****

软件工程专业课程设计报告

课 程： 软件项目开发综合实训

题 目：基于python的基金爬虫可视化

年 级： 2018 级

班 级： 软件工程B2班

姓名学号： 温贤虎 201810610239

姓名学号： 陈晓光 201810610202

起讫日期： 2022年10月14日--2020年10月25日

指导教师： 彭 梅

课程设计任务书

以小组（1~3人一组）为单位，自主选题，完成课程考核作业，要求如下：

综合应用所学的专业知识进行项目设计，所设计的项目工作量符合综合项目开发要求比如网站方向有前端后台，人工智能方向有数据挖掘分析可视化及聚类分析等，所设计项目与生产实际应用相符，有一定的商业价值和自己特色，并且所撰写文档逻辑清晰层次分明，字数达6000字及以上。

格式要求：一级标题：（标题1样式格式为小三号、加粗、中英文均为黑体，段落居中、段前段后间距0行，行距固定20磅，后空一行，二级标题：（标题2样式格式为四号，中英文均为黑体；顶格无缩进，段前段后间距0行，1.5倍行距），正文：宋体小四

|  |
| --- |
| **课程设计报告评语及成绩** |
| **等级（成绩）：**   **签名：** |

目录

[**1 绪论** 1](#_Toc85707737)

[**1.1** **课题的研究背景及意义** 1](#_Toc85707738)

[**1.2** **国内外发展情况** 1](#_Toc85707739)

[**1.3 课题的主要研究内容** 2](#_Toc85707740)

[**1.3 环境介绍** 3](#_Toc85707741)

[**2 相关技术介绍** 3](#_Toc85707742)

[**2.1 技术理论论述** 3](#_Toc85707743)

[**2.1.1 网络爬虫介绍** 3](#_Toc85707744)

[**2.1.2 数据清洗介绍** 4](#_Toc85707745)

[**2.1.3 可视化分析介绍** 4](#_Toc85707746)

[**2.1.4 机器学习预测算法介绍** 4](#_Toc85707747)

[**3 系统分析** 5](#_Toc85707748)

[**3.1 开发背景** 5](#_Toc85707749)

[**3.2 需求的分析** 5](#_Toc85707750)

[**4 设计思想与实现** 6](#_Toc85707751)

[**4.1 基金代码和数据的提取** 6](#_Toc85707752)

[**4.1.1 基金代码** 6](#_Toc85707753)

[**4.1.2 基金数据** 7](#_Toc85707754)

[**4.2 基金数据的处理** 8](#_Toc85707755)

[**4.3 基金的可视化** 9](#_Toc85707756)

[**4.4 基金的预测** 11](#_Toc85707757)

[**4.4.1导入训练集和测试集的数据** 13](#_Toc85707758)

[**4.4.2 将输入转化成[样本数，时间步长，特征数]** 13](#_Toc85707759)

[**4.4.3 将输出转化成[样本数，特征数]** 13](#_Toc85707760)

[**4.4.4做出训练模型** 14](#_Toc85707761)

[**4.4.5做出评估模型** 14](#_Toc85707762)

[**4.4.6进行预测和可视化** 14](#_Toc85707763)

[**4.5 结论** 16](#_Toc85707764)

[**5 总结** 16](#_Toc85707765)

**1 绪论**

本章主要介绍基金的发展背景，及选择该课题的研究目的和意义，并且分析国内外的基金形势来给出基金爬取的意义。

* 1. **课题的研究背景及意义**

随着科学技术的不断发展，大数据应用已经不断渗入我们的基本生活。据统计2021年在支付宝上的购买基金人数已经达到了6亿人数。也就说明我国大概两人中就有一人在支付宝上买了基金。我国购买基金的人口基数这么大，就说明研究的该课题是非常有意义的。虽然购买基金的人数众多，但也不意味着每个人都能投资收益为正数，有盈利就有亏，这是基金市场上的一种常态

证券投资基金作为一种特殊的证券投资方式，是金融资本、投资家的人力资本和高科技项目的融合，需要足够的、便利的资金支持和顺畅的退出渠道来变现自己的投资。证券投资基金也是证券市场众多投资者的一员，在相关金融工具和市场不完备的情况下要想促进证券投资基金和证券市场的发展，必须赋予证券投资基金应有的功能，使其能够享受到和其他投资者一样的待遇。对于火爆的基金市场，我们现在开始研究它数据中的各个特征的含义，并且可以给客户做一个推荐，这能使不懂基金的含义的客户能够更好的了解基金并给客户推荐该基金是否买入。

* 1. **国内外发展情况**

我国基金管理业起步较晚，我国真正意义上发行的第一只[证券](https://bg.qianzhan.com/report/detail/adcba327b2744b79.html" \t "_blank)投资基金是1991年成立的“武汉证券投资基金”。在随后的发展过程中先后经历了野蛮生长、严监管规范以及次贷危机冲击等一系列阶段，目前进入到了多元化发展阶段。

基金可投资范围从股票市场扩展到了[债券](https://bg.qianzhan.com/report/detail/19bdb495171c4920.html)市场和货币市场，基金品种也从股票型基金发展到债券型基金、混合型基金、货币市场基金等。2013年我国第一只互联网货币基金“余额宝”上线，我国基金管理业业态迎来互联网新模式。与我国基金管理业发展历史基本同步，其对应的监管背景也可分为三个阶段。2015年至2017年，基金业监管法律和政策频繁出台，监管涉及的内容基本覆盖了基金运营的全过程。目前我国基金业监管框架已经基本搭建完成，监管规范化逐步加强。随着资管新规的颁布以及国内[金融](https://bg.qianzhan.com/report/zt_jinrong/)市场对外开放稳步推进，基金行业也面临着新的发展契机和挑战。

国外也对基金的研究有很多：国外对证券投资基金环境分析方面已经做了很多相关研究，具体如下：1952 年Markovitz 均值方差理论及投资组合理论，标志着现代证券基金投资理论的开始。Shieifer和Waldman(1990)，Froot，Scharfstein和 Stein(1992)，Lun Marchesi(1999)对非理性所导致的羊群行为进了研究；Cannella Fraser and Lee（1995）证明，经理人市场有能力区分经理经营业绩的好坏；

