题目：材料信息网站数据搜集与分析系统

### 完成人：王睿祺 杜彦璋 曹宏图 黄鑫翔 朱诚

### 学 号：SA19225376 SA19225135 SA19225100 SA19225187 SA19225516

### 任务工作量分配：20% 20% 20% 20% 20%

**本文档按照以下六个方面对所选开发课题进行介绍：**

**一、研究背景**

**二、设计目标**

**三、技术路线**

**四、关键原理**

**1.网页爬虫实现**

**2.数据解析与数据存储实现**

**3.数据可视化实现**

**4.数据分析与预测实现**

**五、结果展示**

**六、收获与感受**

### 一、设计背景

在信息爆炸的时代，如何从海量、多样化的数据中挖掘出有价值的信息，节约大量宝贵的时间和人力，变得越来越重要。在目前的企业管理系统中，员工往往需要人工对接多个网站窗口，多个网页资源来获取有用的新闻、数据等。这样反复操作，耗费大量的人力物力。因此我们设计信息搜集数据分析系统，旨在帮助企业更加便捷地获取目标信息，提高工作效率。

### 二、设计目标

实现爬虫爬取企业指定网站，从当中进行数据解析和数据清洗，筛选出有效信息并存储在数据库中，并设计和开发出面向企业定制的图形界面，给企业提供数据参考。

### 三、技术路线

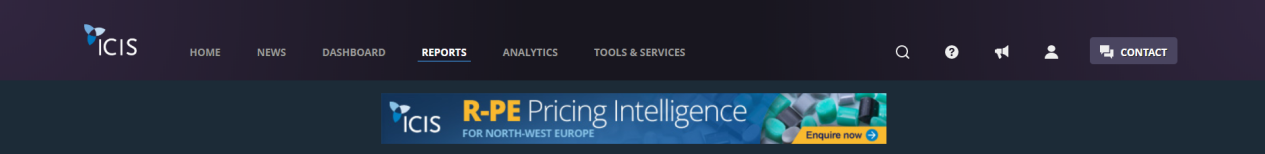
环境配置：Python3.6，MongoDB，Anaconda3，Keras

1. 首先，利用爬虫技术从目标网站：<http://www.pfchina.com.cn/>

和[https://www.icis.com](https://www.icis.com/explore/)获取网页数据。网站信息如下图所示：



[http://www.pfchina.com.cn](http://www.pfchina.com.cn/)网站介绍



[http://www.icis.com](http://www.pfchina.com.cn/)网站介绍

2. 然后，将爬取到的网页数据进行数据解析和数据清洗得到有效数据，并将数据存储到MongoDB数据库中。

3. 之后，将MongoDB数据库的数据进行整合，使用Matplotlib库把数据绘制成图表，使用PyQt5库制作用户交互图形界面，进行数据显示。

4. 最后，进行数据分析：对材料价格进行相关性分析，并使用Keras框架实现基于LSTM的材料价格预测模型。

### 四、关键原理

**1. 网页爬虫实现**

（1）技术介绍

实现一个爬虫首先要做的工作就是获取网页，这里就是获取网页的源代码。源代码里包含了网页的部分有用信息，因此需要把源代码获取下来，就可以从中提取想要的信息了。

实现获取网页源码，需要用到HTTP协议中的请求和响应，向网站的服务器发送一个请求，返回的响应体便是网页源代码。因此最关键的部分就是构造一个请求并发送给服务器，然后接收到响应并将其解析出来。

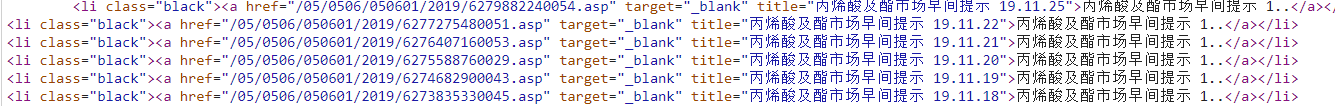
实现这个操作需要Python中的库，比如requests库等。我们用这些库来实现HTTP请求操作，请求和响应都可以用类库提供的数据结构来表示，得到响应之后只需要解析数据结构中的响应体部分，即可得到网页的源代码。

对于使用了Ajax技术的网站，需要对其进行特殊处理。Ajax技术是利用 JavaScript 在保证页面不被刷新、页面链接不改变的情况下与服务器交换数据并更新部分网页的技术。对于使用了Ajax技术的网站，我们有两种应对方式：1.使用分析Ajax接口的方式构造Ajax请求；2.模拟浏览器运行来实现网页加载。在本项目中，我们选择使用Selenium库模拟Chrome浏览器运行，实现网页抓取。

（2）设计思路

a. 在抓取网页之前，首先要实现登录。该网站的登录机制比较特别，使用了Cookie对登录用户数量进行限制。网站使用了两种Cookie：1.登录权限Cookie；2.登录状态Cookie。对于服务器来说，登录权限Cookie的数量有限，在用户数量达到网站限制值时，其余用户再登录就会失败；登录状态Cookie保存了用户的登录状态，一定时间后登录状态Cookie就会失效，需要用户重新登陆。我们针对这两种Cookie机制做了处理。

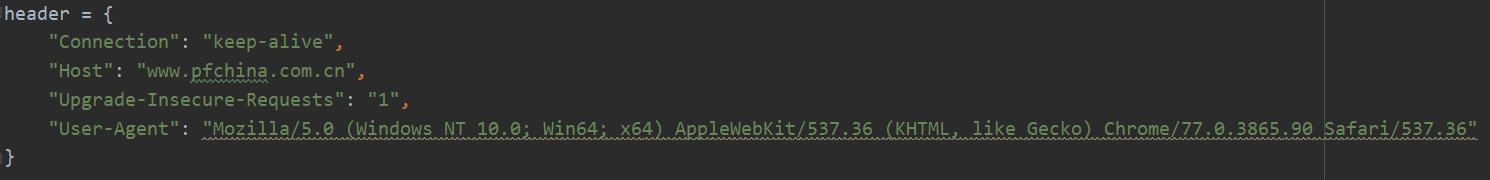
b. 然后实现对网页URL列表的抓取。在目录页面显示了每一个网页的标题，网页的标题对应它的URL。我们需要循环访问每一页目录页面，从目录页面提取网页的URL。在这里，需要对标题的关键字进行筛选，以得到我们需要的网页的URL。网页目录如下所示：



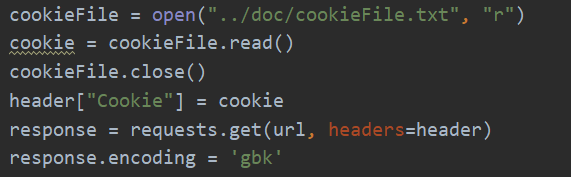
c. 循环遍历URL列表中的网页。在此循环中，实现单个网页内容的抓取，以及之后的数据解析和存储。

