摘要

在当前我国城市化进程的快速推进下，我国的交通规模、能源消耗也在不断扩大，一氧化碳等有毒气体及固体污染物大量增加，严重影响了人们的正常生活，如何减少空气污染、打好污染防治攻坚战，对推动生态文明建设有很强的指导性。所以我从国内某的空气质量记录网站上记录的 384 个城市通过网络信息爬取获得了从 2014 年至今的空气质量记录数据，包括一氧化碳浓度、二氧化硫浓度等参数，然后对数据中的缺失值、异常值进行处理，数据可视化、分析数据特征、接着对每个城市先按照省份进行分组，并将数据保存到数据库中，同时对每个城市采用时间序列分析，在使时序数据变得稳定后，对时序数据进行预测，用户可以通过输入城市和起止日期来预测这段时间的空气质量指数。

关键词：□空气质量预测□ARIMA 模型□网络信息爬取□时序分析

Abstract

With the rapid promotion of our countries urbanization rate、 the urban traffic scale and energy consumption are also enlarging rapidly、 which raise a plenty of toxic gas and solid grain contamination、 such as sulfur dioxide and fine particulate matter、 respirable solid pollutants and carbon monoxide. These pollutant make a serious influence to humans' normal life. So、 reduce the air pollution and winning the Pollute prevention and management battle would have a great instructive for promoting the Ecological Civilization Construction. Therefore、 I got three hundred and eighty four cities air quality index data、

including carbon monoxide concentration 、 sulfur dioxide concentration and so on since 2014 to now through the Network information crawling technology recorded on the air quality records monitor on-line website. Then processing the air quality index data by dealing the missing and abnormal values. After that、 dealing the data by making data visualization and analyses data characterization. The next step is to classify the city by its province、 and save the data in the database. At the same time、 make time series analysis to the each city、 after make the time series stationary、 forecast to the time series. And the user can forecast the air quality index by input the city’s name and the start and end date.

Key words： Air quality forecast ARIMA model Network information crawling Time series analyze

目 录

1. 绪论

1.1 研究背景

自改革开放四十年以来， 我国的工业化和城市化进程的快速发展， 经济得到了飞速的发展，并给我国人民带来了巨大的物质财富和更舒适的生活，然而这一切却对我国的生态环境造成了严重的破坏，从早期的乱砍乱伐造成黄河流域和西部地区的荒漠化，到现在沿海地区的雾霾的严重超标， 无一不对我们的正常生活造成了严重的影响， 甚至对我们的身体健康造成危害， 如一系列的呼吸道、消化道疾病等。尤其是在 2000 年到 2010 年期间，当时我国由于在大气污染防治方面的经验不足，当时许多的雾霾天气都被误报成大雾天气。就在 2004 年，新华网发出来一篇报道，标题是《背景首都机场因雾出现近年最严重的航班延误》，当时所谓的“大雾”天发生后不久，北京居民的短时间内的呼吸道发病率大幅增加，并引起政府和民众的广泛关注。区域性大气污染问题已经对经济和社会的可持续发展以及人类的正常工作和生活造成了严重的影响，一下子成为了尤为突出的社会问题摆在了政府和监管者面前。

目前，我国的大气监测的污染物包括臭氧（O_3）、二氧化硫(SO_2)、二氧化氮(NO_2)、一氧化碳(CO)等有害气体及 PM_{10}、$PM_{2.5}$两种可吸入颗粒，且这些污染物污染分布广泛，主要分布在一些工业化水平和城市化水平较高的区域。城市空气污染主要表现为以下几个方面：

（1）悬浮的颗粒物总浓度在城市的范围内普遍超标，这一现象在以工业为主要产业的城市尤为明显，特别是 $PM_{2.5}$和 PM_{10}的浓度过高。

$PM_{2.5}$ 和 PM_{10} 是指大气中直径分别小于 2.5 微米和 10 微米并可以在空气中悬浮较长的时间的细小颗粒物，这些细颗粒物在空气中的单位体积内的平均浓度越高，就代表此时的空气污染越严重，并且这些细颗粒物对能见度有严重的影响。这些细颗粒物的成分主要包括有机碳(OC)、 元素碳(EC)、 硝酸盐（NO_3^-）、 硫酸盐(SO_4^{2-})、铵盐(NH_4^+)、 钠盐(Na^+)等无机盐以及一些有机物的细小颗粒。这些悬浮的细颗粒物会首先通过呼吸作用由呼吸道进入肺部，然后通过支气管和肺泡进入到血液中，导致混合在其中的如无机物、重金属等有害物质溶解在血液中，这些物质对人体的伤害巨大，甚至可能诱发如哮喘等慢性疾病。

