# Python项目总结小论文

1. **选取背景**

**本次作业为个人原创。**

有人说，21世纪是数据信息时代，的确，随着计算机科学的飞速发展，现今的我们的生活中无处不是信息，无处不是数字，怎么处理海量的信息与爆炸式的数据，已经成为了各大IT公司研究的主要问题了，作为人工智能的分支—机器学习在当今的社会应用得越来越广泛。

在当今的信息社会中，如何预测数据的走势已经成为了数据科学研究的重点，而在预测数据中，回归分析则是目前应用最广的算法之一。

回归分析（英语：Regression Analysis）是一种统计学上分析数据的方法，目的在于了解两个或多个变量间是否相关、相关方向与强度，并建立数学模型以便观察特定变量来预测研究者感兴趣的变量。更具体的来说，回归分析可以帮助人们了解在只有一个自变量变化时因变量的变化量。一般来说，通过回归分析我们可以由给出的自变量估计因变量的条件期望。

回归分析是建立因变量Y（或称依变量，反应变量）与自变量 X（或称独变量，解释变量）之间关系的模型。简单线性回归使用一个自变量 X，复回归使用超过一个自变量（X1,X2,X3…Xi）。

在这次课程设计中我计划使用岭回归算法预测某一股票的价格走势，现今社会中股市是大部分人选择投资理财的一部分，然而股票的价格的涨跌却和许许多多的因素有关，我在想能否利用一些数据作为参数，较为准确的预测某只股票的走势，而scikit-learn则提供了很多用于机器学习的包来进行数据分析，我觉得简单方便是python语言的最大魅力，既然有些算法已经为我们准备好了，那我们为什么不去用它来做一些有趣的事情呢？

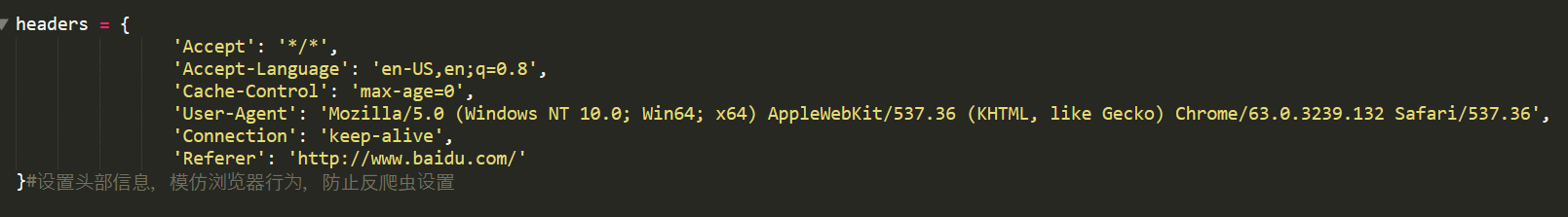
在选题的时候我选择了分析商品价格趋势这一题目，但是在后来的实践中，我发现无论是天猫还是京东在历史价格的保存中都不是那么完善，总是会有各种数据的丢失，不利于后期工作的开展。所以，我后来选择了从网易金融爬取特定股票的价格来进行分析。这样做的话，可以获得尽可能完善的数据，以利于后期工作开展。

