# 哈爾濱工業大學

**HARBIN INSTITUTE OF TECHNOLOGY**

## 2020 年 春 季学期本科生课程考核

（结课报告）

| | | | |
|---|---|---|---|
| 考 核 科 目 | Python 金融大数据分析 | | |
| 学生所在院（系） | 计算机科学与技术学院 | | |
| 学 生 所 在 专 业 | 大数据科学与技术 | | |
| 学 生 姓 名 | 彭钰驯 | | |
| 学 号 | 1170300916 | | |
| 考 核 结 果 | | 阅 卷 人 | |

一. 运用相关库获取上市公司 A 的股票基本数据、计算相应的衍生变量数据、通过相关性分析选取合适的衍生变量、可视化呈现数据、并生成 Excel 工作簿。

1）运用 Tushare 库获取上市公司 A 从 2020-02-03 至 2020-04-01 的股价涨跌幅和前 20 分钟成交量数据。

| 日期 | 名称 | 股价涨跌幅(%) | 20分钟成交量 |
|---|---|---|---|
| 2020-04-01 | 浪莎股份 | -1.35 | 1668.0 |
| 2020-03-31 | 浪莎股份 | -0.43 | 1145.0 |
| 2020-03-30 | 浪莎股份 | -6.62 | 4070.0 |
| 2020-03-27 | 浪莎股份 | -0.79 | 4377.0 |
| 2020-03-26 | 浪莎股份 | 3.61 | 1309.0 |
| 2020-03-25 | 浪莎股份 | 0.07 | 1756.0 |
| 2020-03-24 | 浪莎股份 | 3.67 | 960.0 |
| 2020-03-23 | 浪莎股份 | -3.48 | 1094.0 |
| 2020-03-20 | 浪莎股份 | 2.80 | 671.0 |
| 2020-03-19 | 浪莎股份 | 0.07 | 1152.0 |
| 2020-03-18 | 浪莎股份 | 0.56 | 988.0 |
| 2020-03-17 | 浪莎股份 | -0.07 | 2007.0 |
| 2020-03-16 | 浪莎股份 | -2.27 | 969.0 |
| 2020-03-13 | 浪莎股份 | -0.62 | 1343.0 |
| 2020-03-12 | 浪莎股份 | -2.27 | 790.0 |
| 2020-03-11 | 浪莎股份 | -0.99 | 1361.0 |
| 2020-03-10 | 浪莎股份 | 0.00 | 3078.0 |
| 2020-03-09 | 浪莎股份 | -3.82 | 1721.0 |
| 2020-03-06 | 浪莎股份 | 0.64 | 2276.0 |
| 2020-03-05 | 浪莎股份 | 4.00 | 2759.0 |
| 2020-03-04 | 浪莎股份 | 1.90 | 977.0 |
| 2020-03-03 | 浪莎股份 | 0.82 | 1268.0 |
| 2020-03-02 | 浪莎股份 | 3.11 | 2285.0 |
| 2020-02-28 | 浪莎股份 | -2.81 | 656.0 |
| 2020-02-27 | 浪莎股份 | 0.69 | 1169.0 |
| 2020-02-26 | 浪莎股份 | 0.00 | 831.0 |
| 2020-02-25 | 浪莎股份 | -2.03 | 968.0 |
| 2020-02-24 | 浪莎股份 | 0.82 | 728.0 |
| 2020-02-21 | 浪莎股份 | 0.83 | 717.0 |
| 2020-02-20 | 浪莎股份 | 1.18 | 597.0 |
| 2020-02-19 | 浪莎股份 | -0.28 | 547.0 |
| 2020-02-18 | 浪莎股份 | 1.05 | 1212.0 |

2） 采用当日交易量和昨日交易量的方法:

采用当日交易量和多日交易量的平均值的方法:

```
通过公式1计算的相关系数r值为0.10186356367833241，显著性水平P值为0.5209368691557671
通过公式2相关系数r值为-0.020186253531414608，显著性水平P值为0.8977671137976126
```