（2）二氧化硫浓度普遍较高

二氧化硫(SO_2)是一种主要存在于火山爆发、化石燃料燃烧、含硫矿石的冶炼时产生的无色有刺激性气体， 是大气中的主要污染物之一。由于煤炭和石油等化石燃料都含有硫元素，其中的硫元素化合物在化石燃料燃烧时与氧气反应生成二氧化硫，且由于二氧化硫易溶于水，其溶于水时会形成亚硫酸（H_2SO_3），当亚硫酸在 $PM_{2.5}$存在的条件下，其会被进一步氧化成硫酸，当硫酸与云层中的水混合，并在达到一定的

温度和湿度的条件下形成降水，这就是酸雨的来源。而在我国陕西省盛产煤炭，且在我国的能源消费中，煤炭的消费占据了相当大的一部分(见图 1)，因此我国华北平原地区的二氧化硫浓度比我国其他地区都要高(见图 2)。且从图 1 中可以看到，从 2014 年起，每年我国的能源消费总量也在不断上升，尽管煤炭和石油的消费总量变化不大，但由化石燃料燃烧所造成的污染仍十分严重。且我国的主要产煤地区主要在山西、陕西、河南，这些地区发电主要也是以煤炭作为燃料，因此这些地区的二氧化硫排放量普遍要比其他地区高。

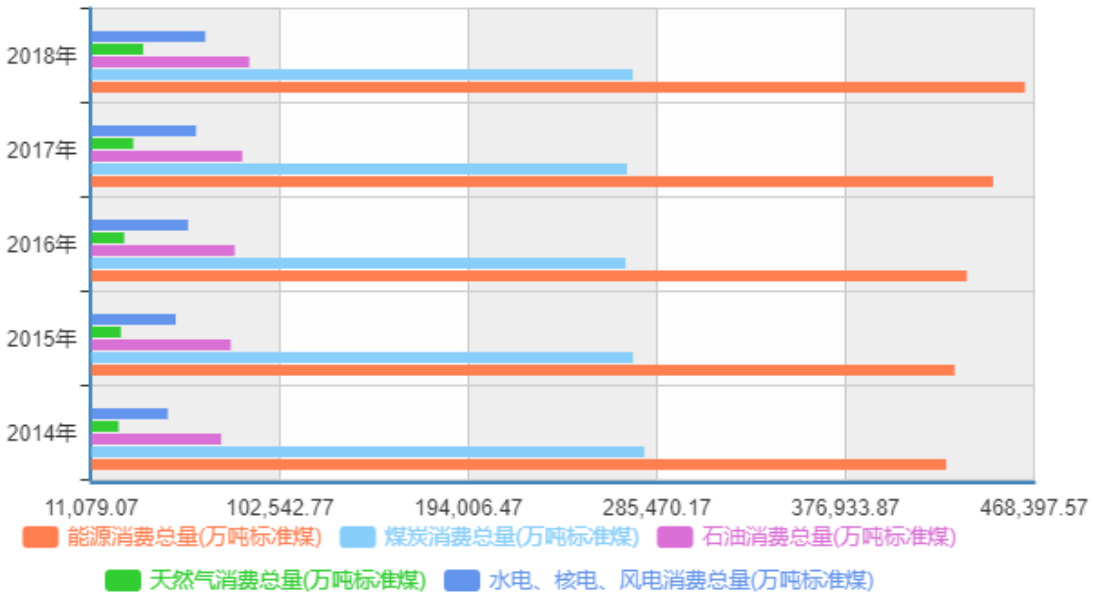

图 1 我国的能源消费总量柱状统计图（来源：国家统计局）

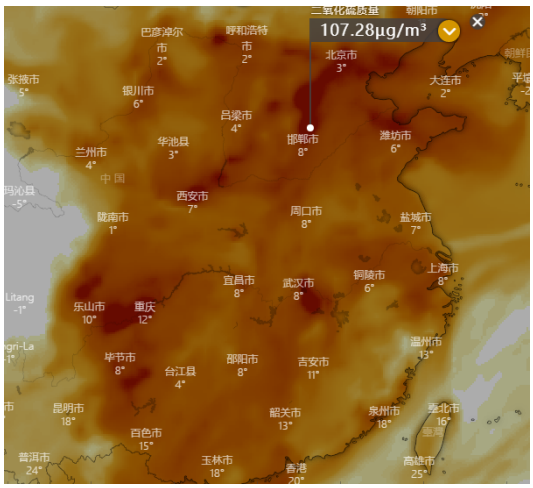

图 2 我国的二氧化硫浓度分布（来源：www.windy.com）

（3）一氧化碳浓度较高

一氧化碳 （CO） 是一种主要出现在汽车尾气、火力发电、金属提炼所的无色无味的气体。 一氧化碳在化学性质上既有氧化性也有还原性，同时还具有毒性，一氧化碳与氧气相比，前者与血红蛋白的亲和力远大于后者，