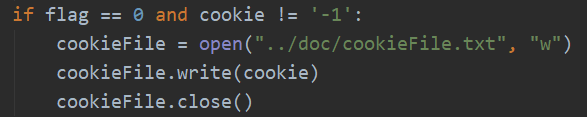
（3）具体实现

a. www.pfchina.com.cn网页抓取

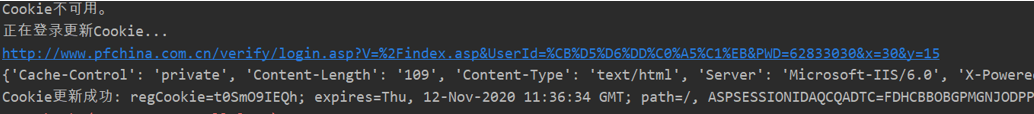
 根据网页的请求头在代码中构造请求头：

Cookies存储与请求：

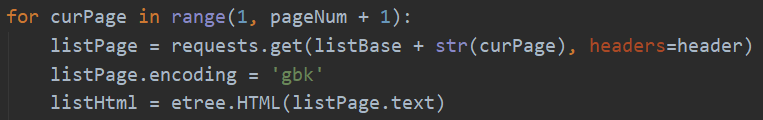




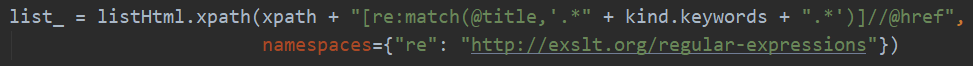
Cookies更新：



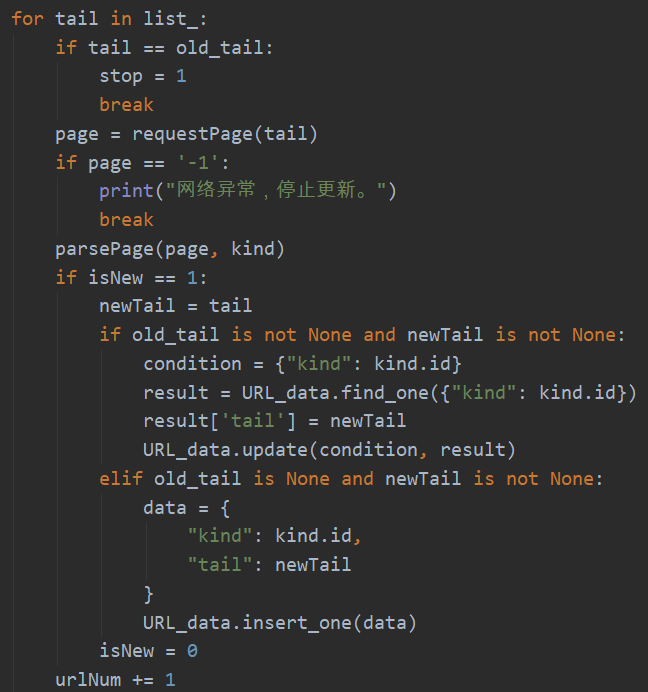
先循环抓取URL列表：



选择URL列表中标题包含关键字的URL：



内层循环遍历包含关键字的的URL，并判断与之前抓取的是否重复：

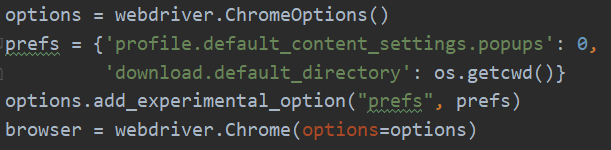


页面抓取过程中的网络延迟处理：

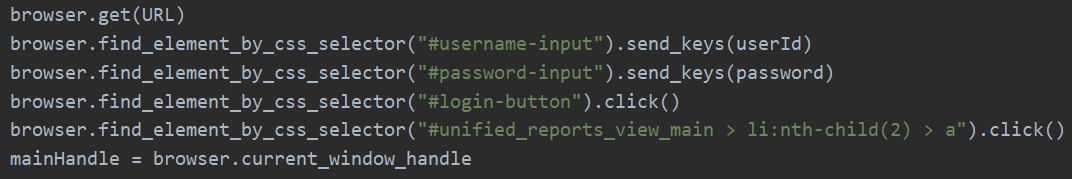


b. www.icis.com PDF下载

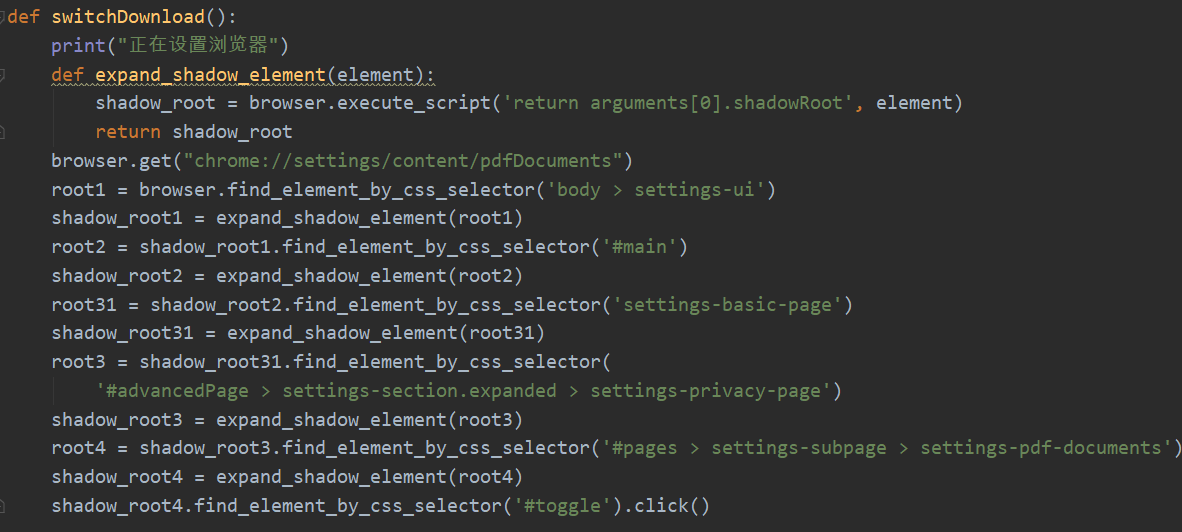
Selenium实现登录，设置webdriver：



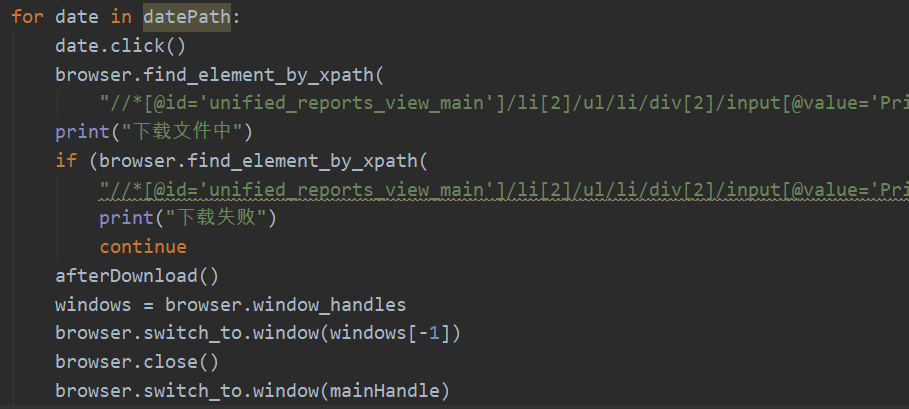
使用Selenium访问网站并登录：



自动设置浏览器以便下载pdf文件。由于html文件中包含Shadow root，需要对Shadow root进行定位：



下载指定pdf文档，每下载一个pdf，解析指定数据并存入数据库后，删除该pdf文档，然后下载下一个pdf文档：



**2. 数据解析与数据存储实现**

（1）技术介绍

对于网页的节点来说，它可以定义id、class或其他属性；而且节点之间还有层次关系，所以对爬虫爬取的网站数据，我们采用XPath技术来提取网站上的某个节点，然后调用相应方法来获取它的正文内容或者属性，从而实现提取网站中我们想要的任意数据；同时把提取到的数据保存到MongoDB数据库中。

XPath的选择功能十分强大，它提供了非常简洁明了的路径选择表达式；另外还提供了大量的内建函数，用于字符串、数值、时间的匹配以及节点、序列的处理等。几乎所有我们想要定位的节点，都可以用XPath来选择。

对于爬虫的数据存储来说，一条数据可能存在某些字段提取失败而缺失的情况，而且数据可能随时调整；另外数据之间可能还存在嵌套关系，如果使用关系型数据库存储，一是需要提前建表，二是如果数据存在嵌套关系的话，需要进行序列化操作才可以存储，这非常不方便，如果使用非关系型数据库就可以避免一些麻烦，更加简单高效；MongoDB就属于非关系型数据库的一种，其内容存储形式类似JSON对象，它的字段值可以包含其他文档、数组及文档数组，同时MongoDB数据读写性能好，内置功能丰富，非常适合存储从<http://www.pfchina.com.cn/>和[https://www.icis.com](https://www.icis.com/explore/)网站爬取到的数据。

（2）设计思路

a. 首先我们要解析<http://www.pfchina.com.cn/>中各种塑膜产品在每天在网站上发布的日评信息等，在日评信息中需要获取的有效信息是当天该塑膜产品在各个地区的收盘价，以及当天对该塑膜产品后期价格走势预测，具体解析网站数据如下图所示：



b. 然后进一步了解网站结构，确定XPath需要定位的节点，确定了定位点有以下3处：

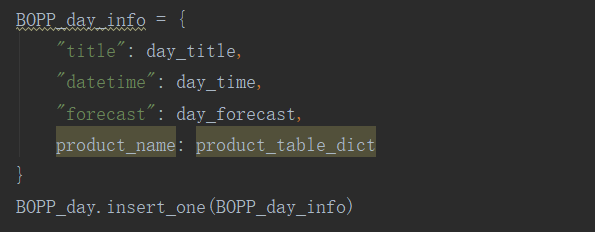
1> 定位塑膜产品日评标题处解析得到日评标题文本信息和发布时间文本信息。

2> 定位塑膜产品各地区收盘价格表Table处解析得到价格文本信息，通过for循环遍历将价格信息和对应的地区信息完整的保存到Python的字典中。

3> 定位塑膜产品走势分析预测处解析得到走势分析预测文本信息。

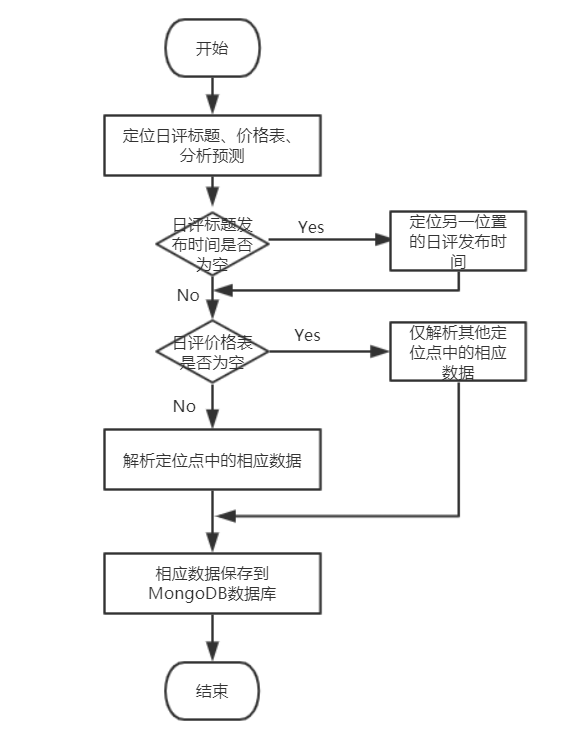
c. 之后测试XPath定位到的节点是否能够在该塑膜产品在每天日评信息网页上都能成功定位并解析到数据，发现以上三部分文本内容在以往的日评信息网页中都可能存在数据丢失、数据变化的情况，在此进行简单的数据清洗工作：对于存在数据丢失的日评信息网页进行非空验证，只去解析未丢失的数据；对于数据内容变化较大的日评信息网页，由于与企业需求不一致将日评信息网页数据做丢弃处理。

d. 最后把解析到的有效数据保存到MongoDB数据库中。下图是向塑膜产品BOPP数据库中插入一条日评信息记录。



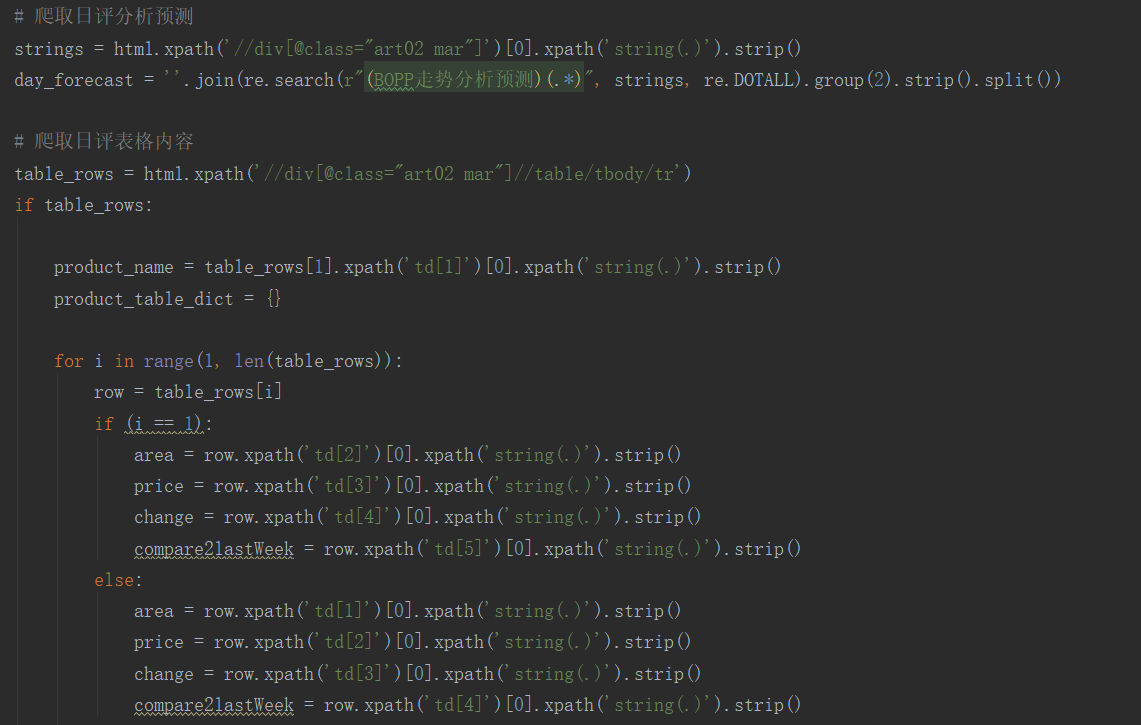
（3）具体实现

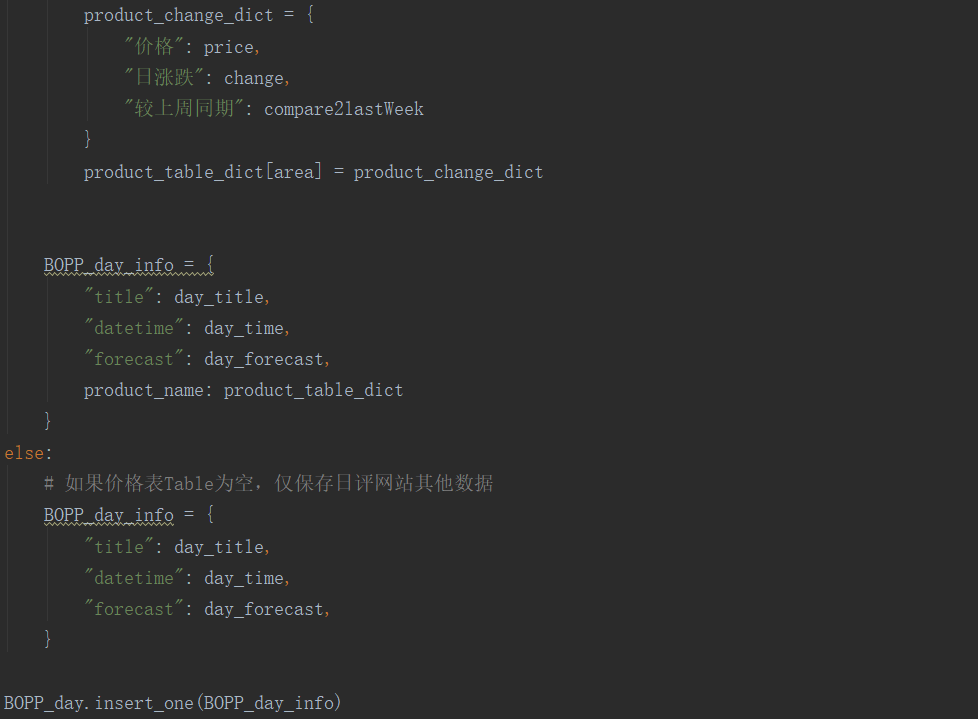
解析数据流程图如下图所示，根据流程图的设计思路编写源码：



同时给出从网站上爬取到的BOPP塑膜产品信息数据解析源码：







**3. 数据可视化实现**

（1）技术介绍

数据可视化部分采用PyQt5库进行GUI程序的开发，图表的绘制采用Matplotlib库整合数据进行绘制。

PyQt5是一套Python绑定Digia QT5应用的框架，它可以用于Python2和3，Qt库是最强大的GUI库之一，pyqt5做为Python的一个模块，它有620多个类和6000个函数和方法。

本课题使用了pyqt5的类别分为以下几个模块：QtCore:包含了核心的非GUI功能。此模块用于处理时间、文件和目录、各种数据类型、流、URL、MIME类型、线程或进程；QtGui包含类窗口系统集成、事件处理、二维图形、基本成像、字体和文本；Qtwidgets模块包含创造经典桌面风格的用户界面提供了一套UI元素的类。

Matplotlib是建立在Numpy数组基础上的多平台数据可视化程序库，Matplotlib具有良好的操作系统兼容性和图形显示底层接口兼容性，Matplotlib支持即使在图形显示接口与输出格式，这使得用户无论在哪种操作系统上面都可以输出自己想要的图形格式。

（2）设计思路

a. 首先确定在图形可视化界面需要显示的内容：

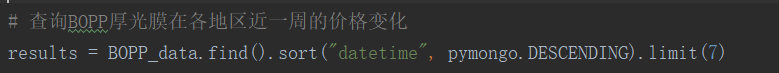
1> 根据企业的需求在图形界面中一共需要展示七个模块的数据，模块之间要能够实现切换效果，因此图形界面要设置菜单栏控件；

