Java 作业说明

2019 年 4 月 27 日

1 作业综述

本次作业实现了一个网络爬虫，从 http://eecs.pku.edu.cn 开始，下载了所有网站中的图片，同时粗略筛选了文章的正文文本，调用 java 的 jieba 分词包，统计了网站中每个词的出现个数。所有的代码，以及 dict.txt 和 prob_emit.txt，都应该被放在同一个目录、package spider 下。（后两个 txt 主要是 jieba 分词使用）

2 实现概述

本次作业主要编写了 spider 类，调用了 HttpKit 类和 jieba 分词的一系列类。其中，HttpKit 类与示例代码相同。

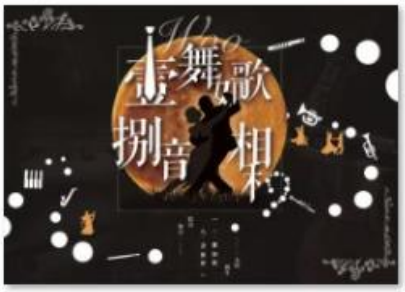
程序运行时不需要输入，直接从给定的 url（即信科网站主页）开始运行。对每个 url，首先用 HttpKit 类获取网站内容。之后，遍历内容，用正则表达式找到所有的 href 和 src。若 href 和 src 的地址以.png,.jpg 或.bmp 结尾，则开启一个新线程，将（atomic int 的）全局计数器+1 并以计数器为文件名，下载该图片到 pics/；若结尾是一个地址或网址，且为 shtml 或该网站下的地址，则递归运行爬虫程序。在全局用一个 Concurrent Map 记录当前已访问的 url。遍历结束之后，再开始一次遍历，找出所有长度不小于 15 的中文字符（包括标点符号，主要是为了粗筛正文内容），调用 jieba 分词，将结果放进另一个 Concurrent Map 计数。最后将 Concurrent Map 的关键词转为 Navigable Set 再转为 array，逐个在 Map 中查询并输出。

3 运行截图

```
学院:1560
北京大学:1508
教育:1412
科学技术:1206
和:1092
与:1070
中国:991
“:976
在:953
”:921
（:788
学会:775
电子:758
计算:664
高等:639
了:620
信息化:591
：:584
国家:574
分会:566
项目:554
所有:547
）:542
高等教育:540
版权:527
版权所有:525
教授:511
系统:504
学生:467
等:453
通信:428
博士:423
年:399
计算机:385
算机:385
；:382
工作:381
领域:361
学术:359
```

14968.png

14969.jpg



14970.png

14971.jpg

14974.png

14975.jpg