且一氧化碳会主动和血液的血红蛋白结合，从而阻止血红蛋白结合并运算氧气。它不仅会使血液的载氧能力降低，还使血液对人体组织的供氧量明显减少，从而使人产生缺氧的现象。吸入少量一氧化碳会导致人出现头痛、头昏、恶心等症状的出现；吸入大量的一氧化碳会使人昏迷，严重的会使人缺氧死亡，甚至产生如神经衰弱、 智力障碍等后遗症。

（4）二氧化氮浓度呈增加趋势，有些城市出现光化学烟雾现象。

二氧化氮 （NO_2） 是一种主要出现在汽车尾气、锅炉废气等高温燃烧过程所释放的在室温下呈红棕色且有强烈刺激性的气体。二氧化氮在被人体吸入后，会对人的呼吸道、眼睛及肺部造成巨大的刺激作用， 使人出现胸闷、咳嗽、咯泡沫痰等症状，吸入后几小时或间隔更长时间可能会出现迟发性肺水肿、呼吸窘迫综合征等呼吸道症状，在迟发性肺水肿消退后两周左右甚至可出现迟发性阻塞性细支气管炎；二氧化氮在慢性影响上主要表现为神经衰弱综合征及慢性呼吸道炎症。同时，它也是形成光化学烟雾的罪魁祸首，会严重降低大气能见度，从而引发交通事故；还能使地表水酸化，水体富营养化（由于氮、磷元素的营养物使藻类大量繁殖），并增加水体中的有害物质，使水中的鱼类因缺氧或有毒物质而大量死亡。

（5）臭氧污染严重

臭氧（O_3）是一种有鱼腥味的具有强氧化性淡蓝色气体，但在大气底层的臭氧并不是天然的，它是受环境污染的产物，这些近地臭氧主要在汽车尾气、锅炉排放的氮氧化物以及挥发性的有机物通过太阳光照辐射催化生成的，甚至连复印件的墨盒在打印时也会排放臭氧。它会使植物叶子变黄甚至枯萎，对植物造成损害；对人体的免疫机能也具有破坏性，使长时间直接接触高浓度臭氧的人出现疲乏、咳嗽、胸闷胸痛等症状。

目前我国正处于全面建成小康社会的决胜阶段，打好污染防治攻坚战，提升国家的生态文明水平，不仅可以满足人民群众美好生活的内在需要，也是落实中华民族永续发展前年打击的关键一步。而且我国地域辽阔，各地区的气候差异明显，因此形成了我国气候的多样化，同时各地区的经济发展也存在明显的不同，各地区的侧重产业也不一样。因此，综合以上因素，可以得出：我国各个地区的空气质量存在明显差异。对于空气质量的预报，是需要非常先进的技术以及各种先进设备的，而随着计算机技术的发展和各种新设备的出现，使得这种工作得以实现。随着人们对空气质量的重视和要求不断提升，人们的需求也从最开始的空气质量报告到现在的空气质量预测，而这种预测工作，也在成为环境科学以及计算机科学的一个重要的研究。[12]

1.2 研究现状

自从空气污染问题产生以来，对污染进行预测一直是空气污染防治的重要话题。

科研人员对空气污染预测方法进行了许多的探索和研究，在发展过程中出现了许多可以预测空气污染的方法，主要分为三大类， 分别为： 数值预测法、 统计预测法以及潜势预测法。其中，潜势预报方法依据气象条件对空气污染物扩散稀释作用，预报未来一段时间的空气污染情况，由于过于依赖对气象条件的判断，因此预报结果往往比较粗糙。而数值预报方法深入探索了污染数据和其他在大气环境中的一系列复制变化，预测结果准确率最高。但是涉及数学、物理、化学、气象等多领域交叉，同时需要丰富的气象数据资料和高性能的计算设备，在普通实验中很难实现。统计预测方法是基于大量现有的监测数据，采用先关的统计方法建立预测模型，模型构造简单易行，通过对相关数据的分析计算可以得出准确的较高的预测结果因此统计预测方法在大多数情况下更具有实用性。

