遇到的错误:

问题1.输入文件的位置作为字符串的时候,遇到Invalid escape sequence (valid ones are  \b  \t  \n  \f  \r  \"  \'  \\ ) 报错

一个正则表达式错误

原因：Java源代码的字符串中的反斜线被解释为Unicode 转义或其他字符转义

解决办法： 在“\”后加（\b   \t   \n   \f   \r   \） 其中的一个就可以了。当然也可以直接把语句中的“\”去掉。

问题2. 东西不对,怎么只有这么一点内容呢?

解决方法:仔细查看具体内容,最后发现,应该是被覆盖掉了, 写入的是(尾声). 应该在后面接着写入而不是覆盖.

追加,是在filewriter中用true参数

我先试试把 pw pw2 放在前面,不要每次都new一个.

但是出现了错误

问题3.Unhandled exception type IOException

这是ioe异常,在read.line方法前加try,catch语句。或者在方法上抛出异常类

解决了之后

还是只有尾声

问题4. 尝试调整了 pw = **new** PrintWriter

reader = **new** BufferedReader

pw.close();

pw2.close();的位置

终于可以全部爬取下来html格式.

接下来还需要选择出文本.

Pw2写入简化后的文本.

问题5. 前面有大量空白,中间有大量重复

解决方法: pw2.println(str);//写入简化后的

把这句放在**while**( ( line = br.readLine() ) != **null** ){

循环外,

问题6. 一行里面有多段html包裹的文本

**while**(i >= 0) {

j = line.indexOf("</span>",i);

str += line.substring(i+"<span class=\"calibre8\">".length() , j);

i = line.indexOf("<span class=\"calibre",j);

}

利用循环来提取.

至此, 基础模块基本完成.

提高模块

使用HTMLParser

要使用该库，在编译和运行时，您需要将htmllexer.jar或htmlparser.jar添加到您的类路径中。htmllexer.jar以线性，平坦，顺序的方式提供对页面上的通用字符串，备注和标记节点的低级访问。htmlparser.jar（包含在htmllexer.jar中找到的类）提供对页面的访问

<https://my.oschina.net/u/175660/blog/85205>

对于大多数使用者来说，使用最多的是通过一个*URLConnection*或者一个保存有网页内容的字符串来初始化*Parser*，或者使用静态函数来生成一个*Parser*对象。

问题1

怎么打开一个txt然后解析?

问题2

报错

忽然br.close();出错了,不知道为何.

java.lang.NullPointerException

解决方法:把close()放在try语句里面,

问题3

然后报错另一个java.io.FileNotFoundException:

应该是因为没有新建目录

解决了之后, java.lang.NullPointerException还是存在

应该是因为每个try语句里不通?

BufferedReader br = **null**;放在try里面

问题4

Unreachable catch block for IOException. This exception is never thrown from the try statement body不可到达的 catch 块异常。try 中的语句，永远不会引发此异常

问题5

NoClassDefFoundError发生在JVM在动态运行时，根据你提供的类名，在classpath中找到对应的类进行加载，但当它找不到这个类时，就发生了

解决方法:

1.存放第三方jar包的目录名改为libs之后，重新设置build path.这样还是出现错误.

2.我把所有的jar都包含进来,好像是htmllexer的问题,我去掉这个包就出现了这个错误.

问题6

显示错误

Error occurred during initialization of boot layer

java.lang.module.FindException: Unable to derive module descriptor for C:\Users\12638\Downloads\htmlparserv16\htmlparser1\_6\libs\sax2.jar

Caused by: java.lang.module.InvalidModuleDescriptorException: SAXDump.class found in top-level directory (unnamed package not allowed in module)

解决方法:

JDK版本：10.0.2

方法1： 新建一个包，将默认包中的类文件拖入新建包中，刷新，默认包会消失

方法2： 将module-info.java删除

但是我没有看到module-info,可能因为我一开始就没有用模块,

我就把sax2.jar删除了

问题7:

出现错误:

Exception in thread "main" java.lang.Error: Unresolved compilation problem

把其他的都注释掉,应该是这里报错:

The package org.htmlparser is accessible from more than one module: filterbuilder, htmllexer, htmlparser, Thumbelina

原因:

在添加Jar包时选择的是Modulepath进行添加。   
更改为Classpath后即可解决。

问题8:

字体忽然变了

解决方法:这个不是你的Eclipse的问题 是你的输入法被你切换为全角模式输入后的情况 只需要把输入法切回半角模式即可.

问题9:

Html解析出来的也没有追加,也只有尾声.

解决方法:

PrintWriter构造函数的第二个参数是表示是否自动刷新缓冲区，FileWriter第二个才是才是追加写文件．加上true就可以了.

问题10 :怎么打包有外部jar 的java.

<https://www.cnblogs.com/lanxuezaipiao/p/3291641.html>

还是不太会,

cmd进入项目目录输入以下命令

//执行以下命令后会在当前目录生成一个warname.war

jar -cvf warname.war ./\*

//有一个弊端就是把整个项目所有的文件都弄到了war里面了，体积巨大。

//解压war包

jar xvf warname.war

我用这个方法了,但是还是不知道怎么打开project,project打开也打开不了.

我的发现:

**new** BufferedReader(**new** InputStreamReader(connection.getInputStream())

可以写在一起,就不用给**new** InputStreamReader 命名然后再调用了.

把reader的URL的名字改一下,不要固定名字,

"C:\\Users\\12638\\Desktop\\javalearning\\homework\\9\_27homework2\\txtdir/"+time+"/"+filename+"\_URL.txt"

然后

new FileWriter("HTMLparaser/"+time+"/"+filename+"\_ByParaser.txt",true),true);

这样就可以在当前目录下创建文件夹然后输出,就可以做到不同电脑通用

Strbuffer可以append(“\n”)就可以换行了.视觉效果很好.