传统的基金公司都是基于既往历史数据，根据基金经理的判断，构建出符合一类人需求的资产组合，但是机器学习技术的发展，能够更好地了解每个投资者的需求，根据每个投资者特定的需求为其配置最合适的资产组合，通过这种方式构建的基金，无疑更加符合投资者的需求。因此，选择该课题是顺应时代的发展。

**1.3 课题的主要研究内容**

互联网的发展推动着人们对基金交易的热情，但大部分的人对基金还是一知半解，因此我们选择了基于python的基金爬虫及可视化这一课题开始研究。

本课题主要研究课题所用的技术，基金的爬虫，基金的数据清洗，基金的可视化和基金的预测四个方面：

（1）基金的爬虫；在现在生活的大数据时代，大数据无处不在，无论是在网上购物还是个人刷抖音，我们的数据都是在网上时刻传输着。在这大环境下，数据的爬虫就非常的有必要，这我对基金进行数据的挖掘，也就是人们所说的爬虫。

（2）基金的数据清洗；在爬虫出来的基金数据后需要对自己无用的、空值的还有杂乱的等等数据来进行清洗，以获取一定规律的数据，为下方的基金可视化来做准备。

（3）基金的可视化；当我们有了数据，我们需要更加客观的观察基金数据的各类数据，因此我们需要用可视化来展现数据，这也可以客观的观察每一类型数据的总体趋势，在对自己认为合适的基金来进行提取数据，提前为基金的预测准备好数据。

（4）基金的预测；这是本次课题研究的重点，预测数据最主要的是算法，本次我才用的是神经网络中的LSTM算法，主要就是对数据循环来进行训练，在对测试集来进行验证其数据。对预测的数据和本来的数据进行对比，再得出结论。

**1.3 环境介绍**

**1.3.1 基本环境**

在window10的系统下，以python为开发的基本语言，通过anaconda的jupyter工具来进行。安装包pandas，numpy，requests，BeautifulSoup，json等等。利用python语言的便于快速开发的特性，来设计基于python的基金爬虫数据处理，可视化等。

**1.3.2 安装anaconda**

在anaconda官网下载最新的exe，然后双击.exe运行安装到C:\ProgramData目录下，在安装选择上选择安装python。这样就可以运行

**2 相关技术介绍**

**2.1 技术理论论述**

**2.1.1 网络爬虫介绍**

网络爬虫是一个自动提取网页的程序，它为[搜索引擎](https://baike.baidu.com/item/%E6%90%9C%E7%B4%A2%E5%BC%95%E6%93%8E)从万维网上下载网页，是搜索引擎的重要组成。传统爬虫从一个或若干初始网页的URL开始，获得初始网页上的URL，在抓取网页的过程中，不断从当前页面上抽取新的URL放入队列，直到满足系统的一定停止条件。聚焦爬虫的工作流程较为复杂，需要根据一定的[网页分析算法](https://baike.baidu.com/item/%E7%BD%91%E9%A1%B5%E5%88%86%E6%9E%90%E7%AE%97%E6%B3%95/15667054" \t "_blank)过滤与主题无关的链接，保留有用的链接并将其放入等待抓取的URL队列。然后，它将根据一定的搜索策略从队列中选择下一步要抓取的网页URL，并重复上述过程，直到达到系统的某一条件时停止。另外，所有被爬虫抓取的网页将会被系统存贮，进行一定的分析、过滤，并建立索引，以便之后的查询和检索；对于聚焦爬虫来说，这一过程所得到的分析结果还可能对以后的抓取过程给出反馈和指导。

**2.1.2 数据清洗介绍**

数据清洗的概念是对数据进行审查和校验的过程，目的是在于删除重复值和纠正存在的错误，在进行数据的一致性

数据清洗用通俗的话来说就是把自己不需要的数据去除掉，是发现并且纠正数据文件中可识别错误的最后一道程序，包括检查数据一致性、处理无效值和缺失值等。从网路爬虫中的数据大多数都是杂乱无章的，这样就不可避免有错误的数据和冲突的数据等等。这些有错误的和冲突的数据显然是我们不想要的。故此我们要按照一定的规律将数据进行清洗，这就是数据清洗。

**2.1.3 可视化分析介绍**

可视化分析是一种分析仪，主要应用于海量数据关联分析，可辅助人工操作将数据进行关联分析，并做出完整的分析图表。

主要应用于海量数据关联分析，由于所涉及到的信息比较分散、数据结构有可能不统一，而且通常以人工分析为主，加上分析过程的非结构性和不确定性，所以不易形成固定的分析流程或模式，很难将数据调入应用系统中进行分析挖掘。借助功能强大的可视化数据分析平台，可辅助人工操作将数据进行关联分析，并做出完整的分析图表。图表中包含所有事件的相关信息，也完整展示数据分析的过程和数据链走向。同时，这些分析图表也可通过另存为其他格式，供相关人员调阅。

**2.1.4 机器学习预测算法介绍**

这里我们采用TensorFlow提供Python语言版本：CPU版本（tensorflow），它是由谷歌人工智能团队谷歌大脑（Google Brain）开发和维护，被广泛应用于各类机器学习算法的编程实现，其前身是谷歌的神经网络算法库DistBelief。

长短期记忆网络（Long-Short Term Memory,LSTM）论文首次发表于1997年。由于独特的设计结构，LSTM适合于处理和预测时间序列中间隔和延迟非常长的重要事件。

LSTM的表现通常比时间递归神经网络及隐马尔科夫模型（HMM）更好，比如用在不分段连续手写识别上。2009年，用LSTM构建的人工神经网络模型赢得过ICDAR手写识别比赛冠军。LSTM还普遍用于自主语音识别，2013年运用TIMIT自然演讲数据库达成17.7%错误率的纪录。作为非线性模型，LSTM可作为复杂的非线性单元用于构造更大型深度神经网络。

**3 系统分析**

**3.1 开发背景**

Python程序设计语言是在近几年开始流行的一种编程语言，它以简单易懂、适合于快速开发、能够跨平台操作吸引人们的关注。经过最新的排名，python的目前地位已经超过了Java，c。这就证明了python有独到的一面，，其固有的可靠性与安全性不仅令程序员放心。Python内建了对网络编程，数据库连接、多线程等高级程序设计任务的支持。