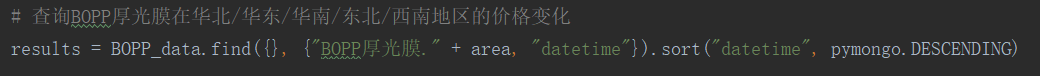
2> 每个模块中需要显示塑膜产品近期价格的变化，既要有直观的变动体现也要具体的数字体现，因此图形界面的每个模块中要设置一个表格来存放价格数据，也要设置一张折线图来体现近期价格波动；

3> 每个模块中需要显示塑膜产品长期价格的变化，因此图形界面的每个模块中要设置另一张折线图来体现该产品长期的价格波动；

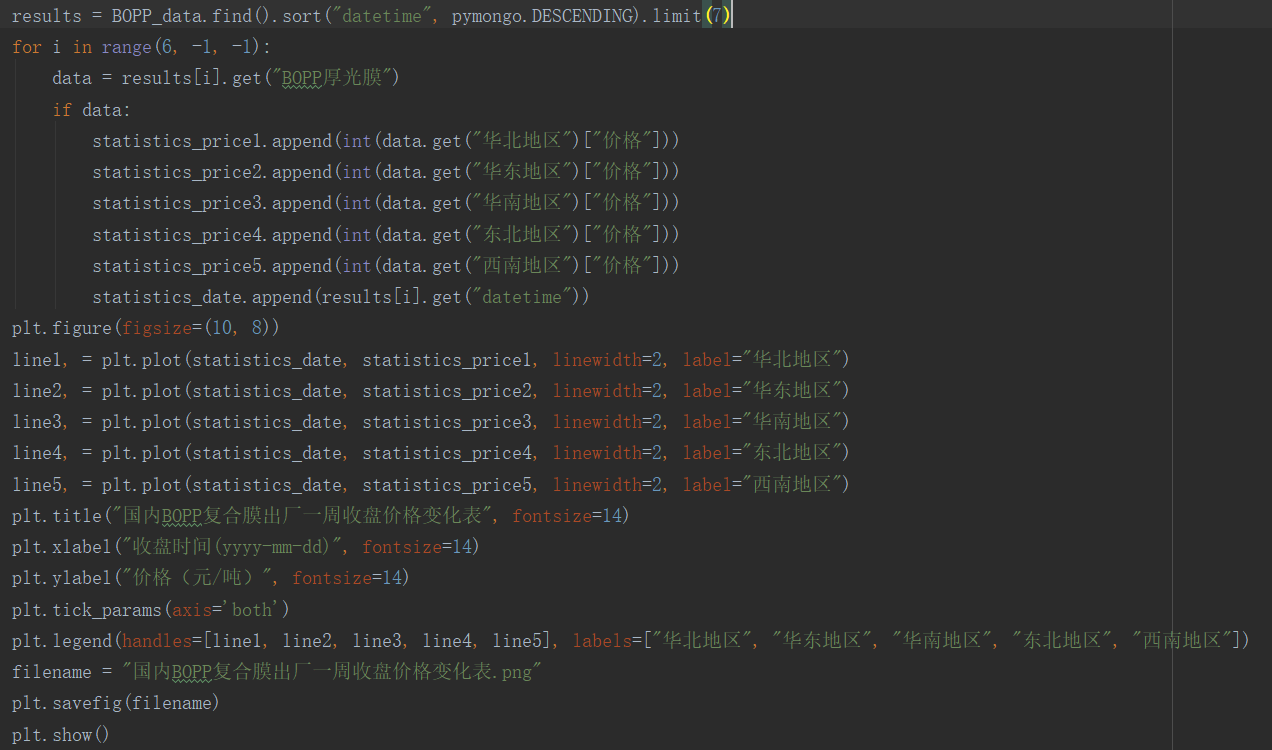
4> 每个模块中需要显示塑膜产品当天日评中对价格走势分析预测评论，因此图形界面的每个模块中要设置一个文本域来显示当天日评的走势分析预测评论；

b. 然后根据上述需要显示的内容，向MongoDB数据库发起相应查询请求获取相应的数据，就需要编写数据库查询语句，比如绘制折线图就需要按照日评的发布时间排序向MongoDB查询数据，部分查询语句如下图所示：



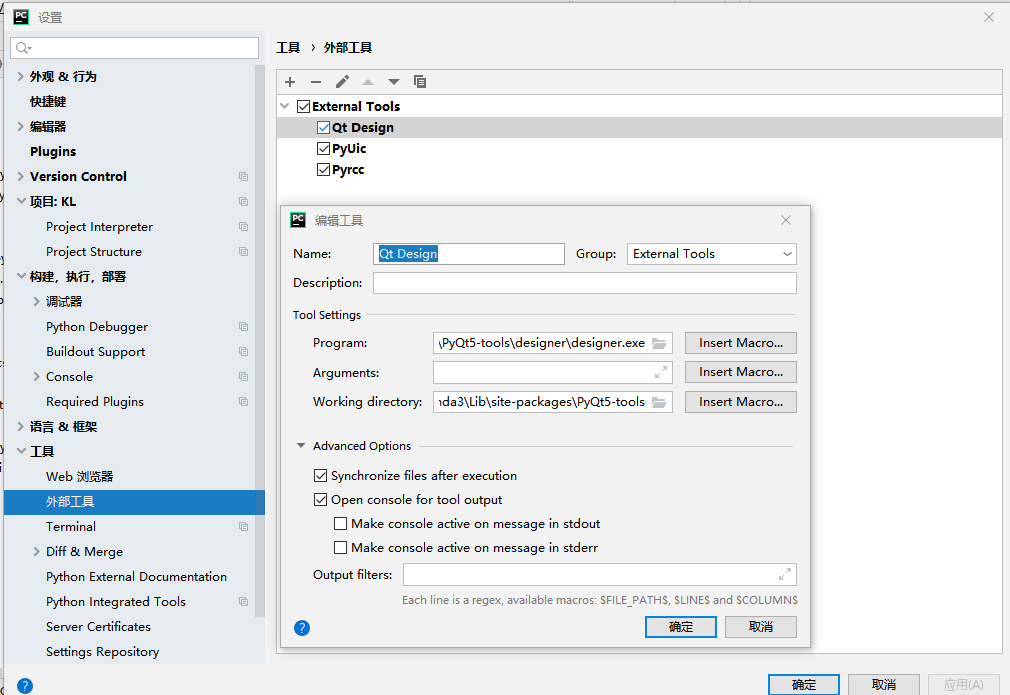


c. 之后采用Matplotlib把查询到的数据绘制成折线图，同时优化折线图的细节显示：设置坐标轴标题、设置坐标轴区间大小、设置折线图标题、给折线图折线添加Label、将图片设置文件名保存等，某塑膜产品绘制折线图关键代码如下图所示：



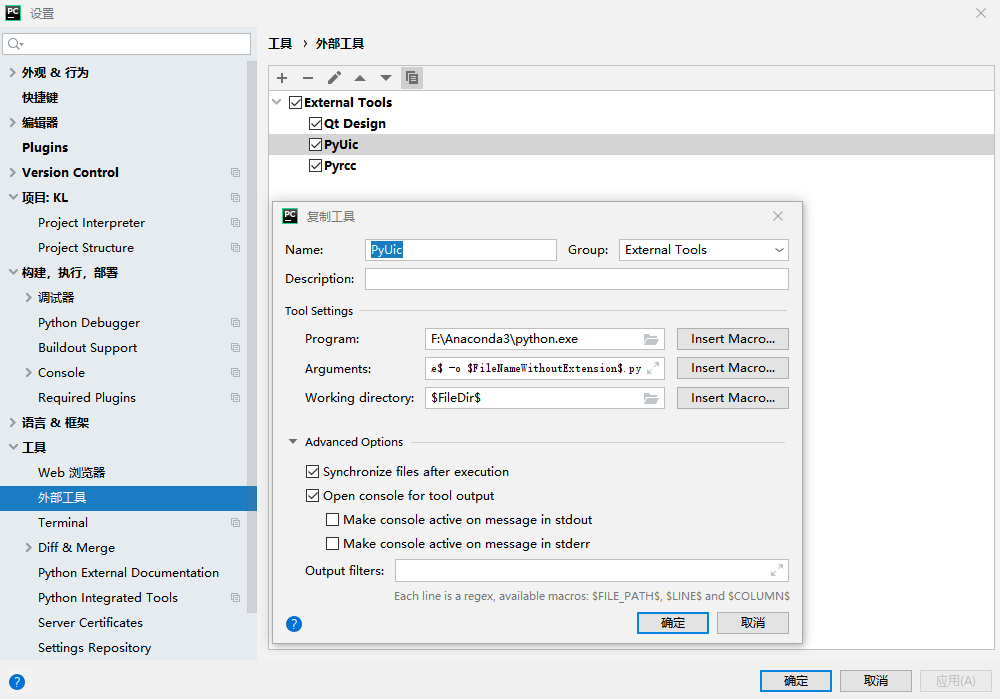
d. 在PyCharm下使用QtDesigner开发图形界面程序：

打开PyCharm，打开File—>Settings—>External Tools，点击加号来添加自己的工具，做如下配置：

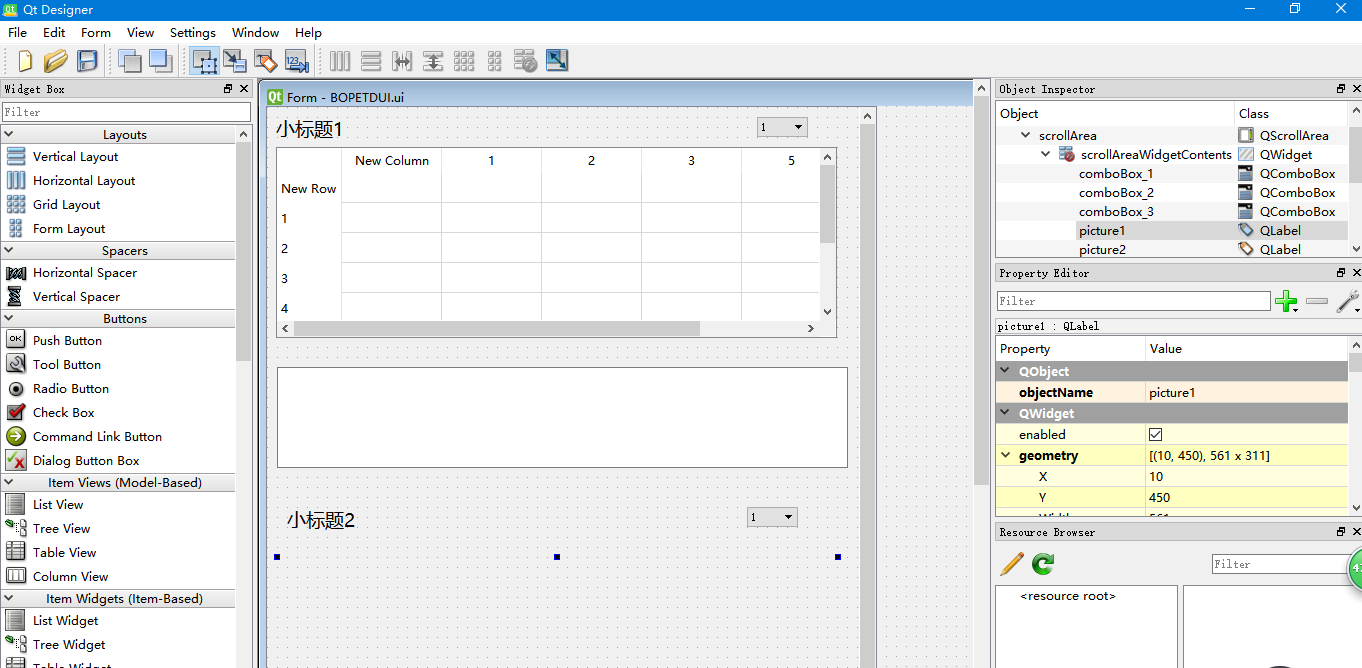


完成后点“OK”，现在QtDesigner就配置好了，就可以直接在PyCharm中打开它来画界面了，但是要在PyCharm中把界面的.ui文件转换为.py文件还需要后面的配置。

同样在External Tools中点击加号来配置pyuic：



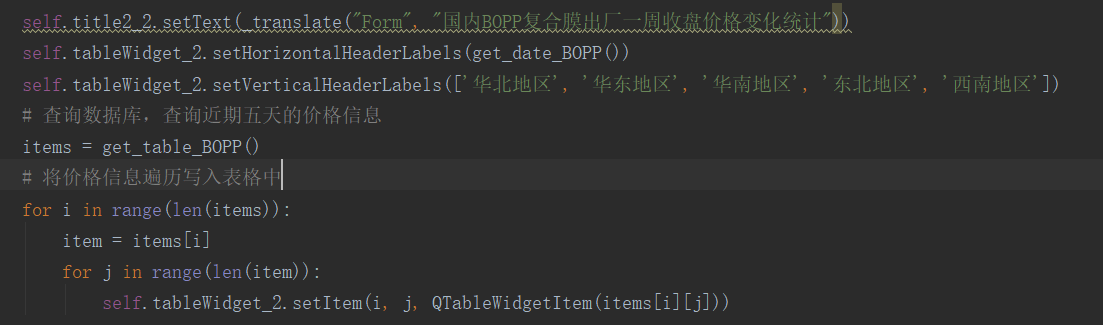
至此就全部配置好了，接下来是test，回到PyCharm，Tools—>Qt—>QtDesigner，点击即可打开designer：