**二、分析、开发、测试过程介绍**

**1、数据准备**

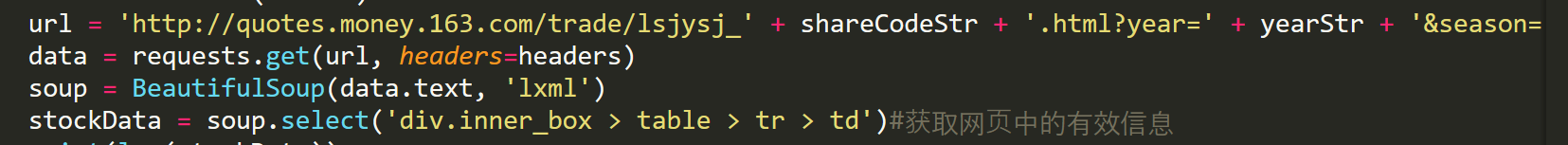
网络爬虫（英语：web crawler），也叫网络蜘蛛（spider），是一种用来自动浏览万维网的网络机器人。其目的一般为编纂网络索引。网络爬虫始于一张被称作种子的统一资源地址（URL）列表。当网络爬虫访问这些统一资源定位器时，它们会甄别出页面上所有的超链接，并将它们写入一张“待访列表”，即所谓爬行疆域（crawl frontier）。此疆域上的诸URL将被按照一套策略循环访问。如果爬虫在他执行的过程中复制归档和保存网站上的信息，这些档案通常储存，使他们可以被查看。阅读和浏览他们的网站上实时更新的信息，并保存为网站的“快照”。

然而，由于爬虫访问网站的过程会消耗目标系统资源。不少网络系统并不默许爬虫工作。因此在访问大量页面时，爬虫需要考虑到规划、负载，还需要讲“礼貌”。 不愿意被爬虫访问、被爬虫主人知晓的公开站点可以使用robots.txt文件之类的方法避免访问。这个文件可以要求机器人只对网站的一部分进行索引，或完全不作处理。所以，我们爬虫的时候为了避免被网站识别出来，我们会在爬虫时添加头部信息，使网站认为我们是一般用户操作：



其中：User-Agent信息可以由chrome浏览器中的extension查看。

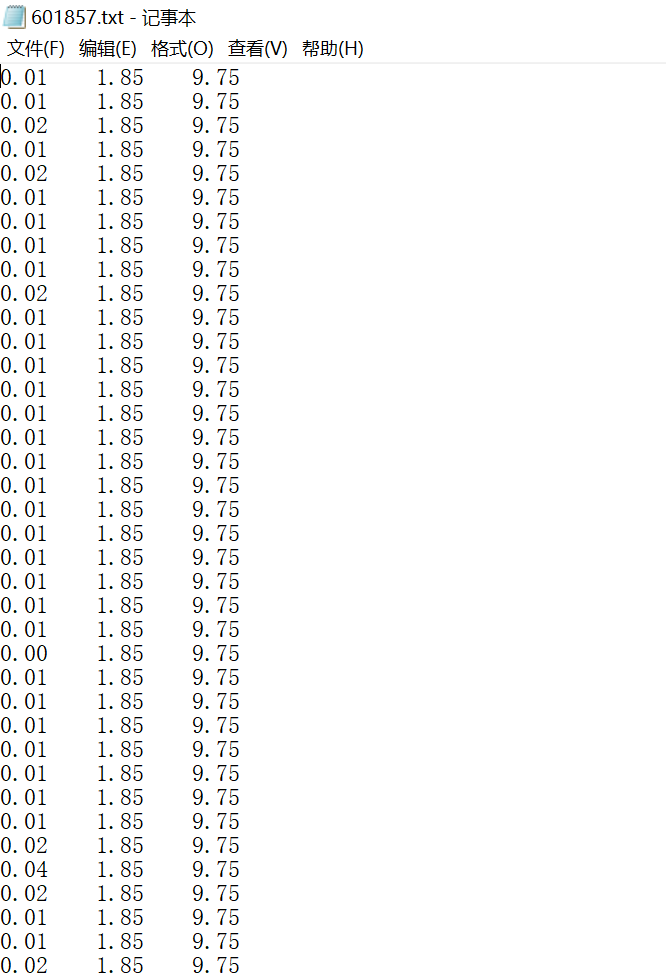
在爬取数据时，我使用了BeautifulSoup库（Beautiful Soup提供一些简单的、python式的函数用来处理导航、搜索、修改分析树等功能。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，因为简单，所以不需要多少代码就可以写出一个完整的应用程序。）来抓取数据，使得工作量大大减轻：