目前，在软件开发活动中，无论何种服务，应用系统的开发仍是开发的主题，对于爬虫的方面，能够快速获取数据，并且对数据进行数据的处理。

**3.2 需求的分析**

传统金融行业受互联网快速发展的驱使，不断自我互联网化和数字化。反之，传统金融行业又对互联网理财平台的丰富和创新具有促进的意义。 因此，不论是传统银行/券商/基金公司的移动化，还是互联网理财平台的“余额宝”、“零钱通”和各类基金产品不断丰富，都给互联网投资理财带来了多元化的选择，促使了整个行业迎来爆发式的发展。 任何一个行业或商业的长足、稳定、健康的发展，都离不开国家的政策、经济、社会文化和基础技术环境。

证券投资基金具有集合投资,分散风险,专业理财等特点,近年获得了快速发展,其在我国证券市场扮演的角色也越来越重要。基金市场中不同的基金品种在风险收益水平上具有较大的差别，即使同一类型的基金由于投资管理水平的差异也使风险收益的差距非常明显。因此，如何选择合适的基金进行投资正逐渐成为广大投资者关注的问题。与股票投资和债券投资一样，基金投资者在进行投资时，首先要做的工作是想清楚自己投资的目标是什么?如果投资者追求的是短期投资目标，并且不想承担太大的投资风险，那么债券型基金和货币市场基金是相对合适的基金品种;如果投资者追求的是长期资本增值，可以承担较大的投资风险，那么股票型基金和平衡型基金是正确的选择。

在中国的资本市场，以人工智能为基础的投资技术以独特的优势正式走向前台。在广发基金网站上发布的“广发百发策略价值基金”的有关信息来看，百度与广发基金合作，正在用智能投资技术，多元化整体提升投资效能、发掘市场价值。百度在人工智能领域一直走在国内的前沿，到人工智能遇上金融，我能看到了中国正在人工智能的投资时代。

在顺应时代发展的前提下，在人们对基金的逐步接受，买基金的人数也开始日益剧增，本课题就是让那些什么都不懂的基金投资者来了解基金的结构，并且对基金的数据进行提取，更将客观的展现基金的各类数据，并且使用人工智能大数据算法来预测基金接下来的走势，根据结果给客户形成一个是否推荐该基金的建议。

**4 设计思想与实现**

现在的网站大多使用瀑布流的json方法来加载，因此在爬取这些信息时，要先对网页进行动态分析，找出对应的json连接后，读取对应的网页内容。本次爬虫主要在anaconda的jupyter上进行。用的是requests和json包来获取网站上的json数据，这是获取基金数据的过程。

然后就是数据处理，数据处理采用的是python中的pandas包，pandas 是一个开放源码、BSD 许可的库，提供高性能、易于使用的数据结构和数据分析工具。接着就是可视化，可视化采用的是pyecharts模块和matplotlib模块Matplotlib 是 Python 的绘图库。 它可与 NumPy 一起使用，提供了一种有效的 MatLab 开源替代方案。

最后是数据预测，这将采用机器学习的sklearn库，这个下面会详细介绍。

**4.1 基金代码和数据的提取**

**4.1.1 基金代码**

最通常的方法是通过r=request.get（url）构造一个向服务器请求资源的url对象。这个对象是Request库内部生成的。这时候的r返回的是一个包含服务器资源的Response对象。包含从服务器返回的所有的相关资源。可以看到，我们将通过一个接口<http://fund.eastmoney.com/js/fundcode_search.js>

利用requests.get来获取数据，并用正则表达式

re.findall(r'"(\d\*?)","(.\*?)","(.\*?)","(.\*?)","(.\*?)"',num.text)

获取基金的代码、基金的名称、基金的类型。如表4-1所示。

表4-1 基金代码



**4.1.2 基金数据**

再一次利用requests.get()方法，利用蛋卷网站的url进行数据的爬取（<https://danjuanapp.com/djapi/fund/derived/>"加上基金的代码）得出end\_d=[]#日期unit\_n=[]#净值nav\_g1=[]#一个月涨幅nav\_g3=[]#三个月涨幅nav\_g6=[]#六个月nav\_gt=[]#今年以来涨幅的数据

还有进行数据的预处理，如果还没发售的基金是没有data这个数据的，所以先要进行数据的预处理，将没有data数据的基金剔除掉。再经过简单的数据预处理，就可以的出基金的各类信息，总共有9885条基金的数据。如表4.2所示

表4-2 基金数据



**4.2 基金数据的处理**

首先利用fund\_sj=sj[sj['值']!="None"]这行代码是将sj值这一列数据不为“None”的值赋值给fund\_sj。将数据有None的数据全部剔除掉，为了数据的美观，用.round（2）方法保存两位小数得到8769条。抽取基金代码、基金名称、基金类型、基金当天日期、基金当天净值、基金近一个月涨幅、基金近三个月涨幅、基金今年以来涨幅，如字段表4-3所示

表4-3 基金数据

|  |  |  |
| --- | --- | --- |
| 字段名称 | 数据类型 | 备注 |
| 代码 | object | 基金代码 |
| 名称 | object | 基金名称 |
| 类型 | object | 基金类型 |
| 当天日期 | object | 基金当天日期 |
| 当天净值 | float | 基金当天净值 |
| 近一个月涨幅 | float | 基金近一个月涨幅 |
| 近三个月涨幅  今年以来涨幅 | float  float | 基金近三个月涨幅  基金今年以来涨幅 |

将所有数据按类型分好类（股票型，混合型-偏股，混合型-灵活，混合型-偏债，混合型-平衡，债券型-可转债，债券型-长债，债券型-中短债，债券型-混合债，指数型-股票,QDII，商品（不含QDII)，货币型），用此fund\_gp=fund\_sj[fund\_sj['类型']=="类型"]一行代码。并将其保存在类型的文件夹。得出如下数据，如图4-1所示