根据QtDesigner左边的控件菜单栏，选用先前确定的图形界面需要的控件，设置合适大小、高度和位置，创建一个简单的图形界面，创建完成之后回到PyCharm，可以看到工程目录下已经产生了\*.ui，右键它，External Tools PyUic点击生成\*.py文件

e. QtDesigner开发图形界面程序中的数据都是静态的，但是在各个模块中显示的数据都是动态的，因此需要修改图形界面控件的部分代码，添加相应的显示动态数据的方法

1> 修改图形界面中表格代码，能够调用查询数据库语句，将查询到的近期五天的价格信息动态添加到图形界面中的表格里，关键代码如下图所示：



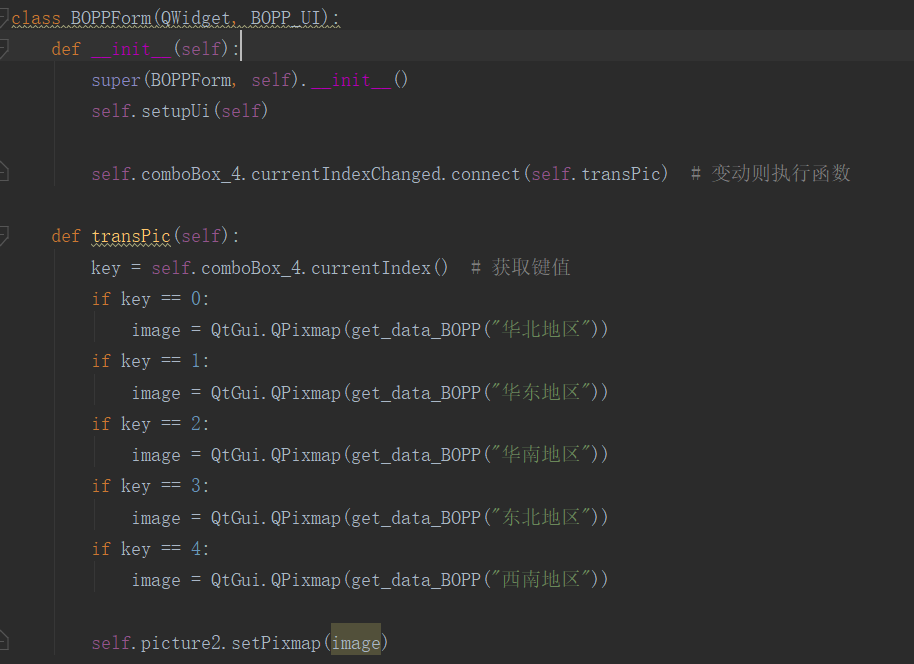
2> 修改图形界面中文本域代码，能够调用查询数据库，将查询到的当天的价格走势分析预测动态添加到图形界面中的文本域中，关键代码如下图所示：



3> 修改图形界面中图片控件代码，能够调用Matplotlib绘制折线图方法，将近期价格走势折线图，或者长期价格走势折线图绘制并保存系统中，图片控件调用该方法返回的图片文件名，将折线图显示出来，关键代码如下图所示：

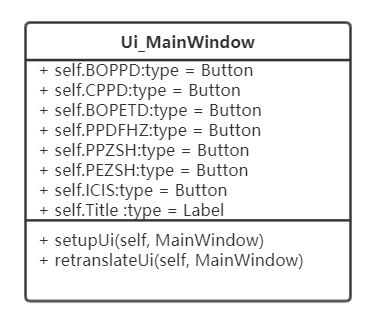


4> 部分塑膜产品可能存在多地区或者多产品类型的价格变动，将全部数据绘制在一张折线图里会导致数据过多影响显示，因此将多地区或者多产品类型的价格变动绘制成多张折线图，折线图的交替显示通过选择图形界面设置的下拉列表中信息实现，关键代码如下图所示：



（3）具体实现

根据上述开发思路设计出如下图所示的数据可视化主要类图，并根据类图编写源码：



UI\_MainWindow类图分析：UI\_MainWindow类是用来初始化图形界面，

其中属性有：

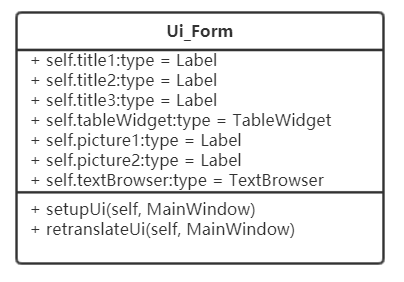
七个按钮Button：作为菜单栏中的Button，实现切换界面；

标签Label：作为图形界面标题；

其中方法有：

setupUI方法用来设置图形界面中各个控件具体细节：字体大小、控件位置等；

retranslateUI方法用来设置图形界面中各个控件具体数据：各个属性为Button的各个按钮名称，属性为Label的标题内容等。



UI\_Form类图分析：UI\_Form类是用来初始化图形界面中对应七个各类塑膜产品界面模块

其中属性有：

五个属性为标签Label：其中title1、title2、title3作为塑膜产品界面标题，指示产品的价格变化表、近期价格变化折线图以及长期价格变化折线图；其中picture1、picture2用来展示产品的近期价格变化折线图以及长期价格变化折线图的图片；

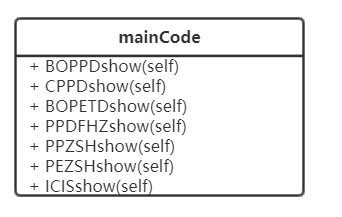
属性为TableWidget：用来展示产品的近期价格变化表，其中TableWidget中的各个item就是该产品的价格具体数据；

属性为textBrowser：用来展示产品的当天日评做出的价格分析预测内容。

其中方法有：

setupUI方法设置塑膜产品界面模块中各个控件具体细节：字体大小、控件位置等；

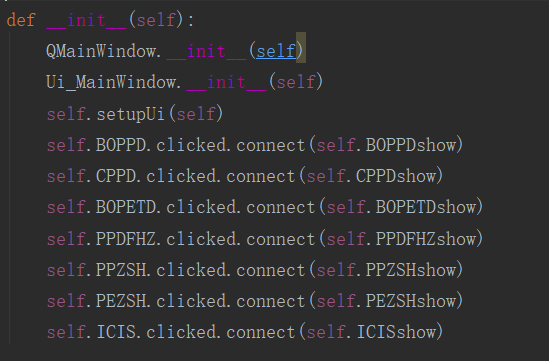
retranslateUI方法设置塑膜产品界面模块中各个控件具体数据：设置具体标题的内容，调用查询数据库方法动态设置价格分析预测内容和近期价格变化表内容，调用Matplotlib生成折线图方法动态设置近期价格变化折线图以及长期价格变化折线图。



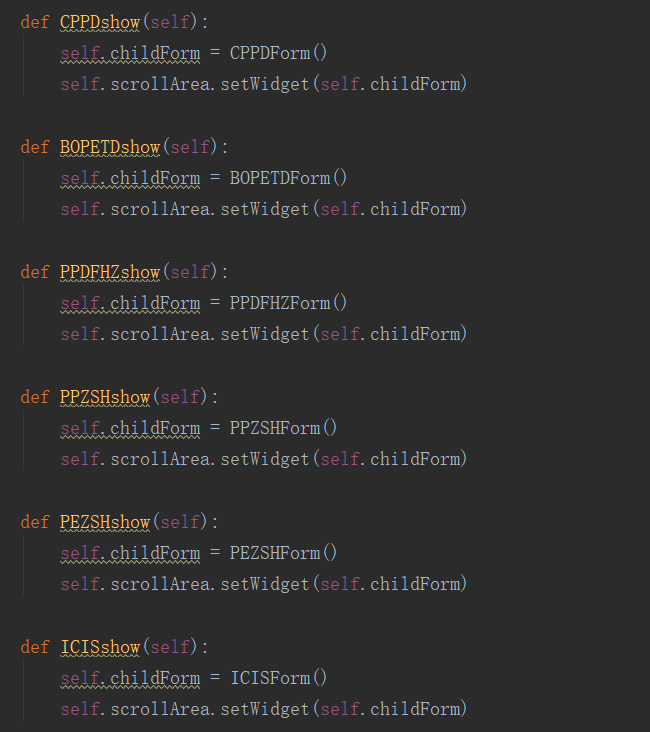
mainCode类图分析：mainCode类是用来初始化图形界面，并监听菜单栏中对应七个各类塑膜产品界面模块的按钮，当按钮被点击就触发相应事件，使图形界面中切换塑膜产品界面。

其中方法有：

\_\_init\_\_(self)：mainCode类的初始化方法，用于初始化图形界面，给各个菜单栏按钮绑定点击事件



BOPPDshow(self)、 CPPDshow(self) 、BOPETDshow(self)、PPDFHZshow(self)、PPZSHshow(self)、PEZSHshow(self)、ICISshow(self)：对应菜单栏按钮被点击后调用的方法，转去初始化该塑膜产品的模块界面，并在图形界面上切换显示



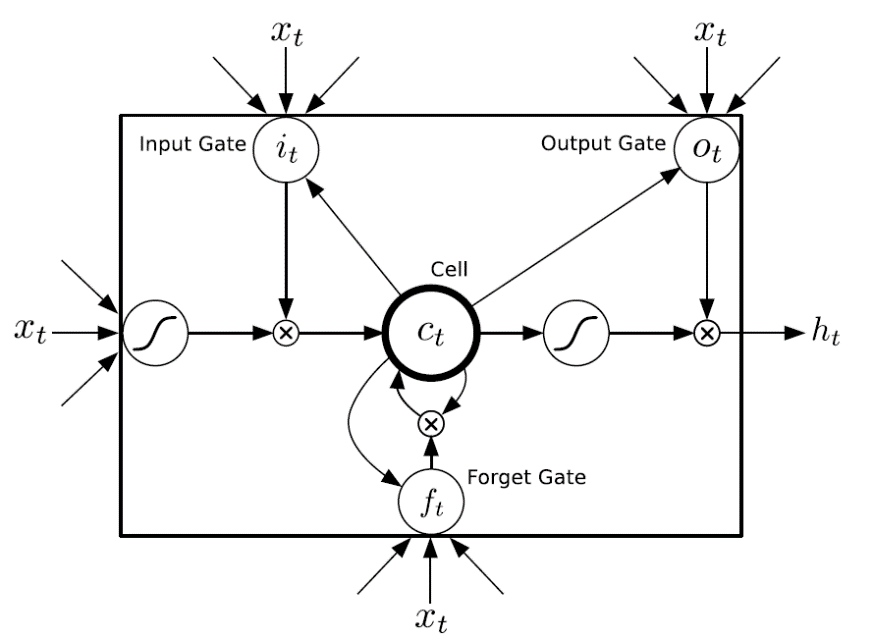
**4.数据分析与预测实现**

4.1基于LSTM的时序数据预测

（1）技术介绍

使用Keras框架进行实现。构建两层LSTM网络，输出层为全连接网络。

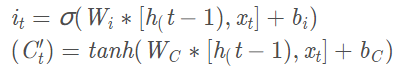
LSTM每个神经元由遗忘门，输入门和输出门组成，有效解决RNN梯度消失和梯度爆炸问题。



遗忘门会读取h\_(t-1)和x\_t ，输出一个在 0到 1之间的数值给每个在细胞状态C\_(t-1)中的数字。1 表示“完全保留”，0 表示“完全舍弃”。



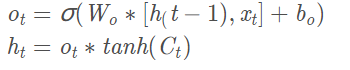
输入门的的 sigmoid 层决定哪些信息需要更新；一个 tanh 层生成一个向量，也就是备选的用来更新的内容。把这两部分联合起来，对 cell 的状态进行一个更新。



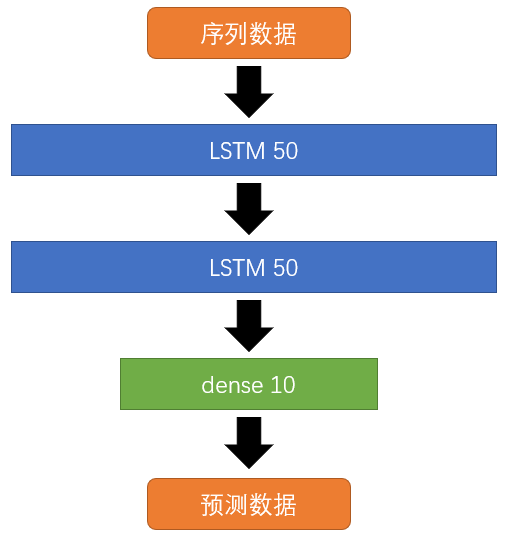
更新CELL状态：



输出门额 sigmoid 层来确定细胞状态的哪个部分将输出出去。接着，我们把细胞状态通过tanh进行处理并将它和 sigmoid门的输出相乘，最终我们仅仅会输出我们确定输出的那部分。