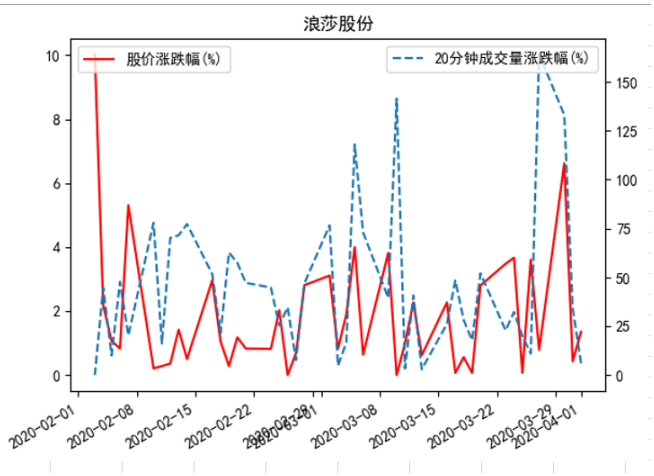
3）通过皮尔逊相关系数分别进行两种前 20 分钟成交量涨跌幅与股价涨跌幅的相关性分析，并选择出更优的前 20 分钟成交量涨跌幅。

| 日期 | 名称 | 股价涨跌幅(%) | 20分钟成交量涨跌幅(%) |
|---|---|---|---|
| 2020-04-01 | 浪莎股份 | -1.35 | -6.012284 |
| 2020-03-31 | 浪莎股份 | -0.43 | -34.201075 |
| 2020-03-30 | 浪莎股份 | -6.62 | 133.064193 |
| 2020-03-27 | 浪莎股份 | -0.79 | 164.144112 |
| 2020-03-26 | 浪莎股份 | 3.61 | -11.012916 |
| 2020-03-25 | 浪莎股份 | 0.07 | 19.945355 |
| 2020-03-24 | 浪莎股份 | 3.67 | -32.287075 |
| 2020-03-23 | 浪莎股份 | -3.48 | -22.857244 |
| 2020-03-20 | 浪莎股份 | 2.80 | -52.066293 |
| 2020-03-19 | 浪莎股份 | 0.07 | -17.840459 |
| 2020-03-18 | 浪莎股份 | 0.56 | -28.114086 |
| 2020-03-17 | 浪莎股份 | -0.07 | 48.408326 |
| 2020-03-16 | 浪莎股份 | -2.27 | -26.177053 |
| 2020-03-13 | 浪莎股份 | -0.62 | -2.923850 |
| 2020-03-12 | 浪莎股份 | -2.27 | -40.746297 |
| 2020-03-11 | 浪莎股份 | -0.99 | 3.356622 |
| 2020-03-10 | 浪莎股份 | 0.00 | 141.402298 |
| 2020-03-09 | 浪莎股份 | -3.82 | 39.482109 |
| 2020-03-06 | 浪莎股份 | 0.64 | 72.981189 |
| 2020-03-05 | 浪莎股份 | 4.00 | 118.016594 |
| 2020-03-04 | 浪莎股份 | 1.90 | -16.591967 |
| 2020-03-03 | 浪莎股份 | 0.82 | 4.806381 |
| 2020-03-02 | 浪莎股份 | 3.11 | 76.611532 |
| 2020-02-28 | 浪莎股份 | -2.81 | -46.923419 |
| 2020-02-27 | 浪莎股份 | 0.69 | -7.696463 |
| 2020-02-26 | 浪莎股份 | 0.00 | -34.664104 |
| 2020-02-25 | 浪莎股份 | -2.03 | -25.413588 |
| 2020-02-24 | 浪莎股份 | 0.82 | -44.783124 |
| 2020-02-21 | 浪莎股份 | 0.83 | -47.193990 |
| 2020-02-20 | 浪莎股份 | 1.18 | -57.465649 |
| 2020-02-19 | 浪莎股份 | -0.28 | -62.677794 |
| 2020-02-18 | 浪莎股份 | 1.05 | -21.409273 |
| 2020-02-17 | 浪莎股份 | 2.96 | 51.763617 |
| 2020-02-14 | 浪莎股份 | -0.50 | -77.260531 |

