### 项目简介

通过爬虫爬取当当网python类图书的相关信息，采集1001条数据，并通过对这些书籍数据进行统计分析，给出购买python类书籍的建议。

### 项目开发环境说明

①Python 3.9.0

②json 2.0.9

③requests 2.24.0

④bs4 4.9.3

⑤re 2.2.1

⑥matplotlib 3.3.3

### 代码说明

①01dangdangbooks\_spider.py

**功能**：该模块用于爬取当当网前1001本书的信息（书名，价格，作者，出版社，出版日期，评论数）。

**核心函数**:

(1)parsePage(html);

通过BeautifulSoup库解析对应的HTML文档，获取书籍对应的书名，价格，作者，出版社，出版日期，评论数。

html: 需要解析的HTML文档。

(2)getHTMLText(url)

根据url获取相应的HTML文档。

url: 需要查询的网站。

②02dangdang\_booksdata\_txtTojson.py

**功能**：该模块用于将数据文件从.txt文件转化为.json文件。

**核心函数：**

dealDataFromTxtToJson(fname,savefilename)

将txt文档转化为json文档。

fname: 需要转化的文档。

savefilename: 转化后保存的文档。

③03dangdang\_datastatistics.py

**功能**: 该模块用于统计分析书籍信息（根据评论数，价格区间，出版社，出版日期）。

**核心函数:**

(1)countCommentNum(filename,hotfilename,totalStatistic\_filename)

统计不同评论数（0，100，1000，10000， >10000）的书籍信息。

filename: 保存所有书籍信息的文件，用于读取。

hotfilename: 用于保存评论数大于1000的所有书籍信息，这些书籍称为hotbooks。

totalStatistic\_filename: 用于保存根据评论划分后的书籍信息。

(2)countPriceRange(filename, savefilename)

统计处于不同价格区间（0，50，100，150，>150）的书籍信息。

filename: 保存所有书籍信息的文件，用于读取。

savefilename: 保存统计后数据的文件。

(3)countPublisherNum(filename, savefilename)

统计书籍中各个出版社出现次数。

filename: 保存所有书籍信息的文件，用于读取。

savefilename: 保存统计后数据的文件。

(4)coutPublishTimeRangeNum(filename, savefilename)

统计处于不同出版日期（<2016,2016,2017,2018,2019,2020）的书籍信息。

filename: 保存所有书籍信息的文件，用于读取。

savefilename: 保存统计后数据的文件。

④04CommentPie.py

**功能**: 根据评论数统计信息绘制相应饼图。

**核心函数**：

（1）plotPieWithCommentNum(datafilename, savefile)

根据评论数的统计信息绘制相应的饼图。

param datafilename: 保存评论数统计信息的文件，用于读取。

param savefile: 保存饼图的文件。

⑤05PriceRangePie.py

**功能**：根据价格区间统计信息分别绘制所有书籍和热门书籍对应的饼图。

**核心函数**：

（1）plotPieWithPrice(datafilename, savefilename, flag=0)

根据价格区间的统计信息绘制相应的饼图。

datafilename: 保存价格统计信息的文件，用于读取。

savefilename: 用于保存饼图的文件。

⑥06PublisherTimes.py

**功能**：根据出版社出现的次数的统计信息分别绘制所有书籍和热门书籍对应的条形图。

**核心函数**:

(1)plotPublisherTimes(datafilename, savefile)

根据所有书籍出版社出现次数的统计信息绘制条形图。

datafilename: 保存出版社出现次数的统计信息的文件，用于读取。

savefile: 保存条形图的文件。

（2）plotPublisherTimes1(datafilename, savefile)

根据热门书籍出版社出现次数的统计信息绘制条形图。

datafilename: 保存出版社出现次数的统计信息的文件，用于读取。

param savefile: 保存条形图的文件。

⑦07Publish\_time.py

**功能**：根据出版日期的统计信息绘制对应的条形图和饼图。

**核心函数**：

（1）plotPublishTime\_Bar(datafilename, savefile)

根据出版日期的统计信息绘制相应的条形图。

datafilename: 保存出版日期统计信息的文件，用于读取。

savefile: 保存绘制的条形图的文件。

（2）plotPublishTime\_Pie(datafilename, savefile)

根据出版日期的统计信息绘制相应的饼图。

datafilename: 保存出版日期统计信息的文件，用于读取。

savefile: 保存绘制的饼图的文件。

### 处理流程说明



### 目录结构

data --- 存放储存数据的文件

program\_module--- 存放爬虫，数据分析，数据可视化程序的文件

Image--- 存放最终统计结果形成数据可视化视图的文件