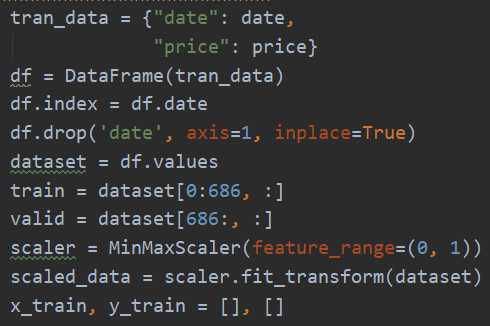
我们的模型使用50个这样的神经元。



（2）具体实现

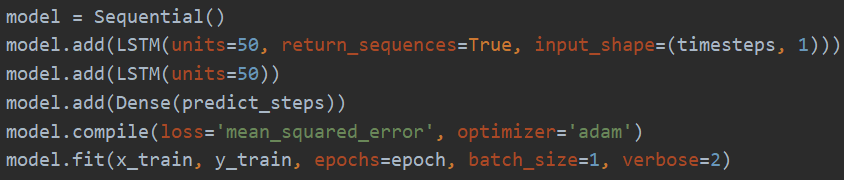
a. 数据预处理

进行特征缩放，将数据压缩至[0,1]区间，加快模型收敛。



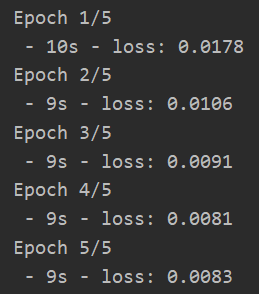
b. 模型构建

损失函数使用均方差(MSE)，优化器使用adam方法。



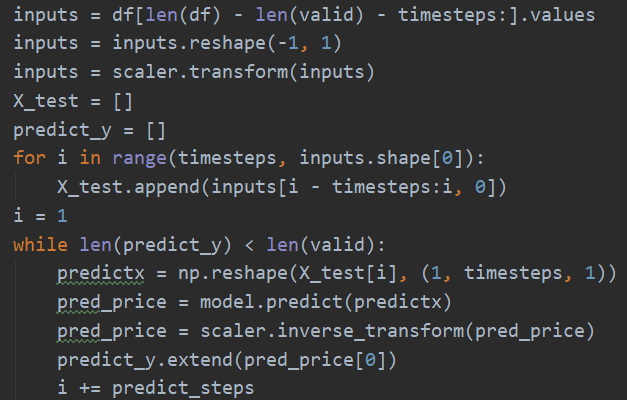
c. 训练模型

将数据划分为训练集与测试集，使用训练集进行训练。



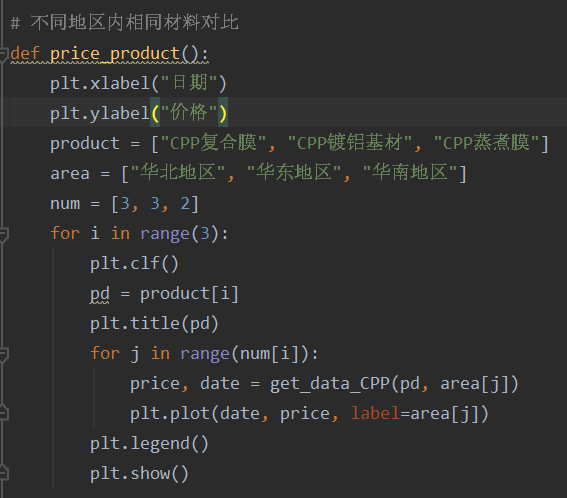
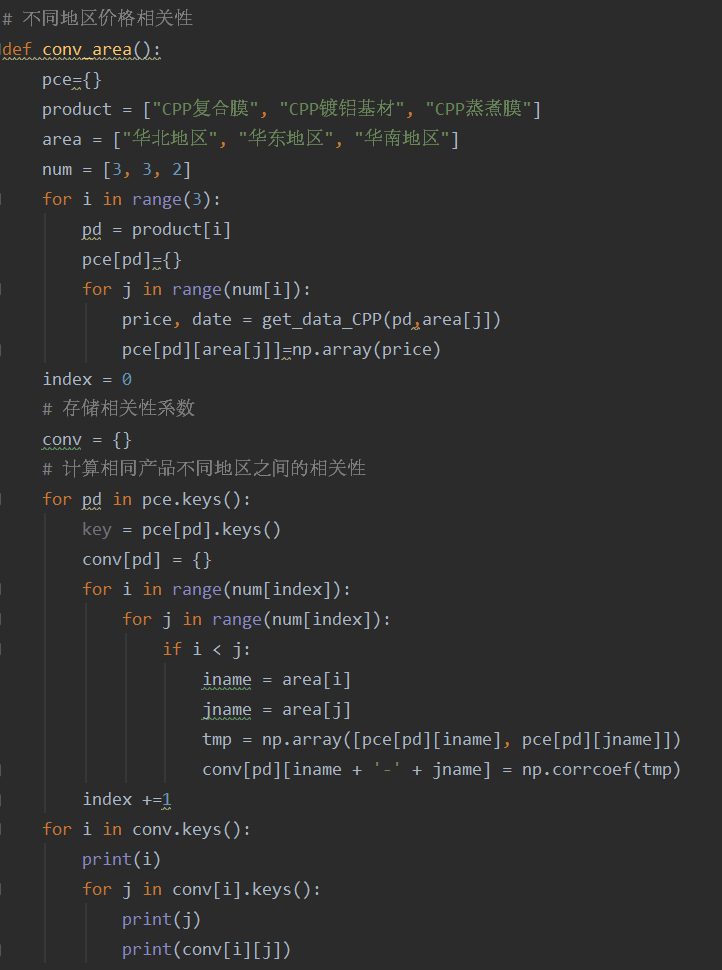
d. 测试模型

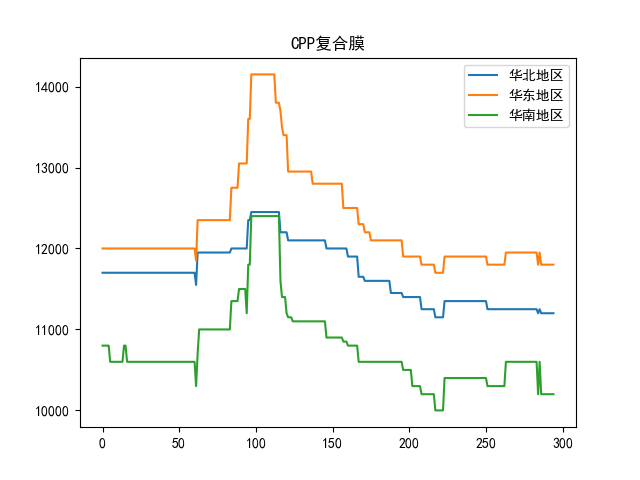
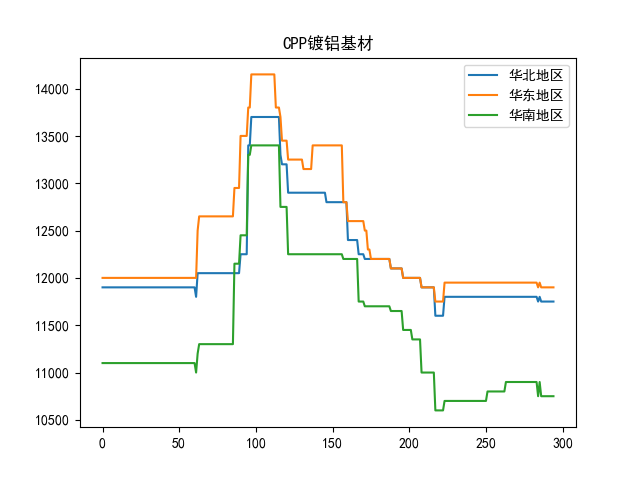
使用测试集进行预测。

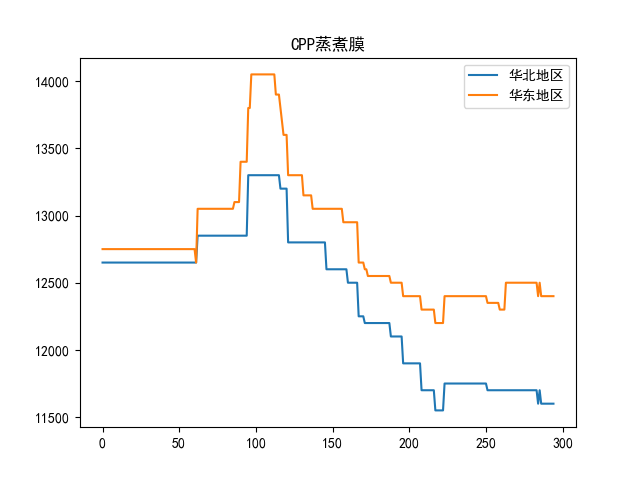


4.2 典型数据相关分析

此处选择数据相对丰富的CPP材料信息进行简单分析，对于部分材料少部分数据的缺失（总量不到十条，且连续缺失不超过四条），此处使用相邻值代替。

1. 查看同种材料在不同地区的价格走势
2. 代码
3. 价格走势

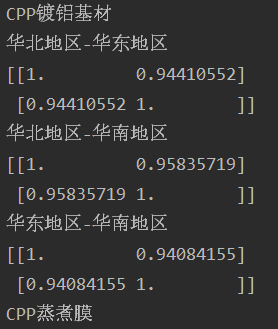
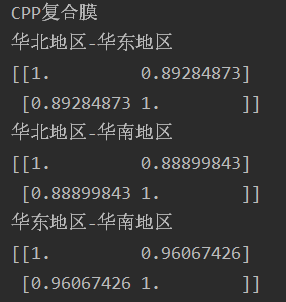


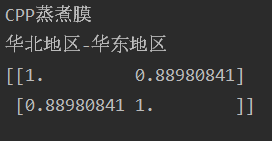


由上图可知，在价格分布上，同种材料在华东地区价格最高，其次是华北地区，华南地区材料价格最低。

对于同一种材料，不同地区的价格走势基本一致，说明材料价格的波动是在全国范围内的波动。

1. 相关系数



由上图可知，CPP镀铝基材三个地区价格之间的相关性系数都在0.94以上，区域间价格的相互影响极大。

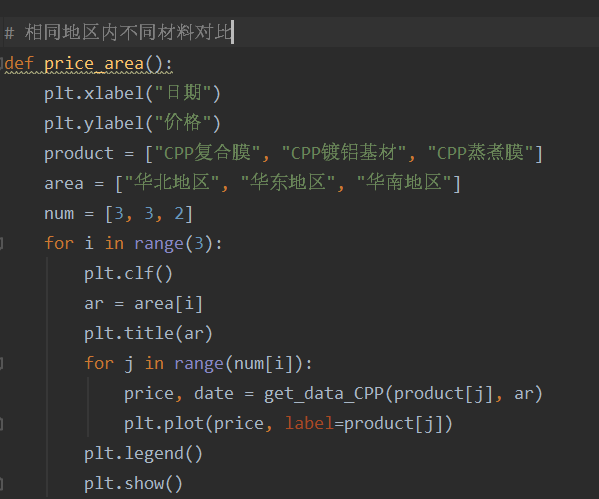
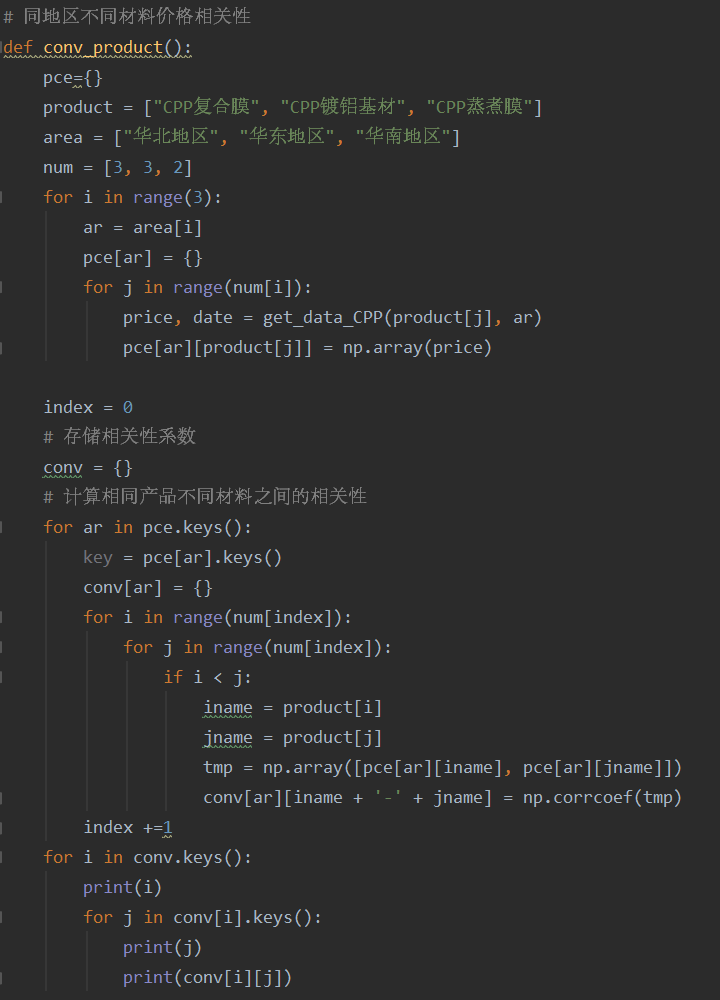
CPP复合膜三个地区价格之间的相关性系数也都较高，尤其是华东地区与华南地区之间。可见对于CPP复合膜而言，区域间价格的相互影响总体而言较大，且华东地区与华南地区之间的相互影响最为明显。