4）在进行简单数据优化后，绘制股价涨跌幅与前 20 分钟成交

量涨跌幅的可视化图表(2 个 Y 轴的形式)。

| 日期 | 名称 | 股价涨跌 | 20分钟成交量涨跌幅(%) |
|---|---|---|---|
| 2020/4/1 | 浪莎股份 | -1.35 | -6.01228 |
| ###### | 浪莎股份 | -0.43 | -34.2011 |
| ###### | 浪莎股份 | -6.62 | 133.0642 |
| ###### | 浪莎股份 | -0.79 | 164.1441 |
| ###### | 浪莎股份 | 3.61 | -11.0129 |
| ###### | 浪莎股份 | 0.07 | 19.94536 |
| ###### | 浪莎股份 | 3.67 | -32.2871 |
| ###### | 浪莎股份 | -3.48 | -22.8572 |
| ###### | 浪莎股份 | 2.8 | -52.0663 |
| ###### | 浪莎股份 | 0.07 | -17.8405 |
| ###### | 浪莎股份 | 0.56 | -28.1141 |
| ###### | 浪莎股份 | -0.07 | 48.40833 |
| ###### | 浪莎股份 | -2.27 | -26.1771 |
| ###### | 浪莎股份 | -0.62 | -2.92385 |
| ###### | 浪莎股份 | -2.27 | -40.7463 |
| ###### | 浪莎股份 | -0.99 | 3.356622 |
| ###### | 浪莎股份 | 0 | 141.4023 |
| 2020/3/9 | 浪莎股份 | -3.82 | 39.48211 |
| 2020/3/6 | 浪莎股份 | 0.64 | 72.98119 |
| 2020/3/5 | 浪莎股份 | 4 | 118.0166 |
| 2020/3/4 | 浪莎股份 | 1.9 | -16.592 |
| 2020/3/3 | 浪莎股份 | 0.82 | 4.806381 |




5) 运用 xlwings 库将优化后的数据以及可视化图表导出至 Excel 工作簿。(除了最终的大作业报告，还需提交相应源代码以及 Excel 文件)

见文件

二. 运用相关库从新浪财经网获取上市公司 A 的网页源代码，提取相应的标题、网址、来源和发布日期信息，进行简单的数据清洗，并生成文本文件。具体要求如下：

1) 运用 requests 库从新浪财经中获取上市公司 A 的网页源代码。

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dt
<html xmlns="http://www.w3.org/1999/xhtml">
<head>


<script async src="https://www.googletagmanager.com/gtag/js?id=UA-162965454-1"></script>
<script>
  window.dataLayer = window.dataLayer || [];
  function gtag(){dataLayer.push(arguments);}
  gtag('js', new Date());


  gtag('config', 'UA-162965454-1');
</script>
<meta http-equiv="Content-Type" content="text/html; charset=gb2312" />
<script src="https://www.recaptcha.net/recaptcha/api.js?render=6LcXY-gUAAAAAOn15cRtgUtGKVdf44bdOk9MKTwj"></script>
<script>
```

2）编写正则表达式获取新闻标题、网址、来源和发布日期信息。

```
p_title = '<h2><a href=".*?" target="_blank">(.*?)</a>'
p_href = '<h2><a href="(.*?)" target="_blank">'
p_date = '<span class="fgray_time">(.*?)</span>'
```

3）对获取的新闻信息进行简单的数据清洗。

```
title[i] = re.sub('<.*?>', '', title[i])
date[i] = date[i].split(' ')[1]
```

4）尝试 从新浪财经 中获取 上市公司 A 的前 5 页网页源代码，同时完成 2 和 3 中的内容（新闻信息的提取和简单数据清洗）。