然而，在整理数据时我发现，爬取的数据保留在了一个列表里面，而其中所有的数据项都带有网页源码的信息，所以我们还需要使用get\_text()函数进行处理，由于我们需要的只是每天该股的收盘价（下一天的开盘价），所以我们只需要特定列的数据，而不是全部数据，与此同时，我发现网易财经给出的数据是按日期从大到小排序的，所以我在获取数据的时候必须倒序获得，才有利于后续工作的开展：



最后，爬取数据的工作已经顺利完成，下一步我计划利用岭回归对获得的数据进行分析，当然，仅仅依靠每日的收盘价格可能并不会很好地对数据进行预测，所以我又加入了振幅和换手率作为参数进行分析，数据爬取情况如下：



**2、设计用户界面**

为了给予更好地用户体验，我设计了用户界面，给出代码如下：  
  
 运行效果如下：  


**3、搭建岭回归模型进行分析**

对于一般地线性回归问题，参数的求解采用的是最小二乘法，其目标函数如下：

argmin||Xw−y||2argmin||Xw−y||2

参数w的求解，也可以使用如下矩阵方法进行：

w=(XTX)−1XTyw=(XTX)−1XTy   
结论：传统的基于最小二乘的线性回归法缺乏稳定性。

岭回归的优化目标:

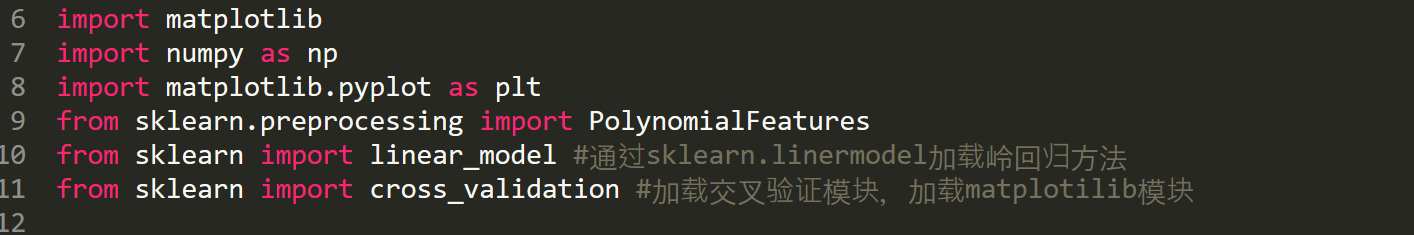
argmin||Xw−y||2+α||w||2argmin||Xw−y||2+α||w||2

对应的矩阵求解方法为:

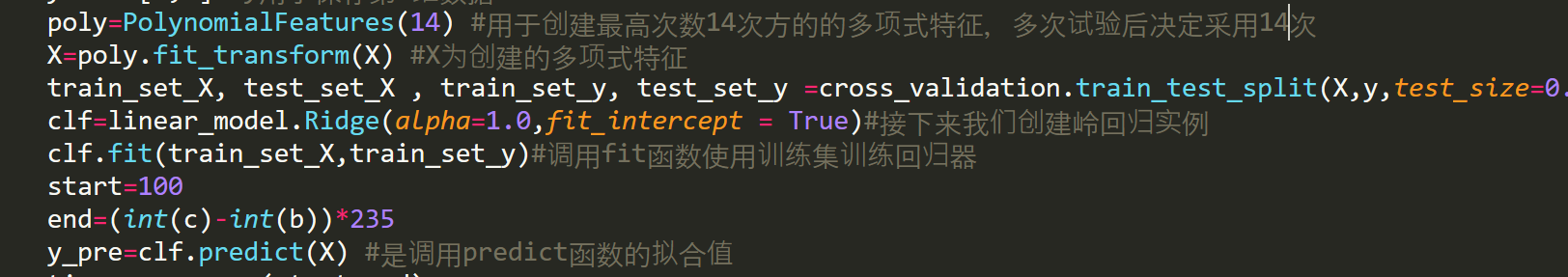
w=(XTX+αI)−1XTyw=(XTX+αI)−1XTy

岭回归(ridge regression)是一种专用于共线性数据分析的有偏估计回归方法,是一种改良的最小二乘估计法，对某些数据的拟合要强于最小二乘法。

对于一般情况的分析数据我们需要导入一下包：

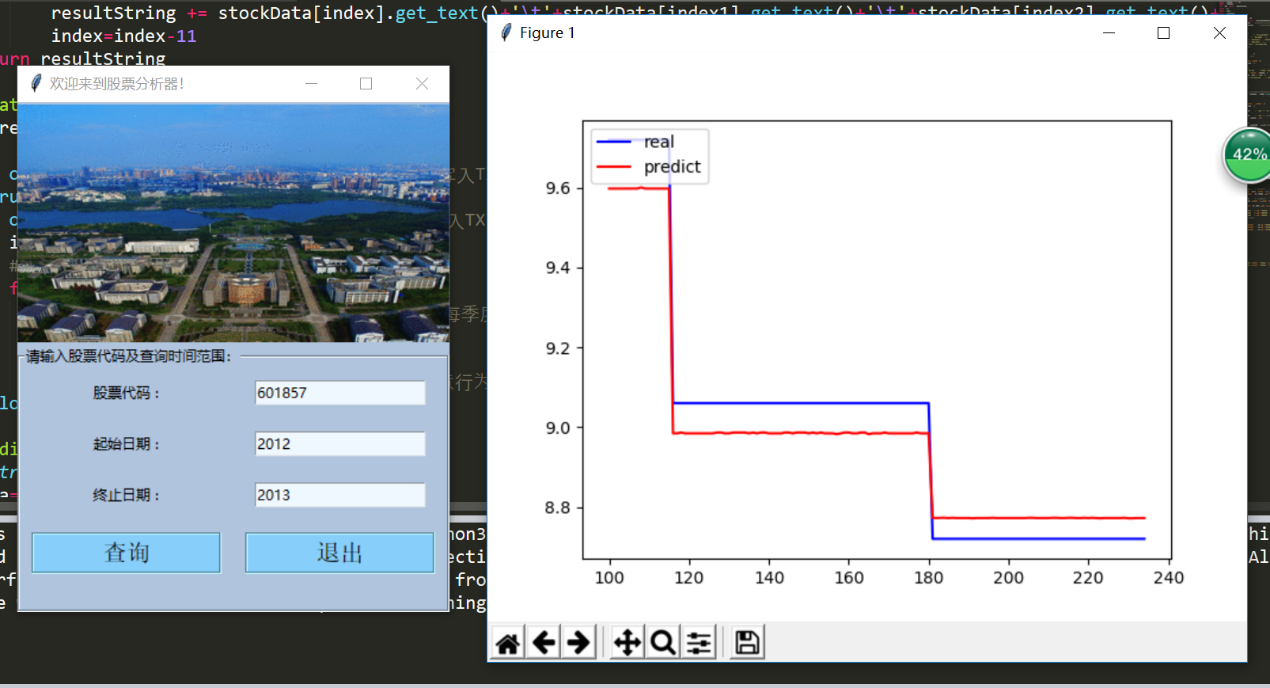


对于处理数据，我们需要用岭回归对数据进行分析，Python的sklearn中提供了大量的对数据进行分析的函数（经过多次试验，我采用14次方的拟合函数）：