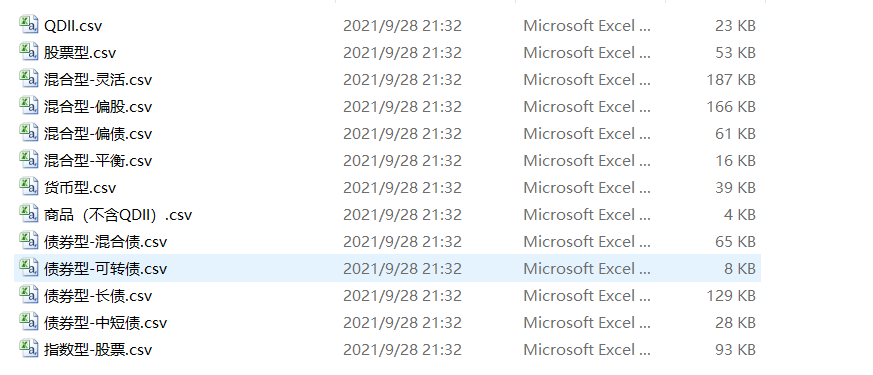


图4-1 分类的csv

**4.3 基金的可视化**

获取各基金的类型分布，得到柱形图如图4-2所示

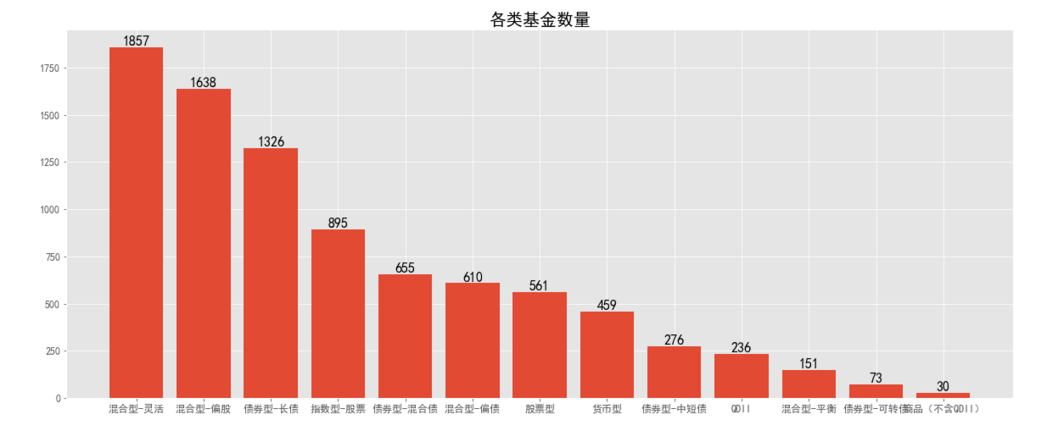


图4-2 类型分布柱形图

由上图可分析到混合型-偏股、混合型-灵活、指数型-股票、股票型四种类型的最多，所以我们选取混合型-偏股、混合型-灵活、指数型-股票、股票型这四种（由于债券型基金没有参考价值，pass）作为分析。

提取该四种基金类型的前十名，以便于做可视化分析，本课题采用的是有三种表（柱形图，折线图，散点图）做出四张小图，如图4-3所示

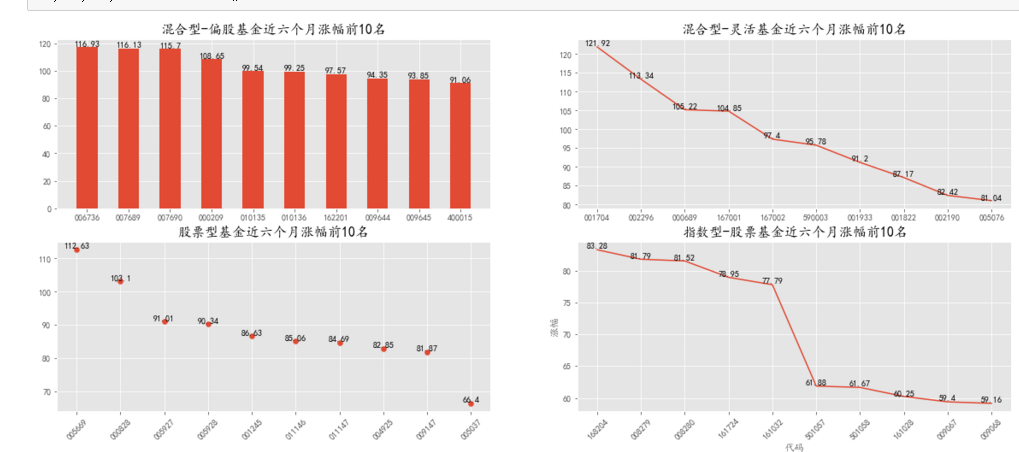


图4-3 类型小图

由上图可以分析出混合型-偏股、混合型-灵活、指数型-股票、股票型这四种的近六个月涨幅第一名的分别为代码001704，05669，006736，168204的四个基金，因此本课题就需要分析这四个基金的涨幅，使用基金的近一个月涨幅、近三个月涨幅、近六个月涨幅的数据为它们做出可视化分析，做出各类基金的涨幅第一的各阶段涨幅。如图4-4所示

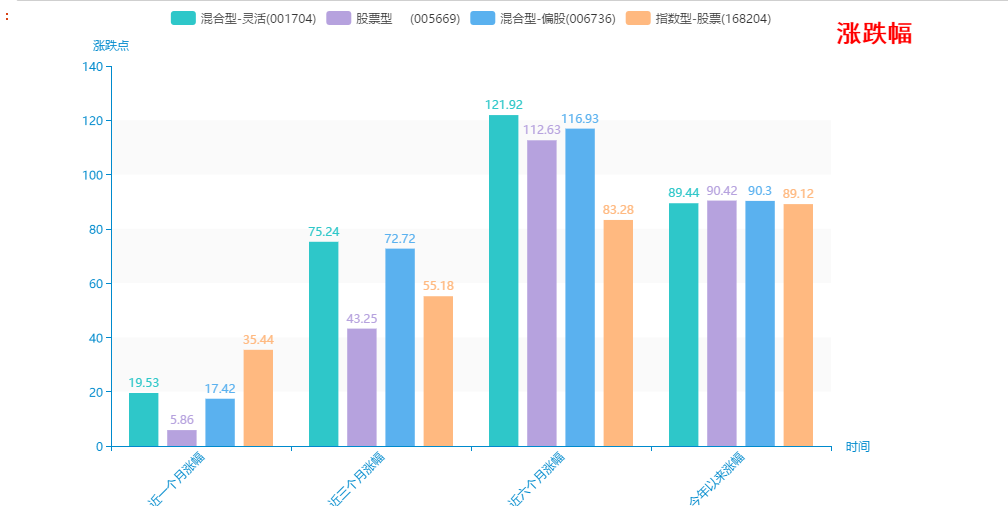


图 4-4 各类基金的涨幅第一

由上图可以观察出，股票型的005669近一个月和近三个月的涨幅都不是很理想，但在基金市场中，这反而是后续的涨幅是更好的，所以就此可以选择05669这个基金，选择性定投。