CPP蒸煮膜两个地区间相关性系数接近0.89，相互影响极大。

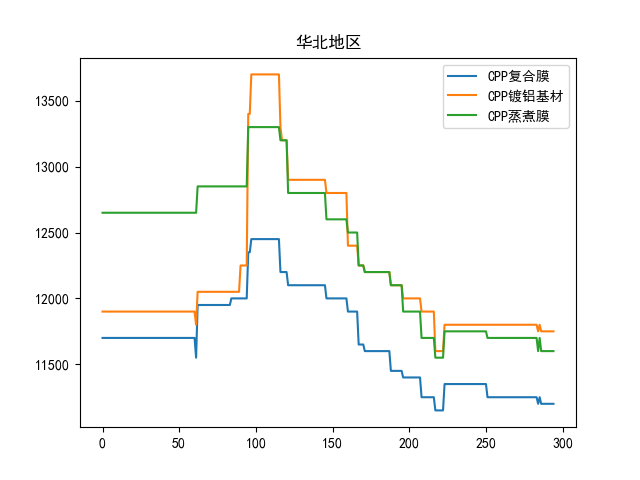
综上可知，CPP镀铝基材区域间价格影响最为明显，CPP复合膜与CPP蒸煮膜相对次之。三个区域中，华东地区与华南地区之间的相互影响最为明显，对于CPP复合膜尤甚。

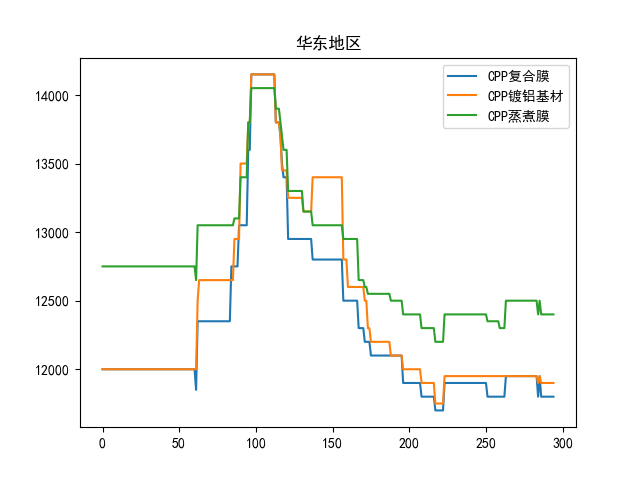
（2） 查看不同材料在相同地区的价格走势

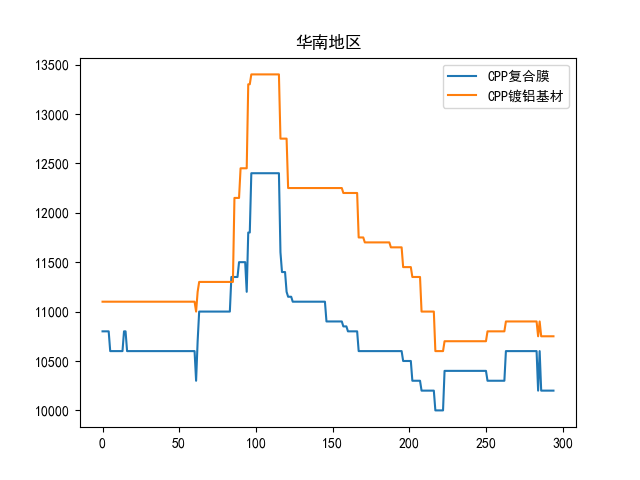
* 1. 代码

* 1. 价格走势







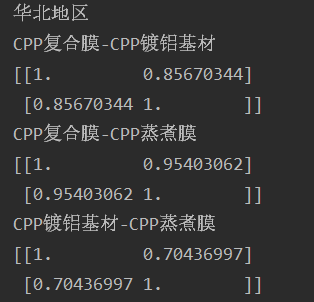
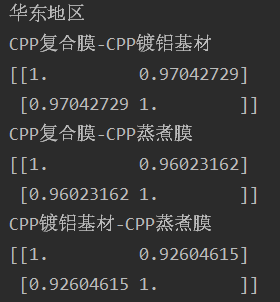
由上图可知，华北地区CPP蒸煮膜与CPP复合膜走势基本一致，价格差异相对平稳。而CPP镀铝基材前期价格与CPP复合膜较为接近，中期涨幅极大，甚至超过了CPP蒸煮膜价格，且后期价格也与CPP蒸煮膜几乎持平，总体而言波动最为明显。

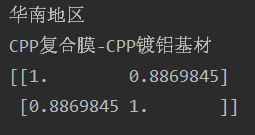
华东地区三种材料涨幅都较大，且高峰期前后三种材料价格几乎持平，后期CPP复合膜与CPP镀铝基材价格走向低谷，与CPP蒸煮膜价格逐渐拉开差距。

华南地区的CPP复合膜涨幅波动介于华北与华东之间。

总体比较而言，华北地区CPP蒸煮膜与CPP复合膜价格波动幅度远小于CPP镀铝基材；华东地区CPP蒸煮膜波动幅度远小于CPP复合膜与CPP镀铝基材；华南地区两种材料波动相对一致，但CPP镀铝基材价格回落速率较慢。

1. 相关性系数



由上图可知，华北地区CPP镀铝基材与其他材料相关性分别略高于0.7和0.85，远小于CPP复合膜与CPP蒸煮膜之间0.95的相关性系数，可见在华北地区CPP复合膜与CPP蒸煮膜之间价格关联极大，而CPP镀铝基材与CPP蒸煮膜关联相对小很多，CPP镀铝基材与CPP镀铝基材价格的相关性介于前面两种之间。

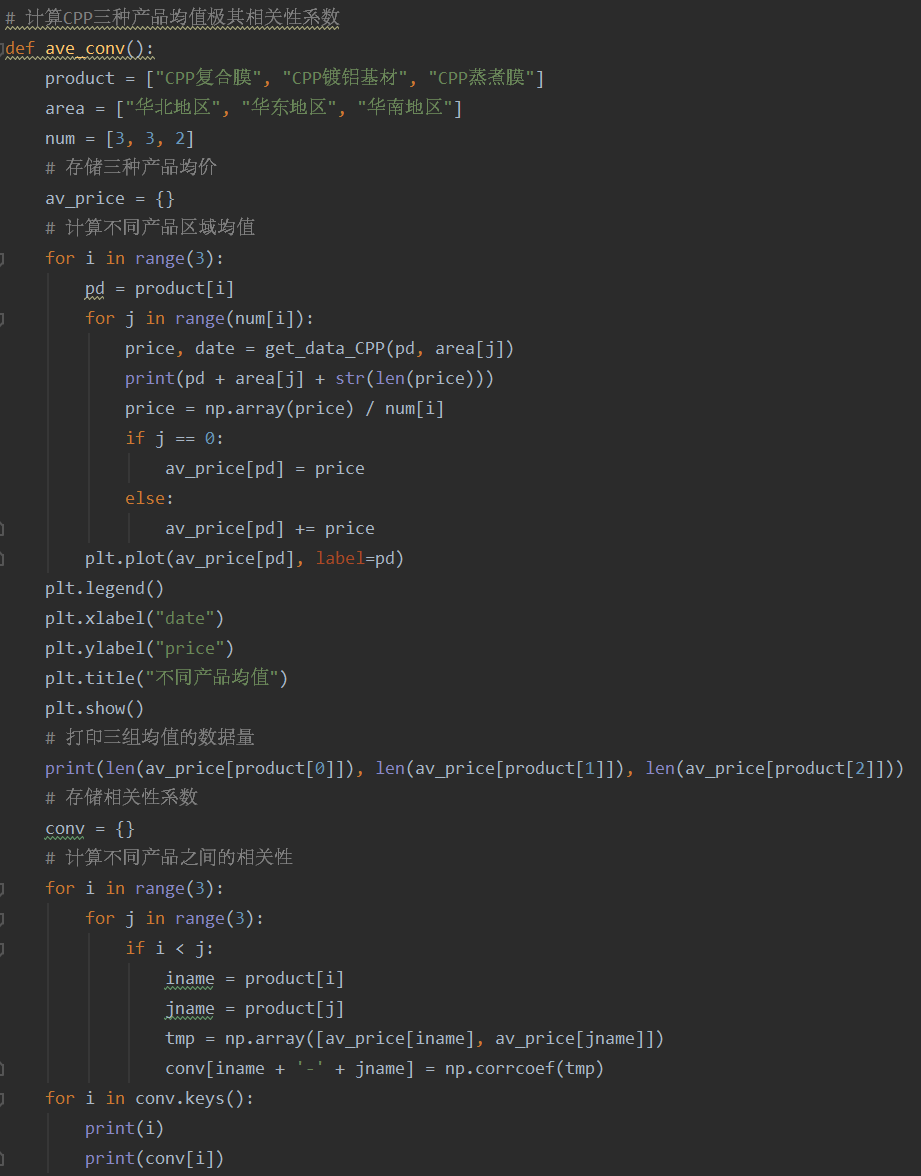
而华东地区三种材料互相之间的先关系数都在0.92以上，关联性都极大，尤其是CPP复合膜，其与另外两种材料价格相关系数都在0.96以上。

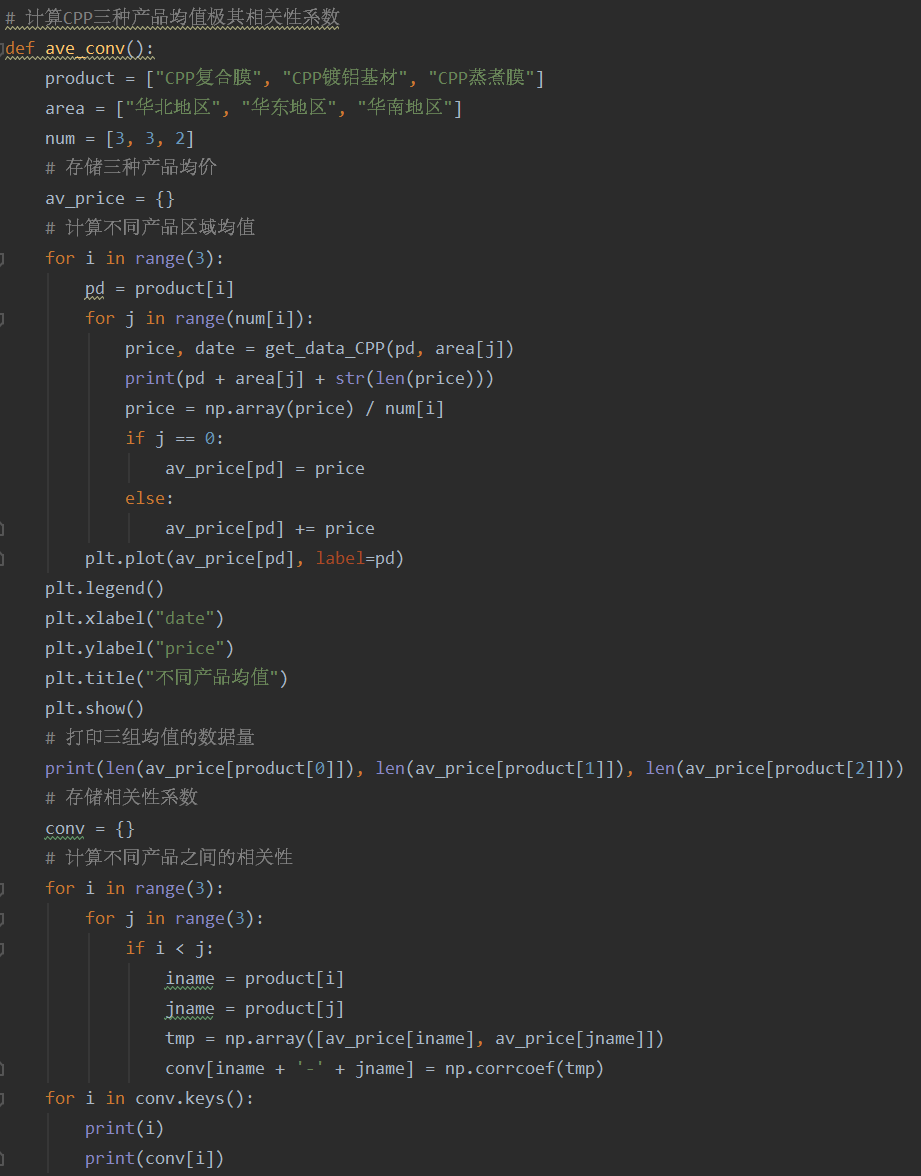
华南地区CPP复合膜与CPP镀铝基材相关性系数在0.88以上，略高于华北地区，但远小于华东地区。