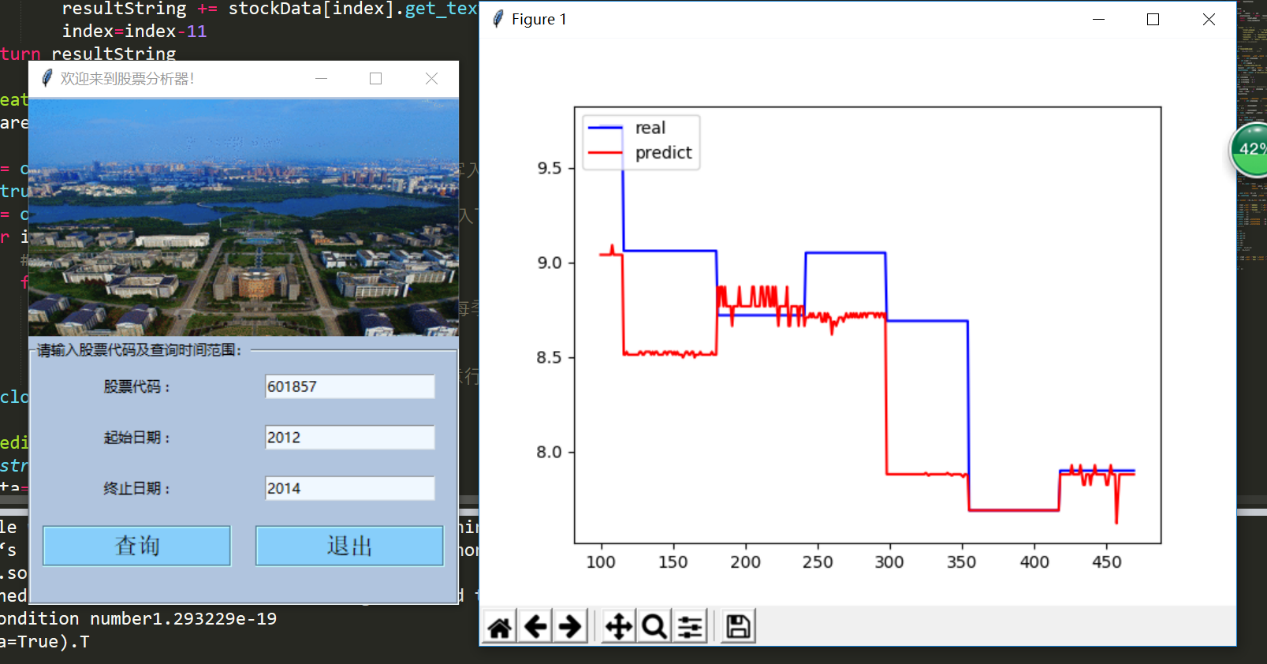


下面为运行结果（请在每次运行之前删除上次运行留下来的TXT文件）：

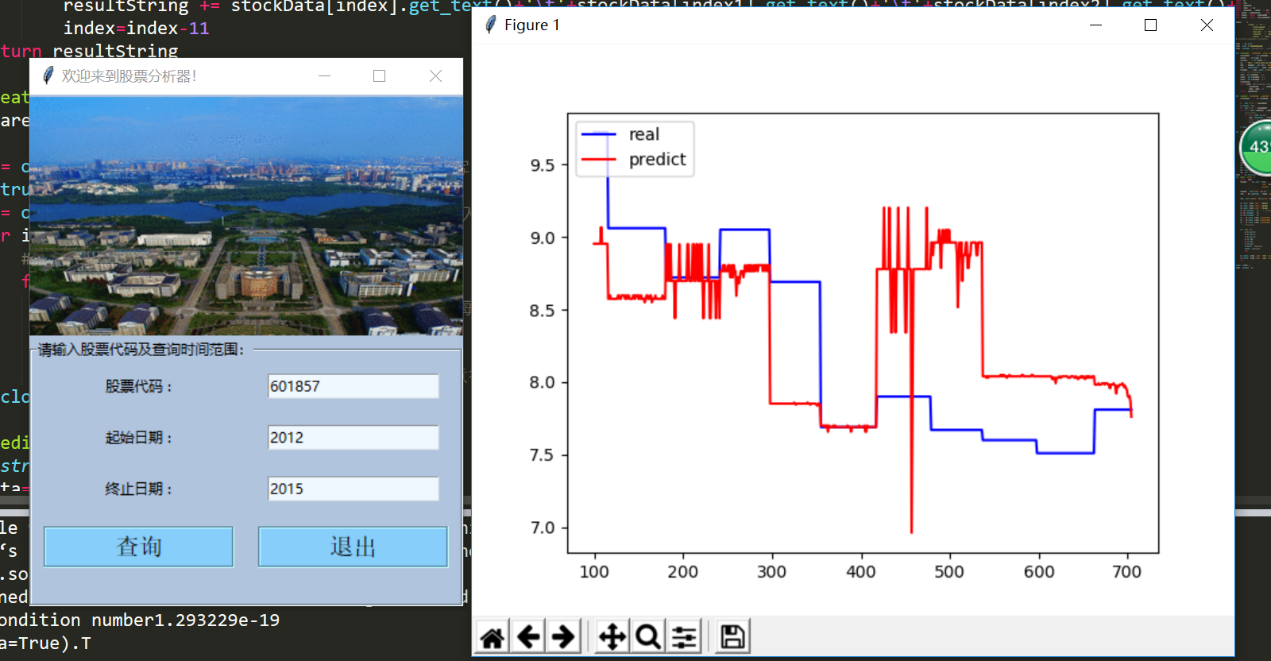
2012-2013对中石油股价进行分析：



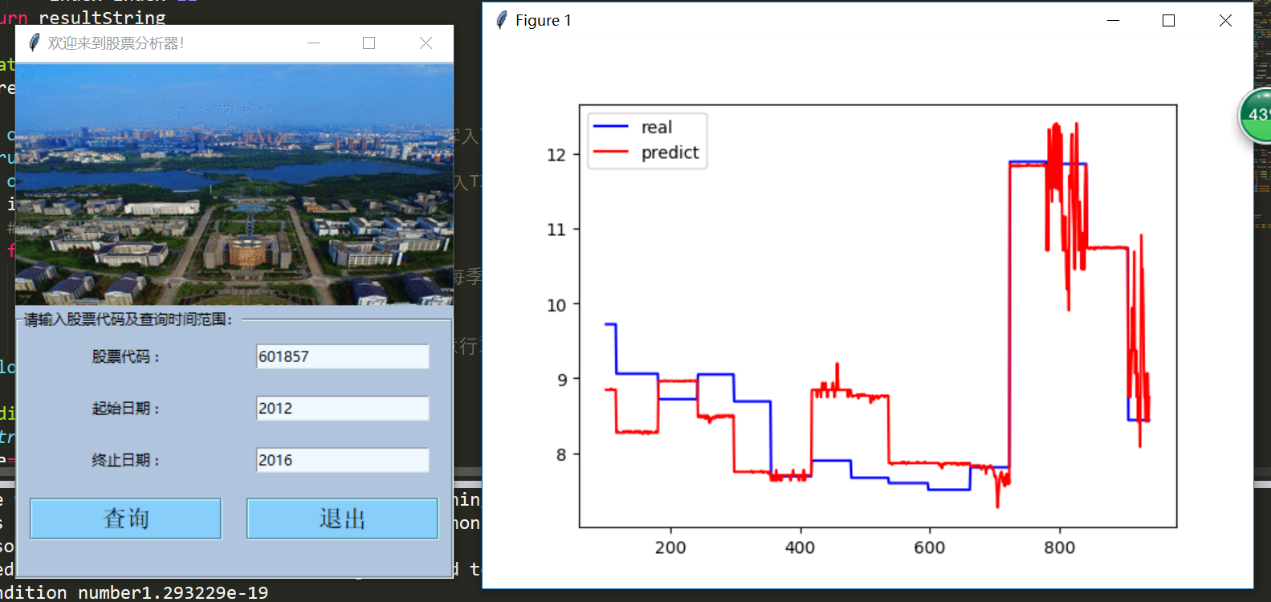
2012-2014对中石油股价进行分析：



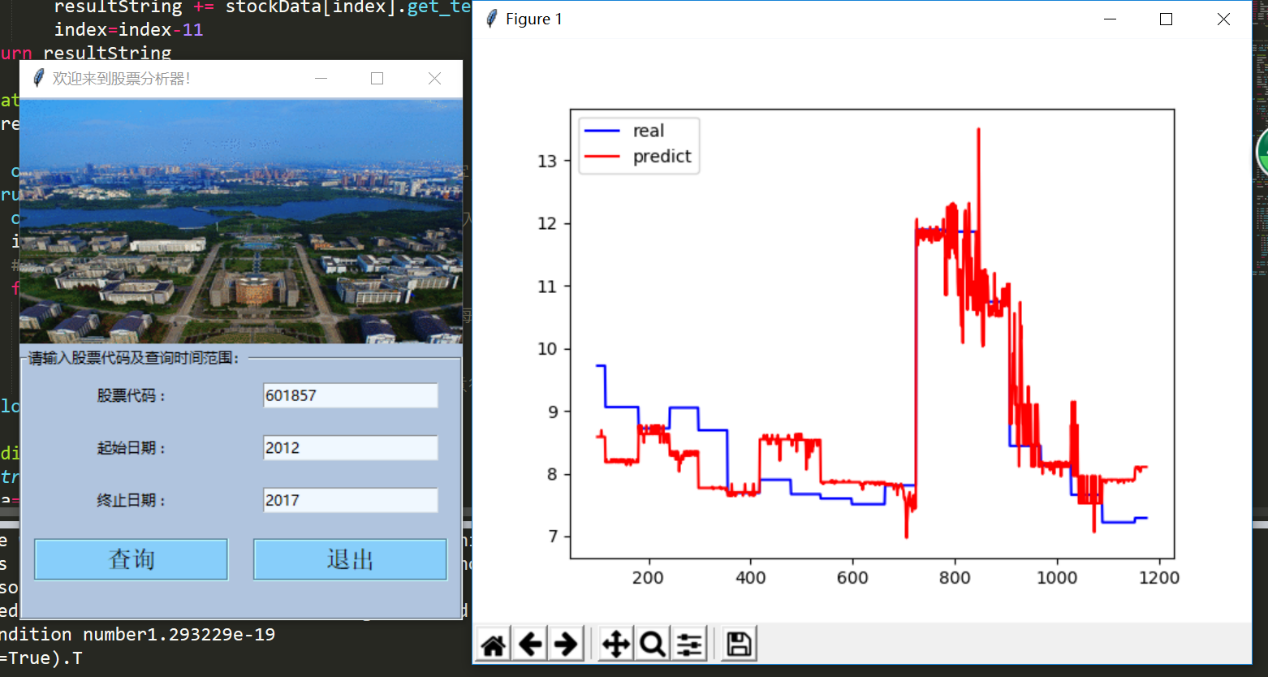
2012-2015对中石油股价进行分析：



2012-2016对中石油股价进行分析：



2012-2017对中石油股价进行分析：