在刚开始研究问题时，时序分析的主要方法是单一的预测模型，但由于各个模型的精度和应用范围都存在差异，所以目前预测领域研究的焦点主要是在如何通过结合各模型的优点对单一模型的局限性进行处理。而组合模型却能扬长避短，从多个角度挖掘信息，并系统全面地进行结论分析，因此组合模型更受青睐[13]。组合预测模型是将不同的模型组合后按照一定的比例来平均权重，以此吸收各模型的优点，使单一预测模型的不足得到了有效的规避，减少由于精度和应用范围所造成的差异，使预测结果更理想。最初，离差或误差是作为组合预测模型的主要的衡量指标，但是由于量纲和各个特征之间存在一定的差异，因此不同序列的离差及误差直接可比性较弱；即使以某种方式消除了量纲和各特征之间的差异所造成的影响，但由于序列本身之间以及数据的波动幅度也会导致出现一定的误差，因此预测方法的有效性也难以统一衡量。

而在我国生态环境部的官网上，可以看到各城市的 AQI 实时发布、AQI 指数日报，以及最长 120 小时的空气质量预报，无法对更长的时间跨度进行预测。因此，较长的时间跨度预测空气质量在国内尚存在一定的空白。国外的空气质量预测普遍选择人工神经网络来预报空气质量，并根据结果表明 MLP 模型（Multi-layer Perceptron）要比回归模型更准确，但对于峰值却无法准确的预测。由于我国的经济发展程度相对较为落后，导致我国在研究空气污染的分析和评价方面与世界发达国家相比都落后了不少， 但在近几年，我国的经济发展速度迅速提高，并推动了一系列的科研发展，使得我国在该研究领域范围内取得了很高的成就，随着大数据、机器学习等技术的广泛推广及应用，国内的环境信息系统也已经逐渐完善成熟，并开始向国际领先水平看齐。目前，国内的一些城市开始将环境信息系统加入到城市的管理作业中，并将其作为城市管理中不亏或缺的一部分。 国内的研究主要是在 BP 模型的基础上，对 NNs（神经网络）加入主成分分析，或者是将 BP 模型与灰色理论相结合，还有就是在 BP 模型中加入遗传算法。这些算法在一定程度上

都使得 BP 模型的不足之处得到了一定程度的解决，但也存在着一定的缺陷。其中有一种是通过将算法相互混合，形成弥补的状态来解决遗传算法原本的缺陷。对于这一设想，许多人尝试去设计混合算法，其中主要将遗传算法和贝叶斯正规化算法混合从而使算法更为合理，尽管在改进的过程中成功使遗传算法的一些缺陷得到了较好的解决，但是仍然还需要对其进行不断的改进。

1.3 研究内容

当今社会已经属于信息化时代，许多的信息都可以通过访问互联网来获取，当然也包括各地的空气质量记录，这是时代发展的必然结果，也是社会进步的重要指标。本设计通过研究分析，主要是通过计算机运算为了空气质量预测提供另一种科学可行的办法。毕竟单凭人工无法处理大量的数据，所以必须依靠计算机来对数据进行处理。然而，空气质量的预测需要大量准确的数据来进行支撑，如果数据不完整或者错误较多，会导致对空气质量的预测不完整和不准确，某种程度对空气质量预测的发展造成了影响。因此，为了对大气污染的动态变化作出及时的反应，并掌握变化规律，使训练模型时使模型更有效，从而提高空气质量预测大师准确性，为空气质量的预测提供另一种更科学合理的可行方法，本设计研究的内容为：

（1）首先通过网络信息爬取技术获取将网站上各城市的空气质量记录的链接爬取下来，并将重复的链接删除。

（2）将链接以文本的形式保存在本地硬盘文本文件中。

（3）读取文件中记录的城市所对应的链接，并逐一进行访问。

（4）将网站上每个城市所记录的空气质量爬取下来，并以表格文件的格式保存在本地硬盘中，以便日后使用。

（5）读取表格文件，对文件中的缺失值和异常值进行处理，并对文字类型的数据进行数据转换。

（6）对处理后的空气质量数据分析特征，并进行时序分析与预测，最后对预测进行验证，并生成可视化结果。

1.4 研究意义

大气环境质量与人们的健康和生活息息相关。在社会以及经济快速发展的同时，人们生活所造成的排放污染对环境产生了很大的影响和破坏，使生态环境持续恶化，空气污染甚至威胁到了人类自身的安全和健康，这将直接导致可持续发展受到破坏。我国作为一个发展中的大国、世界上第二大经济体，一直面临着来自环境问题的种种考验，如何能准确预测大气质量，为各地区的大气防治提供更多可参考的资料，以便提前做好部署和准备，为重大污染事件的发生做好充足的准备，并作出更长远