选取的四个基金数据过多，因此将四个基金选取前三十行数据，做出其折线图。

做出国投瑞银进宝灵活配置混合（001704）和国投瑞银先进制造混合（006736），还有前海开源公用事业股票（005669）和中融中证煤炭指数(LOF)（168204）的折线图如图4-5和图4-6所示

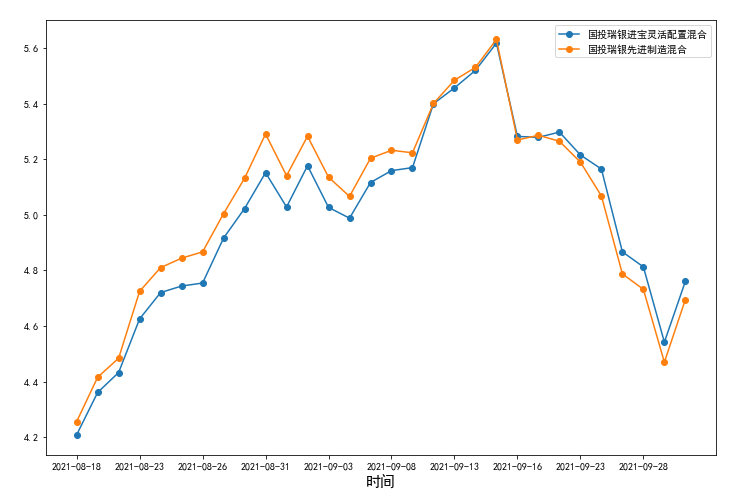


图 4-5 001704和006736折线图

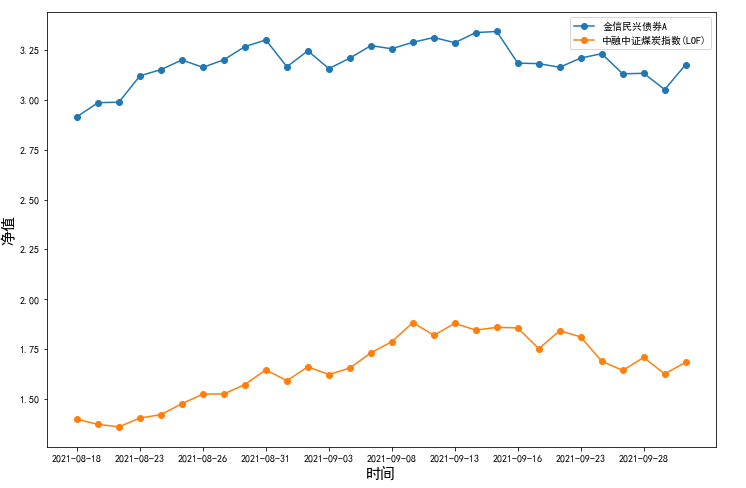


图 4-6 005669和168204折线图

**4.4 基金的预测**

LSTM是一种含有LSTM区块（blocks）或其他的一种类神经网络，文献或其他资料中LSTM区块可能被描述成智能网络单元，因为它可以记忆不定时间长度的数值，区块中有一个gate能够决定input是否重要到能被记住及能不能被输出output。

图1底下是四个S函数单元，最左边函数依情况可能成为区块的input，右边三个会经过gate决定input是否能传入区块，左边第二个为input gate，如果这里产出近似于零，将把这里的值挡住，不会进到下一层。左边第三个是forget gate，当这产生值近似于零，将把区块里记住的值忘掉。第四个也就是最右边的input为output gate，他可以决定在区块记忆中的input是否能输出。

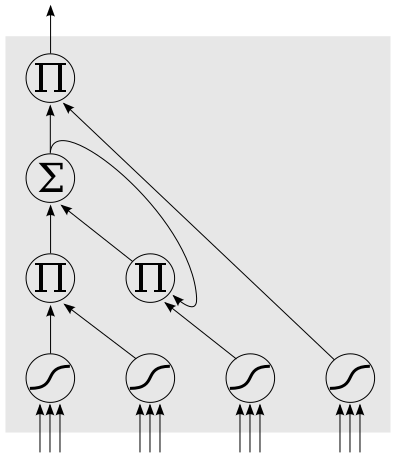
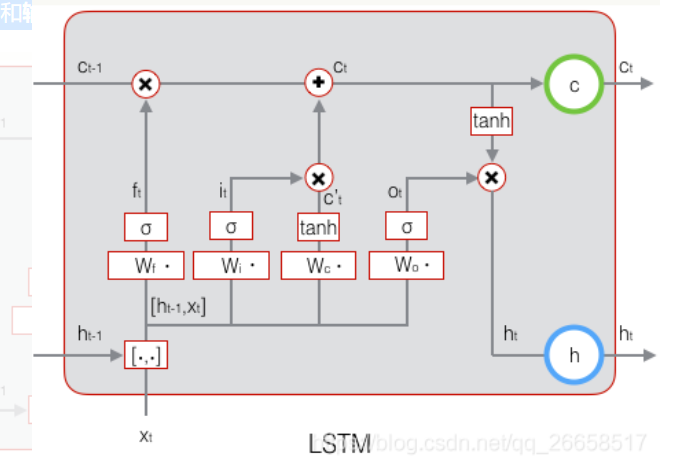
[](https://baike.baidu.com/pic/%E9%95%BF%E7%9F%AD%E6%9C%9F%E8%AE%B0%E5%BF%86%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/17541107/0/14ce36d3d539b600564bd957e550352ac75cb74d?fr=lemma&ct=single)

图1 LSTM模型

LSTM有很多个版本，其中一个重要的版本是GRU（Gated Recurrent Unit），根据谷歌的测试表明，LSTM中最重要的是Forget gate，其次是Input gate，最次是Output gate。

