量化分析 B 期末大作业要求

数管 2301、数管 2302

一、 作业要求：

1. 爬取天财新闻上的所有新闻；
2. 请根据第二部分内容要求完成相关代码并生成相关图片；
3. 所做图片需要符合审美要求，可参考“阿昆的科研日常”微信公众号中的配色，每张图片至少应该标明（title，x_label，y_label、legend、major_tick 等内容）
4. 请撰写你的实验报告
   a) 正文汉字使用楷体、英文使用 Times New Roman、小四、1.5 倍行距、两端对齐；
   b) 一级标题采用标号为：一、二、三；汉字使用楷体、英文使用 Times New Roman；四号、1.5 倍行距、加粗；
   c) 二级标题采用标号为：1. 2. 3 ；汉字使用楷体、英文使用 Times New Roman；小四号、1.5 倍行距、加粗；
   d) 对于内容要求中的每一条，需要在你实验报告中写明你所统计内容对应的公式，公式需要采用公式编辑器或者正确的上下标，禁止插入图片作为公式，居中对齐，并附图给予必要说明；
   e) 插入的图片要求居中对齐，并且图例采用楷体 5 号字，图例在图的下方
   f) 实验报告中禁止粘贴代码；
5. 提交内容打包要求
   a) 提交的内容包含 5 个部分，分别为 实验报告.docx、实验报告.pdf、code 文件夹中存放你所使用的代码，data 部分存放你下载的数据集，figure 部分存放你所使用的所有图片。

b) 以上 5 部分内容放置在一个“学号-姓名-班级”文件夹内，示例：2485-张三-数管 2301

c) 请将上面文件夹打包成.zip 格式进行提交。

6. **请注意，不按照打包要求进行提交的，一律视为未交。包括但不限于“没有提交 pdf 文件，提交文件名为 2485-张三-数据管理 2301”等未按要求提交的情况.**

7. **提交地址和截止时间**

   量化分析 B 期末大作业

   截止时间：2024-07-05 23:59 (该时间可能提前或者错后，建议尽早完成)

   提交地址：https://send2me.cn/9EKYGTz4/SnazcFJtrg740w

二、 内容要求：

1. 按照年/月进行统计，统计学校每年每月会发布几条新闻？**绘制折线图**（横坐标为“年-月”，纵坐标为发布新闻数量；
2. 对题目、摘要中的汉字进行统计，统计在我校新闻当中，统计每一个字出现的次数，将出现字数最多的前 50 个汉字以及他们出现的次数，**绘制柱状图**。（横坐标为汉字，纵坐标汉字出现的次数），请注意，对于单字而言，标点符号不在排序范围之内；
3. 针对第 2 问当中的内容，对两字词、三字词进行统计。（例： 对于句子“我爱中国”来说，“我爱”、“爱中”、“中国”都被视作两字词，即相邻的两个汉字，三字词与之类似，“我爱中”、“爱中国”）；
4. 按照新闻供稿单位进行分类，统计每一个供稿单位发布新闻的数量。绘制饼状图；
5. 按照新闻供稿单位进行分类，统计每一个供稿单位发布新闻的平均阅读量，绘制柱状图。（柱状图横坐标为供稿单位，纵坐标为供稿单位平均阅读量）；
6. 按照 **星期几** 统计平均阅读量，绘制相应图片。（请问，平均周几发布新闻阅读量会大一点呢？）；
7. 请统计新闻供稿单位所发新闻的平均字数（每篇稿件平均多少字，记为 x），统计新闻供稿单位所发新闻的平均图片数量（每篇稿件平均几张图片，记为 y），请根据（x，y）绘制散点图。

8. 其他任何你认为具备可以统计的内容，比如：你可以计算 TF-IDF 相关的内容进行排序，你可以根据新闻供稿单位之间词的相关性进行热力图的生成，等等等等