的监测和预防，尽可能将污染事件的所造成的影响降到最低，并将萌芽扼杀在源头，已成为社会和政府所广泛关注的重要问题。

我国政府在对控制大气污染以及污染物的变化方面，正在提升污染物监测的整体强度，并从中探索出各地的大气污染物的变化规律，从而可以通过更好的方法来预测空气质量和污染物的扩散，对大气污染加强监督监管，将各生产企业的空气排放控制在指标允许的范围内，确保排放出的气体经过无害化处理，务必减少对周边的生态环境及居民的正常生活和身体健康造成影响

对空气质量进行预测的研究，有两方面的意义：对于市民来说，可以通过各种渠道得知的空气质量预测提前做好个人防护，减少由于不知情的原因暴露在空气质量较差的环境下而诱发或感染疾病的可能；二是可以为环保部门提供有关空气质量的各种有关数据，从而对污染物和空气污染之间的关系和影响得出更为准确的判断和分析。同时，城市的空气质量在经过空气质量预测后可以更好地作出评价，使各城市环保部门的治理方案更加具有针对性和独特性，为城市的可持续发展提高评价水平，成为城市发展的重要监督指标。

在空气质量的预测方面，使用更加科学的技术，就是利用针对时间序列产生可靠的预测方法之一，称为 ARIMA，这种技术对空气质量数据所产生的时间序列进行非线性的检测和处理，使预测的效率得到了提高，并且在一定程度上提高了预测的可靠性和准确性。所以利用 ARIMA 模型对空气质量进行预测，具有较高且科学可行的价值及较好的应用前景。与传统的预测方法相比，作为一种对实践更具有针对性的分析预测模型， ARIMA 模型对空气质量的预测效果更为准确， 并有望成为未来空气质量预测的主要方法。

2. 网络信息爬取

网络信息爬取技术（网络爬虫）主要是将需要的网页通过网页下载器下载下来然后转换成字符串数据，字符串数据通过网页解析器解析成树形对象，将需要的数据通过网页解析器进行提取，如：文字、链接、图片等。网页下载器可以将制定的URL 网页下载到本地存储成本地文件或字符串格式，以便进行后续的数据分析，所以网页下载器是整个爬取程序的核心模块。[14]在本章中主要用到 requests，它是一个 python 的第三方库，支持网页下载、登录、文件上传等功能。网页解析器是一个能从网页字符串文件中解析出价值数据的处理器，python 中使用最广泛的是 BeautifulSoup 这个第三方库，BeautifulSoup 最主要的功能是将网页下载器所下载的网页进行解析，BeautifulSoup 自动将输入文档转换为 Unicode 编码，输出文档转换为 utf-8 编码。BeautifulSoup 支持 Python 标准库中的 HTML 解析器，它首先进行网页字符的结构化解析，利用 DOM 和 HTML 之间的映射关系，将 HTML 文档转换成 DOM 树，通过基于语义及基于结果的过滤来进行剪枝操作，通过树形结构能精确地定位到某个节点、属性、文本内容，然后使用 find 或 find_all 方法查询相应的节点，访问节点的属性、名称、文字等信息，从而提取出信息进行分析。在经过 BeautifulSoup 解析后，整个 HTML 文档会被转换成一个复杂的树形结构，且每个节点都是一个对象，所有对象可以归纳为 4 种:NavigableString 、BeautifulSoup、Tag、Comment。BeautifulSoup() 主要用来遍历文档树及其属性，并为此提供了多种方法，比如获取父子节点、兄弟节点等。在本章节中主要是从 BeautifulSoup 树对象中搜索出所需的目标，通过使用 find_all()方法在 BeautifulSoup 树对象中按照标签名称（name）、文本（text）、属性（attrs）等参数对所有 tag 的子节点进行搜索，并判断是否符合过滤条件，将所有符合条件的节点保存并输出。

2.1 获取城市访问链接

首先，打开浏览器，访问所需对其进行爬取的网站，本次是对中国空气质量在线监测分析平台进行爬取，这是一个公益性的软件平台，收录了 367 个格式的空气质量信息(如图 3)，且在经过对比后可得，他与国家生态环境部上公布的数据是一致的，而且，相较于生态环境部的访问限制，以及限定时间范围的下载及查询，该网站爬取信息的难度更低，仅是对访问的 user-agent 作出限制，并没有其他限制措施。