LSTM（长短时记忆网络）神经网络的训练方法：

长短时记忆网络在RNN的基础之上添加了长期依赖信息，主要是在每一个cell（即每一个时刻的输入单元）增添了三个门控----遗忘门、输入门和输出门。整个前向传递过程如下图所示：



LSTM 结构图

最左边的 f\_{t} 是遗忘门，遗忘门用于选择需要记住多少信息

https://img-blog.csdnimg.cn/20191008133348441.png

W_{f}是遗忘矩阵，由0-1组成，以此来决定保留多少当前时刻的状态，C_{t-1}含有前t-1个时刻的信息。

**中间i_{t}**代表输入门，其中的W{_{i} 和W_{c} 分别是输入矩阵和状态矩阵

https://img-blog.csdnimg.cn/2019100813414990.png

https://img-blog.csdnimg.cn/20191008134203379.png

经过当前时刻单元的记忆状态信息C_{t}（有时也被叫作调制输入u_{t}）即为：

https://img-blog.csdnimg.cn/20191008134233930.png

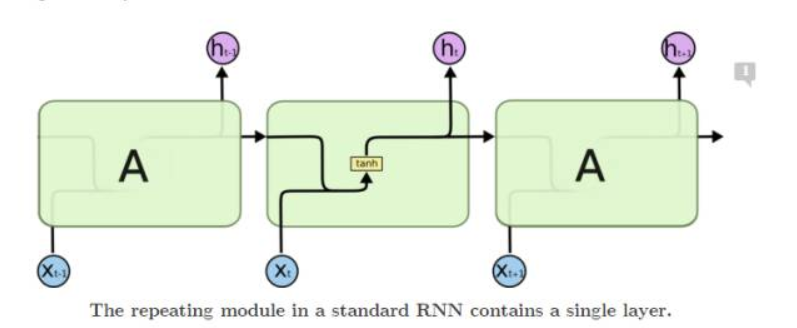
**最右边**的代表输出门W{_{o}是输出矩阵

https://img-blog.csdnimg.cn/2019100813451592.png

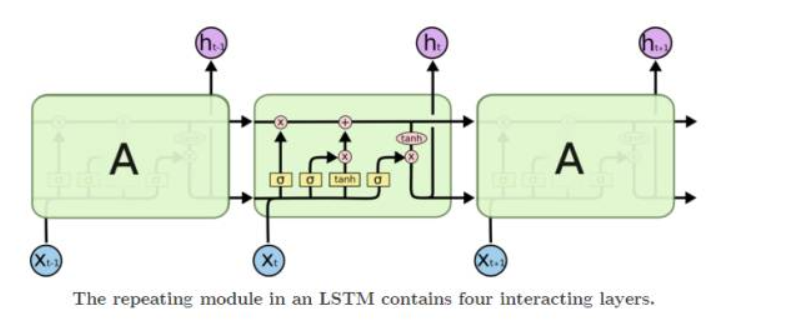
https://img-blog.csdnimg.cn/20191008134607607.png

其中符号 \circ 表示按元素乘（对应位置元素相乘），[A,B]代表A,B的增广矩阵，\sigma代表sigmoid激活函数。

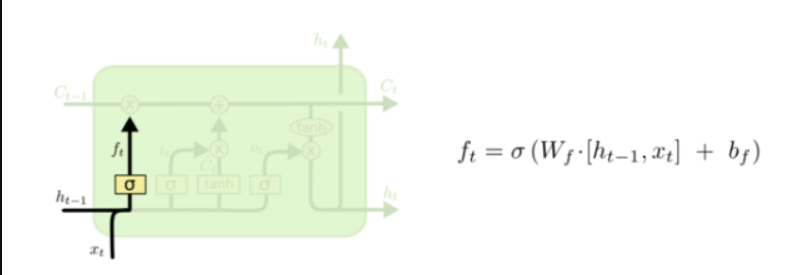
LSTM 刻意的设计去避免长期依赖问题。记住长期的信息在实践中RNN几乎默认的行为，但是却需要很大的代价去学习这种能力。所有的递归神经网络都有重复神经网络模块的链式形式，在标准的RNNs中，重复模块只有一个非常简单的架构，比如只有一个tanh层。

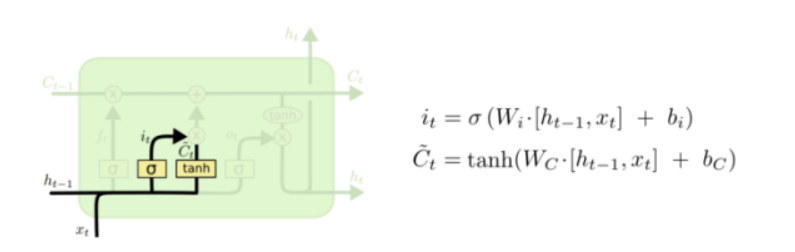


LSTM同样也是这种链式结构，但是重复的模型拥有不同的结构，它与单个的神经网层不同，它有四个， 使用非常特别方式进行交互。



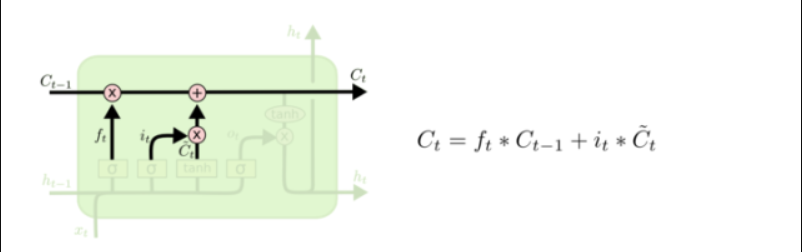
LSTM的第一步是决定我们要从细胞状态(cell state)中丢弃哪些信息。这个决定是由一个sigmoid层叫“忘记门限层”所决定的。它的输入为ht-1和xt, 输出0到1间的数到细胞转态(cell state)Ct−1。其中1表示“完全的保留”而0表示“完全的放弃”。



