas——read_csv方法
https://www.jianshu.com/p/9c12fb248ccc

C和python中%d %.2d %2d %02d的区别
https://blog.csdn.net/sesars/article/details/77448643

Python replace()方法
http://www.runoob.com/python/att-string-replace.html

python 字符串（str）和列表（list）的互相转换
https://blog.csdn.net/roytao2/article/details/53433373