图 3 中国空气质量在线监测分析平台主页

使用检查功能查看网页源代码观察各城市的标签中所包含的地址，网站中所有的链接地址都被保存在＜a＞标签中（如图 4），在导入所需要的requests库和BeautifulSoup库后，就可以开始编写程序来爬取城市的空气质量记录访问链接了。

```
▼<div class="all">
  <div class="top">
      全部城市:
    </div>
  ▼<div class="bottom">
    ▼<ul class="unstyled">
      ▼<div>
          <b>A.</b>
        </div>
      ▼<div>
        ▼<li>
            <a href="monthdata.php?city=阿坝州">阿坝州</a>
          </li>
```

图 4 中国空气质量在线监测分析平台主页网页源代码

(1)首先创建字典UA，用于在访问页面时伪装成正常的浏览器访问，绕开检测。

(2)编写函数get_cityurl()用于爬取各个城市空气质量记录的访问链接。

过程：函数 get_cityurl():

1.填入空气质量记录主页地址 url

2.创建用于保存各城市及其访问链接的列表 city_url_list

3.访问网站主页并使用字典 UA 进行伪装

4.获取网页的源代码

5.使用网页解析器对网页源代码进行解析

6.使用 bs4 库中的 find_all()函数搜索网页源代码中所有包含<a>标签的子节点，

7.并返回列表类型的 url_list

8.For url_list 的每一个值 i：

9. 获取 i 中所包含的链接

10. If i 中包含只有各城市的访问链接特有的字段

11. 将访问链接的前缀补充到 i 中并转换成字符串得到 city_url
12. 将 city_url 添加到列表 city_url_list 中
13. End if
14. End for
15. 将 city_url_list 转换成集合 city_url_set
16. 创建文本文件 city_url.txt 并以写入方式打开
17. For city_url_set 的每一个值 url
18. 将 url 写入到文件中，并且每输入一个换行一次，以便后续使用
19. End for

输出：一个命名为city_url.txt的文本文件

在运行完这段代码后，将会得到一个命名为city_url.txt的文本文件，里面是网站主页上记录的各城市的空气质量记录的链接，在接下来的步骤会用到这个文件。

2.2 爬取城市的空气质量数据

接下来，开始对 city_url.txt 文本文件里面记录的各城市的空气质量记录的链接下的空气质量记录进行爬取，由于每个城市的空气质量记录是按每个月份一个页面进行展示，并且都是在网页中的<td>标签中，而且网站的空气质量数据需要经过动态加载才会显示在页面中，所以我选择使用 selenium 库并结合 webdriver 调用浏览器进行网络信息爬取，并且由于是使用真正的浏览器进行操作，所以并不需要添加特殊字段对爬取操作进行伪装。这次主要用到 BeautifulSoup 库和 selenium 库，在从网上下载webdriver.exe并安装到正确的目录后就可以开始编写程序来爬取各个城市的空气质量记录了，最后还需要将数据以表格形式保存在本地硬盘中。接下来开始编写代码对相关数据进行爬取：

编写函数get_city_aqi()用于爬取各个城市的空气质量数据。

输入：*path* = 文本文件 city_url.txt 的绝对路径

过程：函数 get_city_aqi(*path*)：

1. 创建参数设置对象 *chrome_opt*
2. 将无界面化参数添加到 *chrome_opt* 中，使浏览器无界面化运行
3. 创建 *driver* 对象来启动浏览器并使参数对象 Chrome_option = *chrome_opt*
4. 以只读的形式打开 city_url.txt 对应的 path 从而获得 *file* 对象
5. 使用 *file* 对象下的 readlines 方法获得文件所有行的内容并返回列表 *url_list*
6. 关闭文件 *file*
7. For *url_list* 的每一个值 *url*：
8. 输出 *url* 以便记录已经进行过爬取操作的城市并用 *driver.get* 访问 *url*

9. 程序休眠 2 秒等待页面完成动态加载
10. 使用网页解析器对浏览器获得的网页源代码进行解析
11. 使用 bs4 库中的 find_all()函数搜索网页源代码中所有包含<td>标签的子节点，并返回列表类型的 *tds*
12. 创建列表 *month_url_list* 和 *city_aqi_list* 分别用于储存当前城市每个月的访问链接和城市的空气质量数据。
13. For *tds* 中的每一个值 *td*：
14. If *td* 中包含只有月份的访问链接特有的字段
15. 将 *td* 转换成 str 类型，去除 *td* 中的<td>标签的内容，得到 *month_url*
16. 将 month_u*rl* 添加到列表 *month_url_list* 中
17. End if
18. End for
19. For *month_url_list* 中的每一个值 *month_url*：
20. 用 driver.get 访问 *month_url*
21. 程序休眠 2 秒等待页面完成动态加载
22. 使用网页解析器对浏览器获得的网页源代码进行解析
23. For find 函数搜索<tbody>标签后该标签下的所有子节点 *tr*:
24. If *tr* 与所需的数据类型相同：
25. 生成列表 *tds* = <tr>标签中的<td>标签
26. If 列表 *tds* 不为空：
27. 将列表中的元素用.text 转换成 str 类型并以列表的形式添加到列表 *city_aqi_list* 中
28. End if
29. End if
30. End for
31. If *city_aqi_list* 为空 or *city_aqi_list* 长度小于 365：
32. 输出文本信息提示爬取的城市无法爬取到数据或数据长度过短
33. Else：
34. 输出文本信息提示当前爬取的城市已完成爬取操作
35. 创建以当前爬取的城市命名的 csv 格式的文件并以写入的方式打开
36. 向文件写入列标题并换行
37. For *city_aqi_list* 里的每一行 *line*：
38. For *line* 里的每一个元素 *element*：
39. 向文件写入 *element* 和写入逗号隔开