综上可知，相较于另外两个地区，华东地区材料价格关联性要大很多。

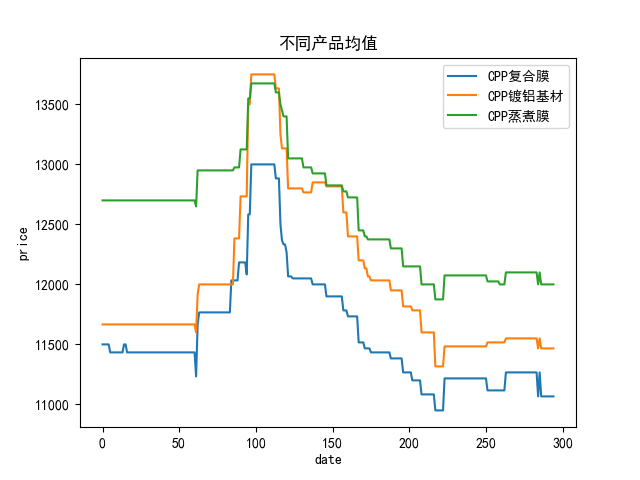
（3）比较不同材料的价格均值

1. 代码

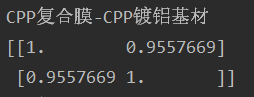
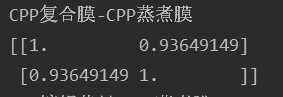
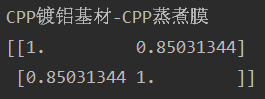




1. 价格走势



由上图可知三种材料在近期的价格走势基本一致，说明影响材料价格的因素是全局性的因素，其对于不同地区，不同种类的材料价格有着同的影响效果。

1. 相关系数

由上图可知，CPP复合膜与CPP镀铝基材价格之间的相关性系数最高，达到了0.95以上；CPP复合膜与CPP蒸煮膜之间的相关性系数也较高，在0.93以上；CPP镀铝基材与CPP蒸煮膜相关性系数最低，略高于0.85。

可知，三种材料价格之间的相关性都较高，其价格波动有相当大的内在联系。

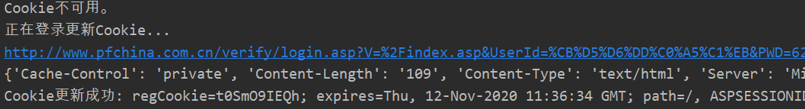
其中CPP复合膜与另外二者的相关系数都超过的0.93，而CPP蒸煮膜与CP镀铝基材之间的相关系数在0.85左右，说明CPP复合膜的价格波动对于其他材料的价格走势都有极大的参考意义，而另外二者的价格波动对于其他材料的价格走势具有比较大的参考意义。

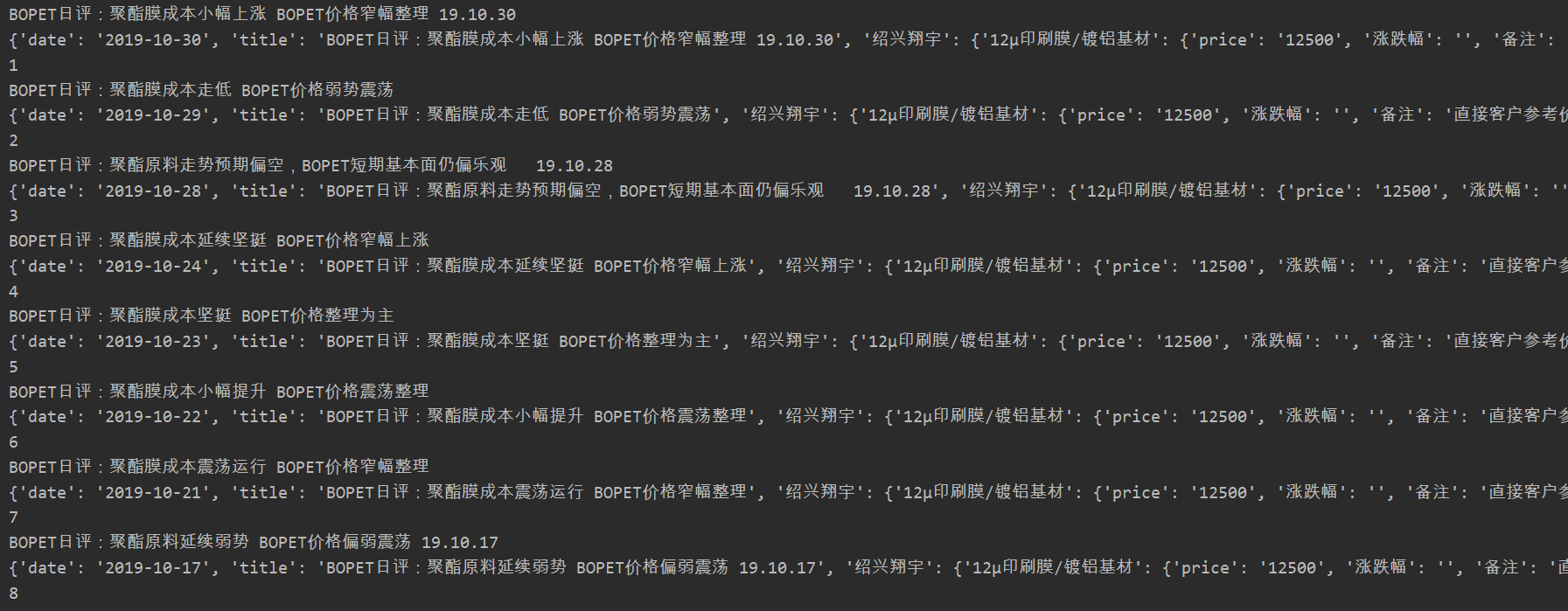
### 五、结果展示

（1）爬取结果展示

a. www.pfchina.com.cn

下图为Cookie自动更新的过程：



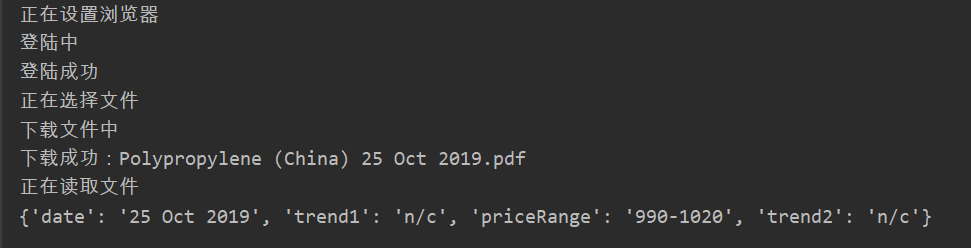
 以BOPP日评栏目作为示例，下图为实现页面抓取以及解析数据，并保存在字典结构中的结果：

存放至数据库：

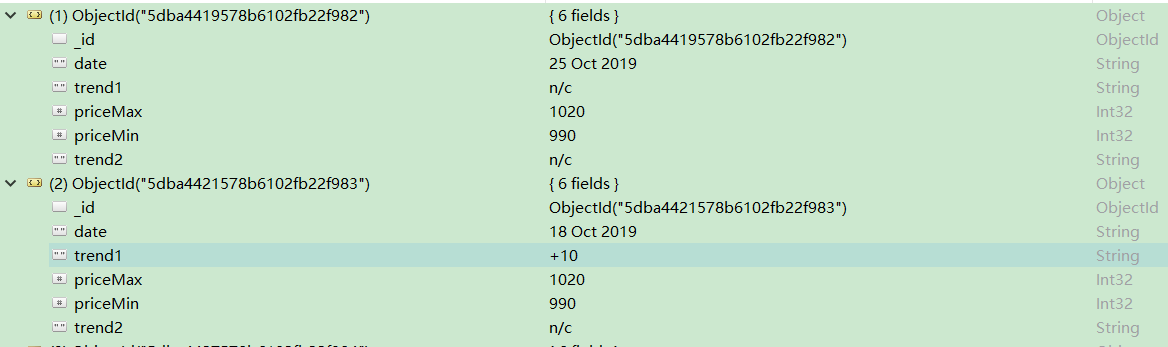


1. www.icis.com

登录、下载pdf并提取指定数据：



数据存储：



（2）数据可视化与界面展示

界面内容包括七个栏目，每个栏目包含表格，文字以及数据可视化图片，具体包括四个部分（每个部分根据不同栏目内容进行调整）：

1. 近期不同地区材料价格展示
2. 材料价格走势文字评价
3. 不同地区材料价格一周走势可视化
4. 不同地区材料价格长期走势可视化

界面左侧按钮可选择查看栏目，栏目内容较多，可以以滚动条的形式查看。不同地区以及不同材料的数据可视化以下拉框的形式切换。界面内容根据数据库内容生成，即会根据数据更新而对应更新。

以BOPP日评、CPP日评和ICIS为例，展示界面。近期数据以表格形式呈现给用户，可以看到五个地区近期的具体价格；表格下方是专家的文字预测；之后是一周的价格走势可视化；最后以下拉条目的方式选择查看不同地区的价格可视化。