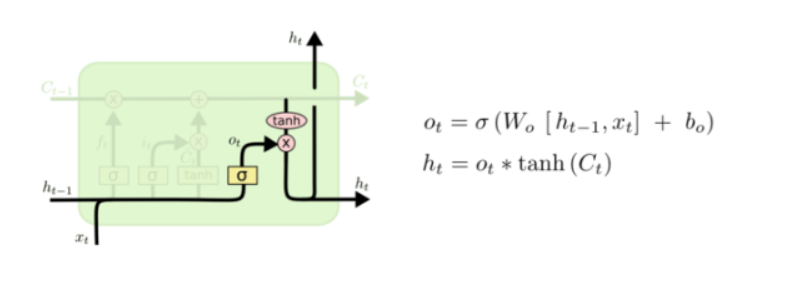
下一步就是决定哪些新的信息需要存储在细胞状态中。这包含两部分：第一部分，一个sigmoid层叫“输入门限层”决定哪些值我们需要去更新。第二部分，一个tanh层，创建一个新的候选向量C^t,会被加入到这状态上。下一步，我们将组合这两个部分产生一个新的更新状态。在我们的语言模型中，我们想添加这新代词的主语类别到这细胞状态，去取代我们要忘记的老的代词。

现在是时候去更新老的细胞状态Ct−1, 到一个新的细胞状态Ct。前面的步骤已经决定了将要做什么，我们仅仅需要去实现它。

我们将旧状态与ft相乘，需要丢弃的我们需要更早的丢弃。然后加上it\*Ct。 这就是这新的候选值，规格化了我们需要去决定有多少状态值需要去更新。



但是也有一个过滤后的版本。首先，我们运行sigmoid层，这层决定细胞转态哪些部分需要输出。然后，我们把细胞状态通过tanh(输出-1到1之间的值)进行处理，然后再乘上sigmoid输出。最后我们仅仅会输出我们确定输出的那部分。在语言模型例子中，因为它仅仅看到了一个代词，也许我们想去输出与动词相关的信息，在这种情况下会发生什么呢？例如，也许输出是否这代词时单数还是复数。如果是动词的话，我也知道动词是否需要进行词形的变化。



首先，引入所有必须的lib。Keras在建立神经网络的时候非常非常方便。由于是对于基础简单的神经网络。并且我们从Keras.dataset 中引入MNIST数据库。同时，我们从Keras自带模型中导入Sequential，在导入Sequential之后本课题在后面接着引入各种层。将建立一个Sequential Model并且增加卷积层、池化层。同时可以在每层增加了Dropout层。最后增加了Dense层，并且使用的是Softmax可以用于分类。用来建造模型build\_model()。

通过上面的基金可视化就可以找到四种基金类型的前三名数据如图4-7所示

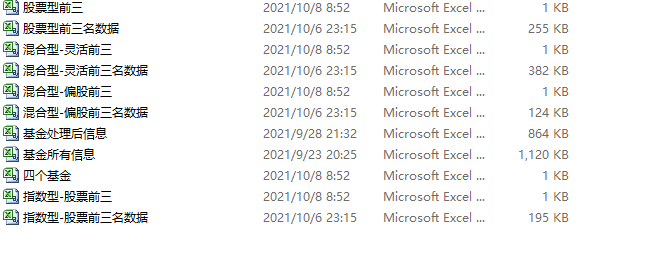


图4-7 种基金类型的前三名数据

有了数据就可以开始预测

本课题采用的是Tensorflow的LSTM模型，因为对于基金净值的有序数据来说，采用循环神经网络更加切合。

**4.4.1导入训练集和测试集的数据**

测试集的数据用的是前海开源公用事业股票（005669），训练集的数据采用的是泰达转型机遇股票A（000828）创金合信新能源汽车股票A（005927）还有就是混合型-偏股类型的国投瑞银先进制造混合（006736）三个基金的数据，其原理就是参考三个基金的数据来对目标基金的数据进行预测。并将其存取csv文件中。

**4.4.2 将输入转化成[样本数，时间步长，特征数]**

采用这两行代码X\_train = X\_train.reshape(-1, time\_step, input\_size)

X\_validation = X\_validation.reshape(-1, time\_step, input\_size)

.reshape函数是MATLAB中将指定的矩阵变换成特定维数矩阵一种函数，且矩阵中元素个数不变，函数可以重新调整矩阵的行数、列数、维数。将测试集的数据转化为样本数，时间步长，特征数，为预测做准备工作

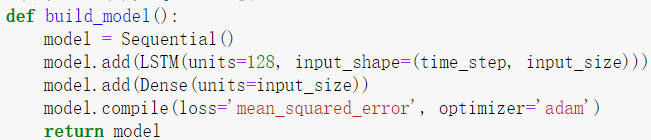
**4.4.3 将输出转化成[样本数，特征数]**

采用这两行代码y\_train = y\_train.reshape(-1, input\_size)

y\_validation = y\_validation.reshape(-1, input\_size) 将测试集的数据转化为样本数，特征数，为预测做准备工作

**4.4.4做出训练模型**

训练的模型利用的是Sequential()， Keras库中有有两种类型的模型，序贯模型（Sequential）和函数式模型（Model），函数式模型应用更为广泛，序贯模型是函数式模型的一种特殊情况。序贯模型（Sequential):单输入单输出，一条路通到底，层与层之间只有相邻关系，没有跨层连接。这种模型编译速度快，操作也比较简单，所以本课题选取Sequential()模型作训练。构建build\_model()。如下图所示



model = build\_model()

model.fit(X\_train,y\_train,epochs=epochs,batch\_size=batch\_size,verbose=1, validation\_split=0.25, shuffle=True)

**4.4.5做出评估模型**

利用train\_score = model.evaluate(X\_train, y\_train, verbose=0)

validation\_score = model.evaluate(X\_validation, y\_validation, verbose=0)将该训练的模型进行评估，得出的数据可以很好的说明该模型的准确度，得出的结果为



分数是越趋于0，就说明训练集和测试集越好，由我们得出的分数看出训练集的分数是不错的，但测试集的分数就不行，但最主要的是看预测的数据和本身数据的差距才能说明该模型的好坏。

**4.4.6进行预测和可视化**

采用这行代码predict\_validation = model.predict(X\_validation)，将预测数据存入predict\_validation里面，可以将预测的7天数据和实际的数据打印出来进行对比。得出图4-8