40. End for
41. 向文件写入换行符
42. End for
43. 关闭文件
44. 输出文本提示完成文件写入
45. End for

输出:全国 27 省及 4 个直辖市共计 353 个城市自 2014 年 1 月 1 日至今的空气质量记录的 csv 表格文件。

2.3 将数据导入到数据库中

在获得上述的空气质量数据文件后，因为数据量较大，且不利于管理，需要将其导入到数据库中，使开发效率得到明显的提升，令数据的调用更为方便，程序规模得到简化，减少了程序的维护和修改的频率，对数据进行了集中化的管理，使冗余得到有效的控制，从而使数据的利用率和一致性得到了提高，对应用程序的开发和维护起到了积极的作业。Navicat 作为一款为降低系统管理成本及简化数据库的管理流程及操作的专业数据库管理软件，用户可以创建、组织、访问并共用信息以安全简单的方式，无需通过冗长复杂的指令来对数据库进行操作，并且可以对本机或远程的 MySQL、SQL Server、SQLite、Oracle 及 PostgreSQL 数据库进行管理及开发。

首先,打开 Navicat,并输入本地数据库正确的用户名及密码，连接到数据库后新建一个数据库并命名为 air_quality_data，用于存放城市的空气质量数据，然后选择导入数据，将本地的 csv 文件导入到数据库中，数据库设计如下：

字段名称	类型	字段说明	备注
Date	Varchar	日期	主键
AQI	Int	AQI 指数	
air_quality_level	Varchar	质量等级	
PM2.5	Float	PM2.5 浓度	
PM10	Float	PM10 浓度	
SO2	Float	二氧化硫浓度	
CO	Float	一氧化碳浓度	
NO2	Float	二氧化氮浓度	
O3_8h	Float	臭氧浓度	

然后开始将数据进行导入。导入完成后，Navicat 将会输出信息表明完成导入（如图 5）。

图 5

最后再随机打开几个表进行核查，检验是否有出现错误或空缺。可以看到，表中数据(如图 6)，与原数据表格式一致，且和原数据表对比并没有存在缺失值或异常值。

2014-12-31	60	良	26	70	12	0.6	11	60
2015-01-01	54	良	28	58	21	1	26	51
2015-01-02	63	良	45	74	25	1.8	48	54
2015-01-03	72	良	52	81	32	1.9	42	58
2015-01-04	101	轻度污染	76	109	25	2.5	59	60
2015-01-05	126	轻度污染	96	134	20	3	63	57
2015-01-06	117	轻度污染	89	107	16	2.5	21	67
2015-01-07	55	良	38	59	13	1.5	22	70
2015-01-08	72	良	53	82	17	1.9	41	56
2015-01-09	104	轻度污染	78	117	23	3.2	69	60
2015-01-10	100	良	75	114	20	3	65	62

图 6

3.数据处理

本章节是对已经保存在本地的空气质量记录数据进行数据清洗，包括缺失值和异常值处理、数据转换以及数据归一化，以便训练模型时提高模型准确度，减少误差，以及对数据进行可视化处理，观察数据特征，找出各个城市变化规律和差别。

由于原始数据集的情况并不清楚，所以需要对原始数据先进行了解然后再进行数据处理。以下以七台河市的空气质量记录为例，首先打开记录七台河市的空气质量的文件，使用代码对数据进行描述，可得：七台河的空气质量记录共有 1779 行 9 列数据，列名称分别为：日期、AQI、质量等级、 PM2.5 浓度、 PM10 浓度、 SO2 浓度、 CO 浓度、 NO2 浓度、 O3_8h 浓度。每一列的数据类型分别为：日期和质量等级为 object 类型，其他均为数值类型。在对数据使用统计分析函数时发现若干个数据列都存在最小值等于 0 的情况，说明数据存在异常值。因此要对数据进行缺失值和异常值检测和处理。