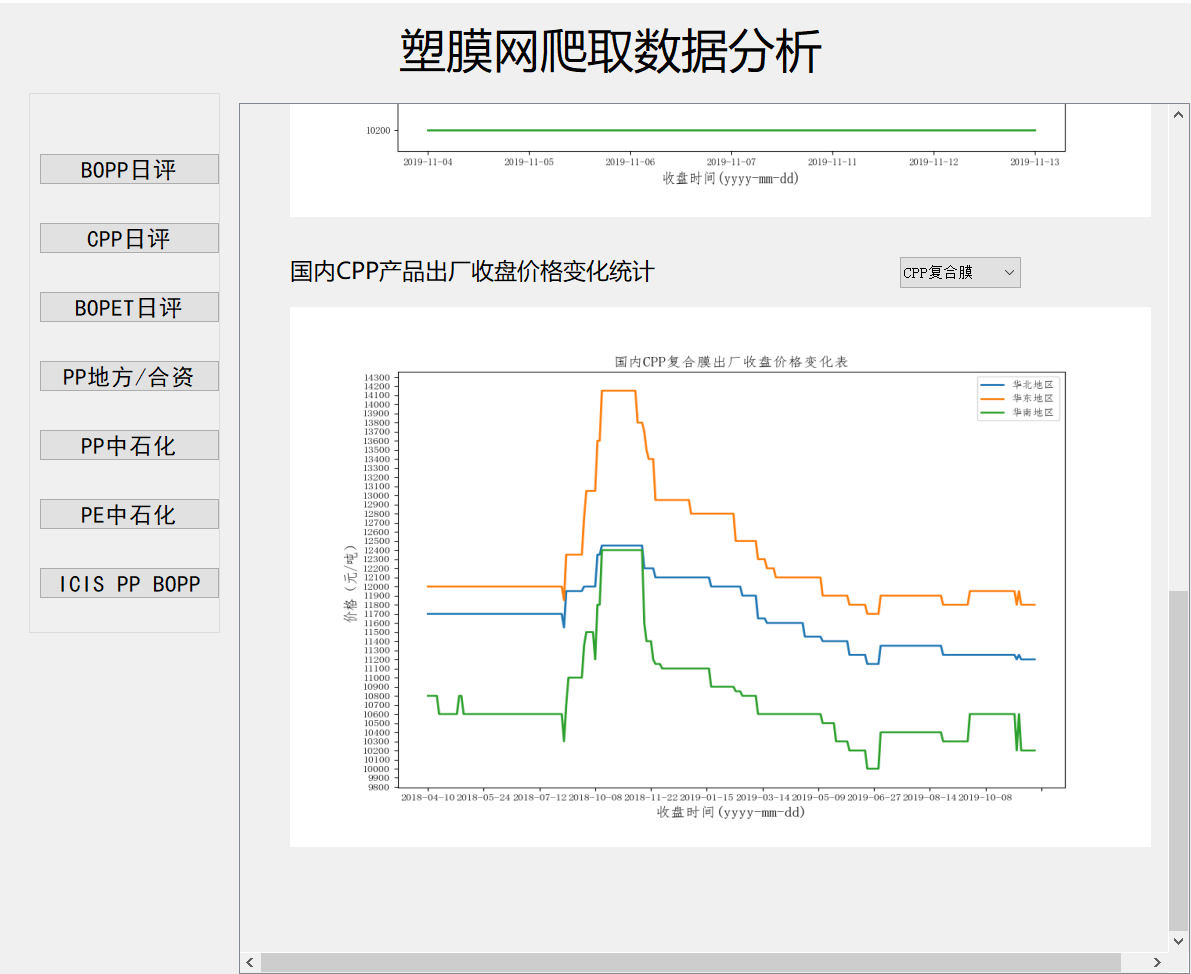
BOPP日评1



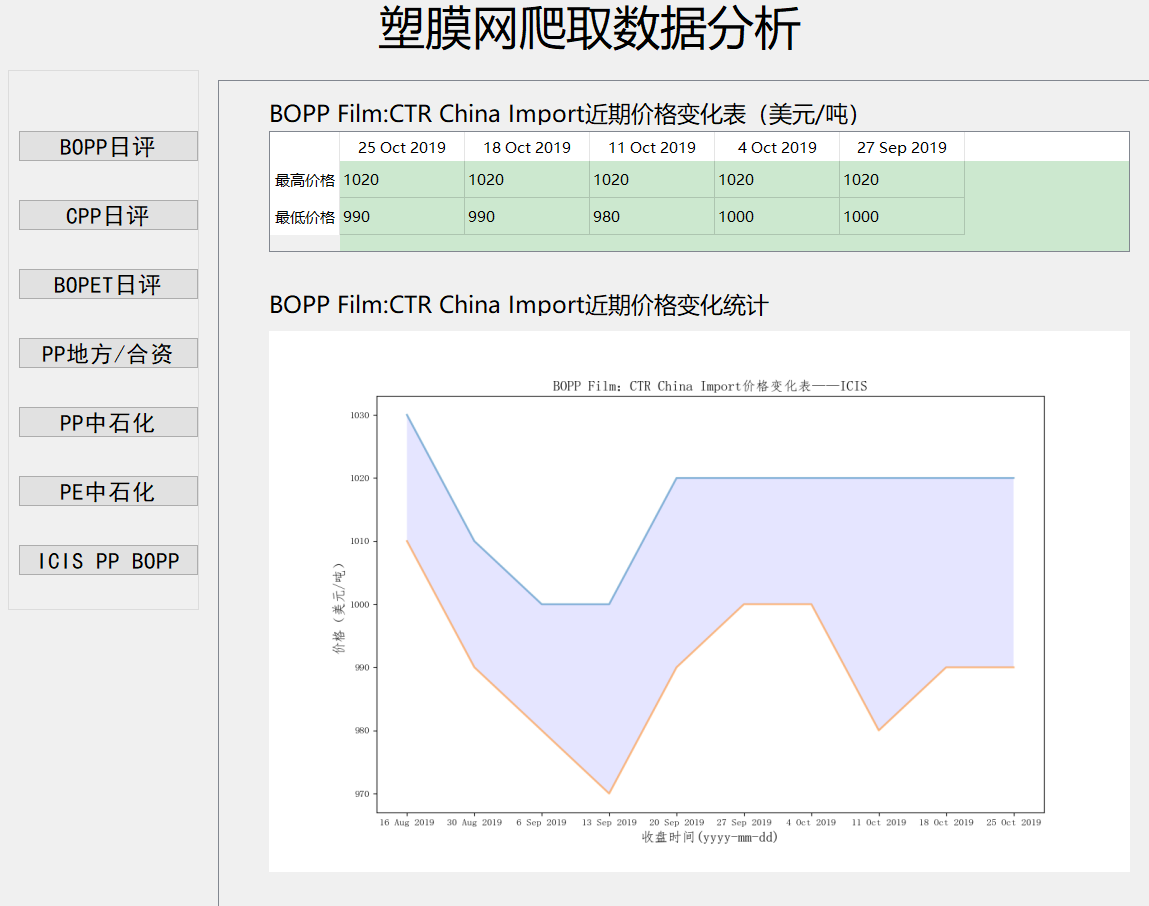
BOPP日评2



CPP日评1



CPP日评2



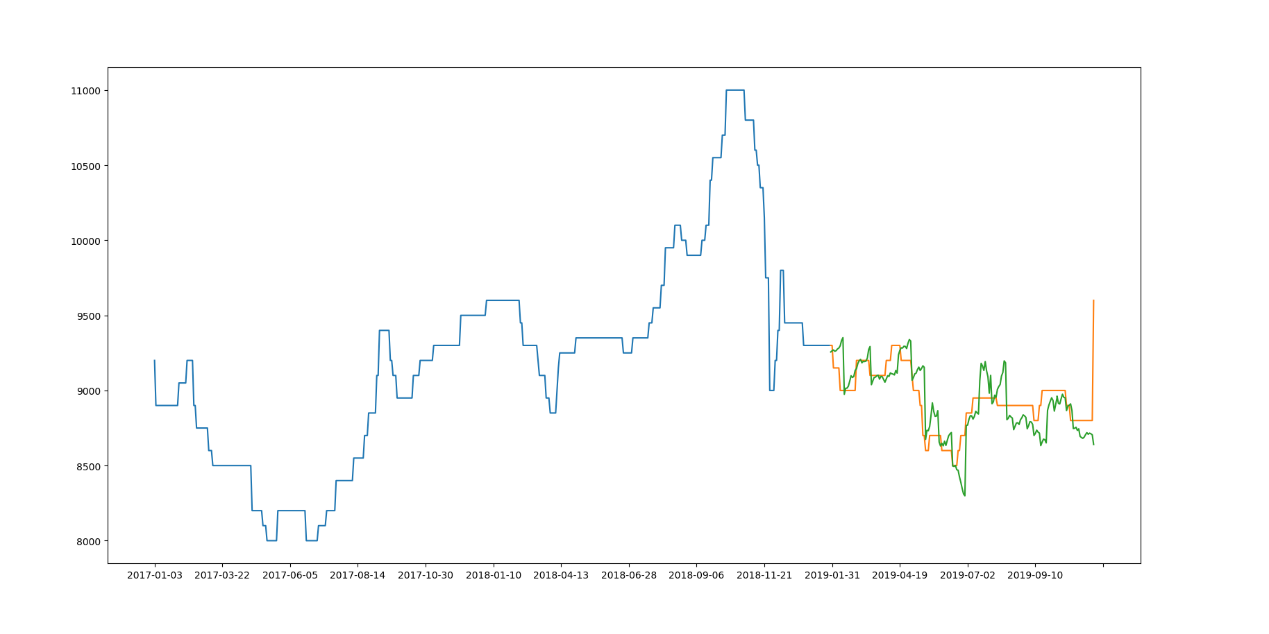
ICIS PP BOPP

（3）数据分析与预测结果

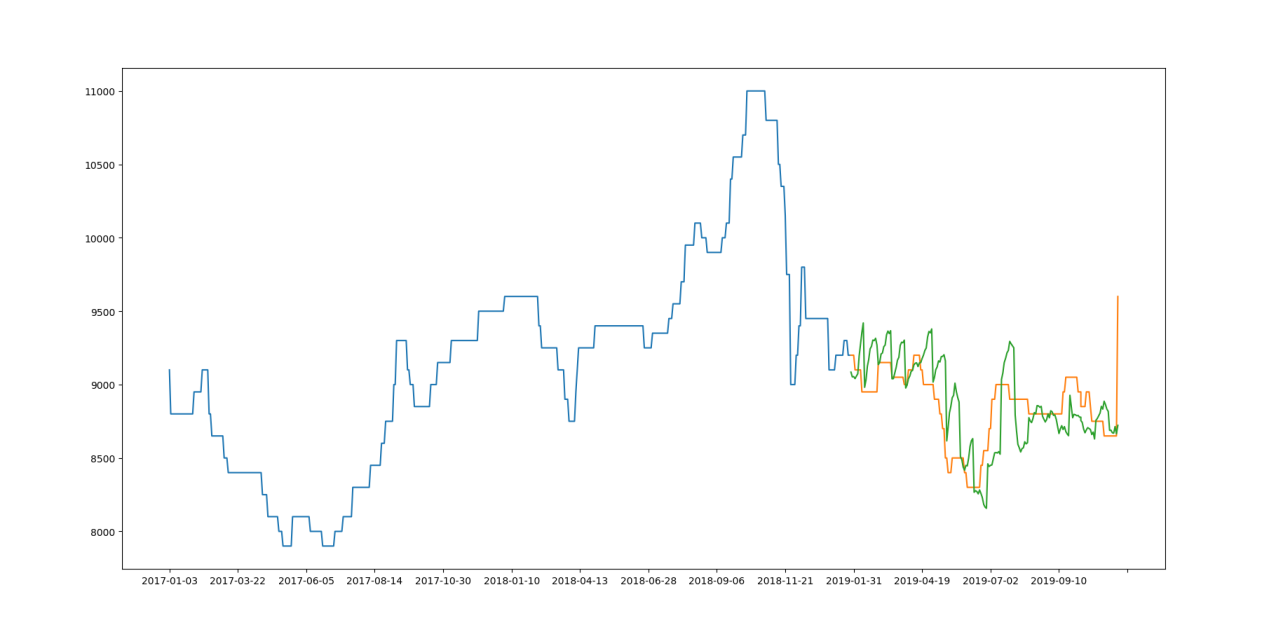
a.数据预测结果

由于材料的种类及地区比较多，这里列举了两种材料的预测结果，图中蓝色线条为训练集，橙色线条为测试集真实数据，绿色线条为测试集预测数据。

国内中石化出厂：BOPP模料：



国内中石化出厂：拉丝



由于影响材料价格的因素很多，尤其受外界因素影响较大，比如国际交易政策等因素影响，因此这里的预测结果准确率还有待提高。

b．数据分析结果

具体分析过程以及分析结果在关键原理4.2部分详细说明，以CPP为例，这里做一下简单总结：

1. 在价格分布上，同种材料在华东地区价格最高，其次是华北地区，华南地区材料价格最低。
2. 对于同一种材料，不同地区的价格走势基本一致，说明材料价格的波动是在全国范围内的波动。
3. CPP镀铝基材区域间价格影响最为明显，CPP复合膜与CPP蒸煮膜相对次之。三个区域中，华东地区与华南地区之间的相互影响最为明显，对于CPP复合膜尤甚。
4. 华北地区CPP蒸煮膜与CPP复合膜价格波动幅度远小于CPP镀铝基材；华东地区CPP蒸煮膜波动幅度远小于CPP复合膜与CPP镀铝基材；华南地区两种材料波动相对一致，但CPP镀铝基材价格回落速率较慢。
5. 华北地区CPP复合膜与CPP蒸煮膜之间价格关联极大，而CPP镀铝基材与CPP蒸煮膜关联相对小很多，CPP镀铝基材与CPP镀铝基材价格的相关性介于前面两种之间。
6. 相较于另外两个地区，华东地区材料价格关联性要大很多。
7. 三种材料在近期的价格走势基本一致，说明影响材料价格的因素是全局性的因素，其对于不同地区，不同种类的材料价格有着同的影响效果。三种材料价格之间的相关性都较高，其价格波动有相当大的内在联系。
8. 其中CPP复合膜与另外二者的相关系数都超过的0.93，而CPP蒸煮膜与CP镀铝基材之间的相关系数在0.85左右，说明CPP复合膜的价格波动对于其他材料的价格走势都有极大的参考意义，而另外二者的价格波动对于其他材料的价格走势具有比较大的参考意义。

### 六、收获与感受

在这次项目中，我们系统学习了python的爬虫开发、数据处理和计算以及机器学习等方面的内容。我们实现了一个完整的爬虫项目，对数据进行了可视化，同时根据数据特点进行数据分析，使用了深度学习框架对时序数据进行预测，最终实现了数据清晰完整、用户操作便捷的交互界面。

具体来说：

1. 网页抓取：在这里由于登陆限制，导致前期进度难以推进，对登录机制进行研究后，找到了解决方案；网页抓取方面由于没有验证码等机制，实现相对顺利。
2. 数据解析与存储：由于网站结构并不是特别规范，解析时需要对不同的页面结构分别进行处理。
3. 数据可视化与界面制作：使用PtQT制作界面相对便捷，但项目中材料种类以及数据类型较多，所以逻辑部分代码的编写较繁琐。
4. 数据分析与预测：由于数据受外界因素影响较大，预测方面准确率还是有所欠缺的，后期可以加入更多数据来进行分析。

### 完成时间：2019年11月26日