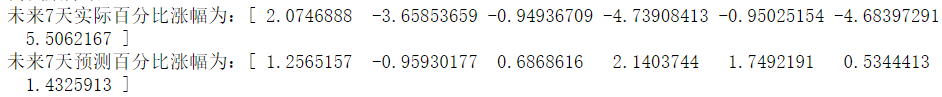


图 4-8 数据对比

由上图可以分析，对涨幅的正负来看，3/7的概率预测是准确的，这也符合了基金市场瞬息万变的规律，再看看下方的可视化图，处理后的图为图4-9所示。

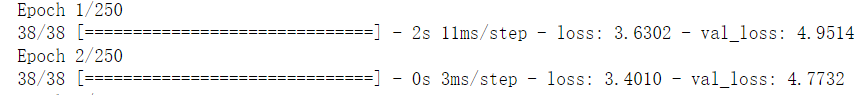


图 4-9

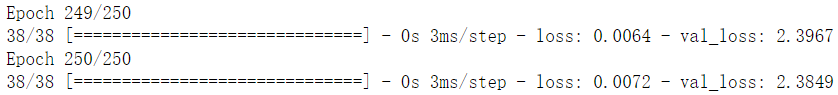
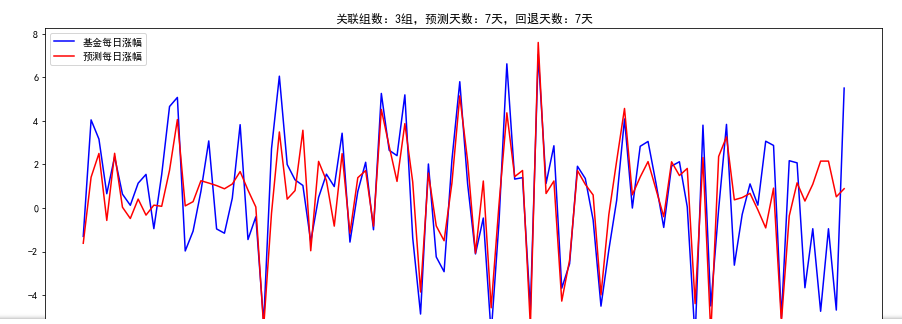


图4-10

由上图4-10和图4-11所示，其中loss是训练集上的损失，val\_loss是测试集上的损失。首先是loss从一开始的3.6484降至0.0097，然后是val\_loss由4.9909降至2.5735.由此可以看出这个模型是能发挥作用的。但val\_loss测试后的值还是有点高，这就证明测试集的数据还是要优化的。

图 4-11 基金预测涨幅图

如上图4-11所示，图中主要看基金的每日涨幅和预测的涨幅的重合拟合度，这是判断该模型预测的数据重要依据。图中的重合拟合度基本上重合，预测的数据也基本正常。

净值的图对比，如图4-12所示。

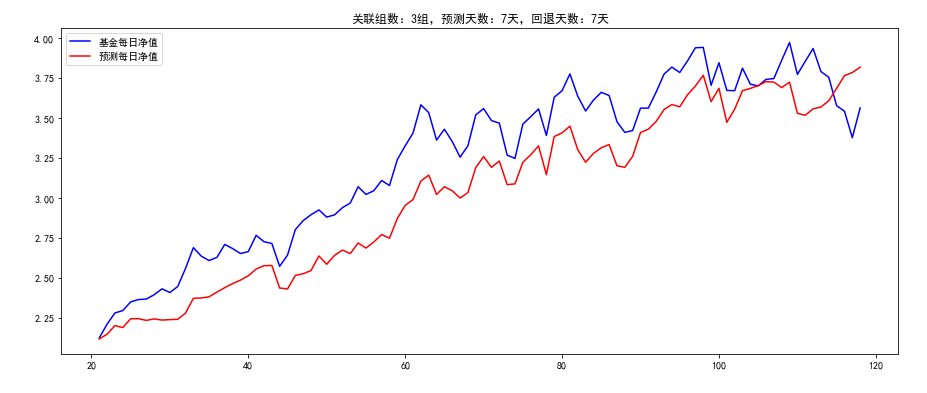


图4-12 净值预测图

由上图可以看出，预测的净值和基金的每日净值形状类似，证明涨幅也是类似的。

**4.5 结论**

通过神经网络LSTM算法，得出预测的数据和实际数据的对比图，可以看出预测的数据还是没那么准确。但也已经可以经过该数据给用户来进行参考，但这也仅仅只是参考，基金的市场是瞬息万变的，没有人可以说明天的会发生什么，以我的经验推荐该基金还是可以买入的。

**5 总结**

通过本学期的学习，我对数据的爬虫，数据的清洗，数据的可视化有了更深层的了解。

个人觉得特别喜欢Python的地方是对字符串操作特别的灵活、采取缩进的方式简单明了、以及简单的语法。其内嵌数据库中有很多常用函数，需要使用时调用即可，这一点让我觉得相当方便，可以说少写了一半代码。其简洁方便的编辑方式也很对我胃口，相对比所学过的其他语言来说确实很灵活。

在选择本次课题时，我们团队选择了基于python的基金爬虫可视化，本身我也是基金的购买者，因此我迫切的想了解基金的知识，所以我利用request.get()方法对基金的数据进行爬虫，再使用数据清洗的方法对数据的空值，None值，不需要的数据清洗掉，得出的数据再进行可视化分析，在这一步就能找到往年来涨幅最好的几只基金，在对其进行预测，得出的结果可以很好的反应该基金是否值得购买，本课题预测的是代码为005669的前海开源公用事业股票，得出的结果还算不错，所以给出是推荐购买的建议。但这也仅仅只是建议，因为在基金市场是瞬息万变的，谁也不能知道明天会发生什么，所以也需要客户具有自行判断的能力。

在这些过程中，我不仅知道了职业生涯所需具备的专业知识，而且让我深深体会到一个团队中各成员合作的重要性，要善于团队合作，善于利用别人的智慧，这才是大智慧。靠单一的力量是很难完成一个大项目的，在进行团队合作的时候，还要耐心听取每个成员的意见，合作带来的往往是更完美的成果。

在对于走向工作岗位的我们更要不断加强自己的专业技能，社会不会要一个一无是处的人，所以我们要更多更快的从一个学校人向社会人转变。为此我们将会在以后的日子里继续努力，不断激励经验，不断磨砺自己，早日走向工作岗位。