

Java应用技术第二次作业报告

作者：葛帅琦

班级：计科1505

学号：3150102193

作业描述：跟踪特定网页，下载该网页中所有链接的指定内容，去除广告等无关内容，组合成单一文件。主要作广度搜索，深度暂为1。

在我的代码中，正文使用了substring的截取方式，目录链接用了正则表达式进行截取。
HTML也可以利用第三方库HTMLparser（进阶）来爬取。写了该模块，但是毕竟只要选一种就行了，我就把这段代码给注释了

1. 函数架构

```
main ()  
    输入网站url  
    爬取目录中章节url链接  
    for循环爬取各个章节正文内容
```

1. 关键函数讲解（

■ main函数入口

```
public static void main(String[] args)  
{  
    String catalogUrl = "http://www.kanunu8.com/files/yuanchuang/201102/1530.html"; // 努努书坊《诛仙》目录  
    websiteUrl  
    novelName = "诛仙.txt"; // 可自行设置小说名字  
    GetCatalog(catalogUrl); // 建立目录  
    String url;  
    for(int i=0;i<rootCatalog.length;i++)  
    {  
        System.out.println("第"+(i+1)+"章"+",共"+count+"章..."); // 获取第i+1章url链接  
        url = rootCatalog[i]; // 爬取该章节  
        ReadContentAndWrite(i+1,url); // 读取文章内容  
        并存储  
    }  
    return;  
}
```

- 目录链接获取并存储（利用正则表达式）

函数输入参数为对应章节url链接。
根据网站源文件格式正则爬取

```
Pattern pattern = Pattern.compile("<td width=\"25%\"><a href=\"(.+?)\">.+?</a></td>");
```

整体关键部分如下：

```
static void GetCatalog(String catalogUrl){
    .....
    // 正则处理章节链接
    Pattern pattern = Pattern.compile("<td width=\"25%\"><a href=\"(.+?)\">.+?</a></td>");
    // 定义一个matcher用来做匹配
    Matcher matcher = pattern.matcher(result);
    // 遍历整个网站
    while (matcher.find()) {
        // 打印出结果
        //System.out.println(i+ " " + matcher.group(1));
        StringBuilder t = new StringBuilder(root + matcher.group(1) + "!");
        Catalog.append(t);
        count++;
    }
    .....
}
```

- 爬取正文并存储
爬取正文我用了最暴力的字符串处理：

```
int start = result.indexOf("<p>");
int end = result.indexOf("</p>");
String str = result.substring(start+3,end);
str = str.replace("<br />","");
//获得正文内容
```

- 利用第三方库HTMLparser来进行爬取
HTMLparser的使用网上文档说明不多，使用起来并不方便。它是按照HTML的元素进行分割节点，然后会有个过滤的选项。

```

try{
    //建立parser对象
    Parser parser = new Parser( (HttpURLConnection) (new URL("http://www.kanunu8.com/files/yuanchuan
g/201102/1530.html")).openConnection() );
    parser.setEncoding("GB2312"); // 解码方式
    NodeIterator Nodes = parser.elements ();
    String HTML = ""; // 遍历节点
    for(Node node = Nodes.nextNode();Nodes.hasMoreNodes();node = Nodes.nextNode()){
        //System.out.println("toHtml:"+node.toHt
ml());
        HTML = node.toHtml().toString();
    }
    System.out.println(HTML);
}
catch( Exception e ) {
    System.out.println( "Exception:" +e );
}

```

2.运行结果



```

118
119
120
121
}
    // 打印出结果
    return matcher.group(0);
}

```

The screenshot shows a Java IDE interface with a 'Run' tab selected. The code area contains the provided Java code. The output window displays the following text:

```

118
119
120
121
}
    // 打印出结果
    return matcher.group(0);
}

```

小说截图

```
C JavaCraw.java × 啦啦.txt × 诛仙.txt ×
34542
34543 萧逸才应了一声，迅速转过身来，右手一挥，自己当先飞起，跟在他身后的是将近百人的正道中人，人数虽然没有云海广场上的多，但法宝毫光之闪亮耀
34544 轰然雷鸣，电芒在天空苍穹乱窜，仿复又回到了多年之前的那一场雨。只是不知怎么，就算是这个雨天，天际上竟然还有着那么一轮诡异的月亮，很亮很白。
34545 雨水打在脸上的感觉，那么的凉.....。
34546
34547
34548
34549 张小叉木然回首，风雨潇潇，那一个小小村落，终究稍稍隐去。他不由自主的伸出手去，想要抓住些什么，但空空如也。只有身后，普智那一双眼睛，静静
34550 下一刻，他已经置身在那个熟悉的房间，大竹峰上特有的气息，在四周泛地，那么的亲切与熟悉。远处有诸位师兄们的谈笑声，有大黄和小灰的嬉闹声，还
34551 有那山风的呼啸声。
34552 他全身飞抖突然之间，数十年来在心间筑起的心防堤坝破碎了，崩溃了。
34553
34554 他泪流满面！
34555 枯槁的手掌从背后伸出，轻轻拍打他的肩，那个和蔼的声音低声问道：“怎么了，孩子，为什么要哭呢？”
34556 张小凡忽然回头，看着那个慈悲的脸庞，身子忍不住的绷紧。他深深的盯着面前那双眼睛，直欲看到这个慈悲老和尚的深心处，只是普智的眼神从来是那么平
34557 和。
34558 他一字一字地、仿佛是低吼一般地问道：“为、什、么，为、什、么你要选我，为什么你要这么做？”
34559 普智没有回答，他只是依旧那么悲天悯望着张小凡，眼神中除了慈悲还是平和，看不到任何的情绪波动，更不用说是什么后悔！
34560 身旁的一切又再一次消失了，整个世界又只剩下了他们两个人。张小凡，不，现在看去他整个人已经仿佛化身恶魔，凶厉的血红目光再一次占据了他的神，人
34561 普智的目光终于震动了一下，慢慢向那件凶煞之物望去。噬魂顶端，那颗正大放光芒的“噬血珠”，一点一丝遍布珠身的暗红血丝，仿佛也都在凝望着他，带
34562 沛不可挡的血腥气息，突然从从鬼厉身上凭空出现，继而排山倒海般冲了过来，如狂风吹过，普智僧袍猎猎飘舞，怔怔望着，那狰狞中带着绝望的红芒，如
34563 他没有丝毫回避的意思，站在那里，一动不动，下一刻，那绝望而凶狠的红芒穿过了他的身子，慢慢在他身后停下，凝聚出鬼厉的身影。
34564 苍老的和尚缓缓低头，慢慢看了一眼自己的身体，然后，他叹息一声，头颅垂下，身子缓缓跌倒一旁。在他身后，鬼厉猛的转过身子，看着普智，脸上神色
34565 “啊！.....”
34566
34567
34568
34569
34570
34571
34572
34573
34574
34575
```

实验心得：

操作起来，java自带的URLConnection用起来要比第三方库HTMLparser轻松很多。对于简单的文本，其实直接字符串处理就够了。如果爬取要求条件较多，可以考虑使用正则表达式。当网页结构变化较多时，正则表达式也有一定的问题。爬取这本小说数百万字，总共4.7M耗时大概十分钟左右，不是很满意。有机会再研究下，如何提升java的性能。