**三、项目自我评价**

对于本次项目，我给自己打85分左右。

Python的本质，我个人认为是利用丰富的代码包（轮子），去解决一些原本很复杂的问题，就像本例中，对于从网站爬取数据，还是搭建岭回归分析模型，我的代码都十分短小精悍，总体下来也就是100多行代码（包括界面部分）。所以我觉得我的课程作业基本上体现了Python编程语言的较大优势。

但是，由于缺少足够的金融学知识，以及股票样本数据过于离散化，数据规律与较多复杂因素有关，所以导致最后的模拟情况只能拟合数据的大体变化情况，而不能精准预测股票价格走势；由于我现在正在准备研究生考试，时间有限，所以没能继续扩展下去，对此我觉得十分遗憾，如果以后有时间，我会继续努力，争取继续优化数据拟合结果。

**四、课程总结与建议**

经过一学期对于Python课程的学习，我基本了解了Python语言的语法与基础编程思想，并学会使用Python语言进行一些实际问题的操作；就像任何一门其他语言一样，只掌握理论知识是没有用的，还需要大量的实践，才能掌握一门编程语言。所以我认为这门课以一个Python小项目的形式结课是一个非常好的评判方式，让我们可以把理论知识应用到实际应用中去，而不是“纸上谈兵”。

段老师在上课的时候对于语法得讲解十分透彻，让我们觉得受益匪浅；如果能在最后的几节课中，多讲一些前沿知识，或者举出Python在数据处理中的应用，或许我们可以更好地在以后的学习中运用Python语言！