AQI(Air Quality Index)是环境空气质量指数的缩写，用于描述该环境的空气受污染程度的以及对健康的影响的一个参数。环境空气质量指数的重点是判断暴露在数小时或数日受到污染的空气对人体所造成的生理影响。环保局通过以下几个主要污染标准来计算空气质量指数：地面臭氧(O_3)，颗粒物污染（也称颗粒物），一氧化碳(CO)，二氧化硫(SO_2)，二氧化氮(NO_2)。我国环保局为保障人民的身体健康在 2012 年均已对上述污染物成立了新的空气质量评价标准。我国空气质量取 24 小时平均值作为发布标准；同时，由于我国与美国采用的空气质量指数及污染物浓度指标不同，导致存在污染物浓度相同而空气质量指数的计算结果也可能存在一定的差异，因此在查阅实时数据经常与会与媒体公布的结果不一致。

3.1 缺失值和异常值的处理

首先对数据进行缺失值检测，发现数据并没有缺失，因此不需要对数据进行缺失值处理。然后对数据进行异常值的检测，获取数据集中 AQI 为 0 或质量等级为无的行并将整行输出。在七台河的空气质量数据集中，一共有 4 行数据是 AQI 为 0 或质量等级为无，以及有部分数据经过核算后发现 AQI 指数是和污染物浓度不匹配的，因此需要对异常值进行处理，通过 AQI 的计算公式：

$$I = \frac{I_h - I_l}{C_h - C_l} \times (C - C_l) + I_l \quad (1)$$

并结合空气污染物浓度限值表(如表 1)计算得出各项污染物对应的 AQI 指数，最后取数值最大的为最终的 AQI 值。其中 I 为空气质量指数， 即 AQI 指数；C_l 、C_h 为该污染物浓度限值，I_l 、I_h 为 AQI 限值；C 为该污染物浓度，即输入值。

AQI	SO_2 浓度	PM_{10} 浓度	O_3 浓度	NO_2 浓度	$PM_{2.5}$ 浓度
0	0	0	0	0	0
50	50	50	160	40	35
100	150	150	200	80	75
150	475	250	300	180	115
200	800	350	400	280	150
300	1600	420	800	565	250
400	2100	500	1000	750	350
500	2620	600	1200	940	500

表 1　空气污染物浓度限值表

在计算完 AQI 指数并填入数据集中后，再对异常的质量等级进行处理，将 AQI 指数所对应的质量等级填入后，这样对数据的缺失值和异常值的处理就完成了。

3.2 数据转换

在对数据集进行缺失值和异常值处理后，便可以开始对数据集进行数据转换了，由于计算机无法对字符串类型的数据进行数据处理，因此我们需要将数据转换成计算机可以识别的数据类型，因此需要对质量等级一列进行数据转换，因为质量等级的取值没有大小意义，所以这里使用 pd.get_dummies()函数以独热编码方式对数据进行转换，将质量等级一列拆分成 6 列，每列分别对应一个等级，对应的列为 1，其他列为 0，以便计算各特征的关联度。在对数据集进行数据转换后整个数据集有 14 列 1779 行(如图 7)

	日期	AQI	PM2.5 浓度	PM10 浓度	SO2 浓度	CO 浓度	NO2 浓度	O3_8h 浓度	质量等级_严重污染	质量等级_中度污染	质量等级_优	质量等级_良	质量等级_轻度污染	质量等级_重度污染
0	2014-12-31	130	99	121	31	1.1	19	59	0	0	0	0	1	0
1	2015-01-01	61	42	71	24	0.9	9	95	0	0	0	1	0	0
2	2015-01-02	72	53	73	21	0.7	21	119	0	0	0	1	0	0
3	2015-01-03	103	78	112	30	1.3	27	121	0	0	0	0	1	0
4	2015-01-04	246	196	260	38	2.5	37	74	0	0	0	0	0	1

图 7

在完成数据转换后，开始对空气质量数据进行规范化（归一化）处理，这是数据转换的一项基本工作。由于不同的参数或指标存在各自的计算单位及量纲，数据间的差异性较大，如果不对数据进行处理有可能会对数据的分析结果造成影响。为了消除这一影响，需要对各项数据进行归一化处理使各项参数之间的量纲不受影响。在对数据进行标准化处理时，需要对数据按比例缩放后分布在指定的范围内，以便对空气质量数据进行综合分析。由于数据集的各项参数不存在负值，所以选择