```
1.2020年新疆维吾尔自治区劳动模范和先进工作者拟表彰人选名单 - 2020-06-15
https://k.sina.com.cn/article_1784473157_6a5ce64502001wec2.html?from=news&amp;subch=onews
2.负债率201%！英利进入司法重整 能否涅槃重生？ - 2020-06-15
https://finance.sina.com.cn/money/bond/research/2020-06-16/doc-iircuyvi8623240.shtml
3.沪深股市交易提示（6月15日） - 2020-06-14
https://finance.sina.com.cn/stock/relnews/hk/2020-06-15/doc-iircuyvi8497450.shtml
4.6月15日沪深两市最新交易提示 - 2020-06-12
https://k.sina.com.cn/article_1704103183_65928d0f02001rd7m.html?from=finance
5.章盟主、超短帮、小鳄鱼齐聚！王府井肥了谁？原创野马财 - 2020-06-12
https://finance.sina.com.cn/stock/relnews/cn/2020-06-15/doc-iirczymk7003631.shtml
6.恒指本周下跌1.89% 机构：寻找跌出来的机会 - 2020-06-12
https://finance.sina.com.cn/stock/stockptd/2020-06-14/doc-iircuyvi8437046.shtml
7.最新！宝安区2019年度346个商住小区评星结果出炉！ - 2020-06-12
https://k.sina.com.cn/article_1924738303_72b92cff01900s4in.html?from=news&amp;subch=onews
```

5）将 4 中的数据信息保存到一个文本文件中。（除了最终的大作业报告，还需提交相应源代码以及 txt 文件）

见文件

三. 运用相关库从东方财富个股吧中获取上市公司 A 的网页源代码，提取相应的标题、网址、来源和发布日期信息，进行简单的数据清洗，并生成文本文件。具体要求如下：

1）运用 selenium 库从东方财富个股吧中获取上市公司 A 的网页源代码。

```
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>搜索结果 - 东方财富网</title>
<!-- make at 6/17/2020, 3:32:07 PM, s_num: production, use 1ms -->
<link rel="stylesheet" href="/newstatic/css/style_reset.css">
<link rel="stylesheet" href="/newstatic/css/style_s.css">
<script src="https://hm.baidu.com/hm.js?e834f0bcb11ce14253b9eba75492b597"></script><script>
    var _hmt = _hmt || [];
    (function () {
        var hm = document.createElement("script");
        hm.src = "https://hm.baidu.com/hm.js?e834f0bcb11ce14253b9eba75492b597";
        var s = document.getElementsByTagName("script")[0];
        s.parentNode.insertBefore(hm, s);
    })();
</script>
<style>body {
  margin-top: 43px;
}
```

2）编写正则表达式获取新闻标题、网址、 来源 和发布日期等信息。

```
p_title = '<div class="news-item"><h3><a href=".*?">(.*?)</a>'
p_href = '<div class="news-item"><h3><a href="(.*?)">.*?</a>'
p_date = '<p class="news-desc">(.*?)</p>'
```

3）对获取的新闻信息进行简单的数据清洗。

```
title[i] = re.sub('<.*?>', '', title[i])
date[i] = date[i].split(' ')[0]
```

4）将数据信息保存到一个文本文件中。（除了最终的大作业报告，还需提交相应源代码以及 txt 文件）


四. 运用相关库获取上市公司 A 和上证指数的股票价格数据，分析其基本特征，进行正态性检验，并实施贝叶斯回归。具体要求如下：

1) 运用 Tushare 库获取 A 公司和上证指数(000001)从 2018-01-01 至 2019-12-31 的股票价格数据收盘价。

2) 将 A 公司和上证指数的股票价格数据规范化并绘制股票规范化价格数的线图和散点图。

3) 计算 A 公司和上证指数的 对数收益率，并 进行正态性检验 (3 种方法选其中 2 种)。

4) 运用 pymc3 库 对 A 公司和上证指数 的股票价格 实施贝叶斯回归 (A 公司的股票规范化价格为 x 值,上证指数的股票规范化价格为 y 值)，输出统计结果，并绘制后验分

布及轨迹图，和贝叶斯回归线。（除了最终的大作业报告，还需提交相应源代码以及 txt 文件）