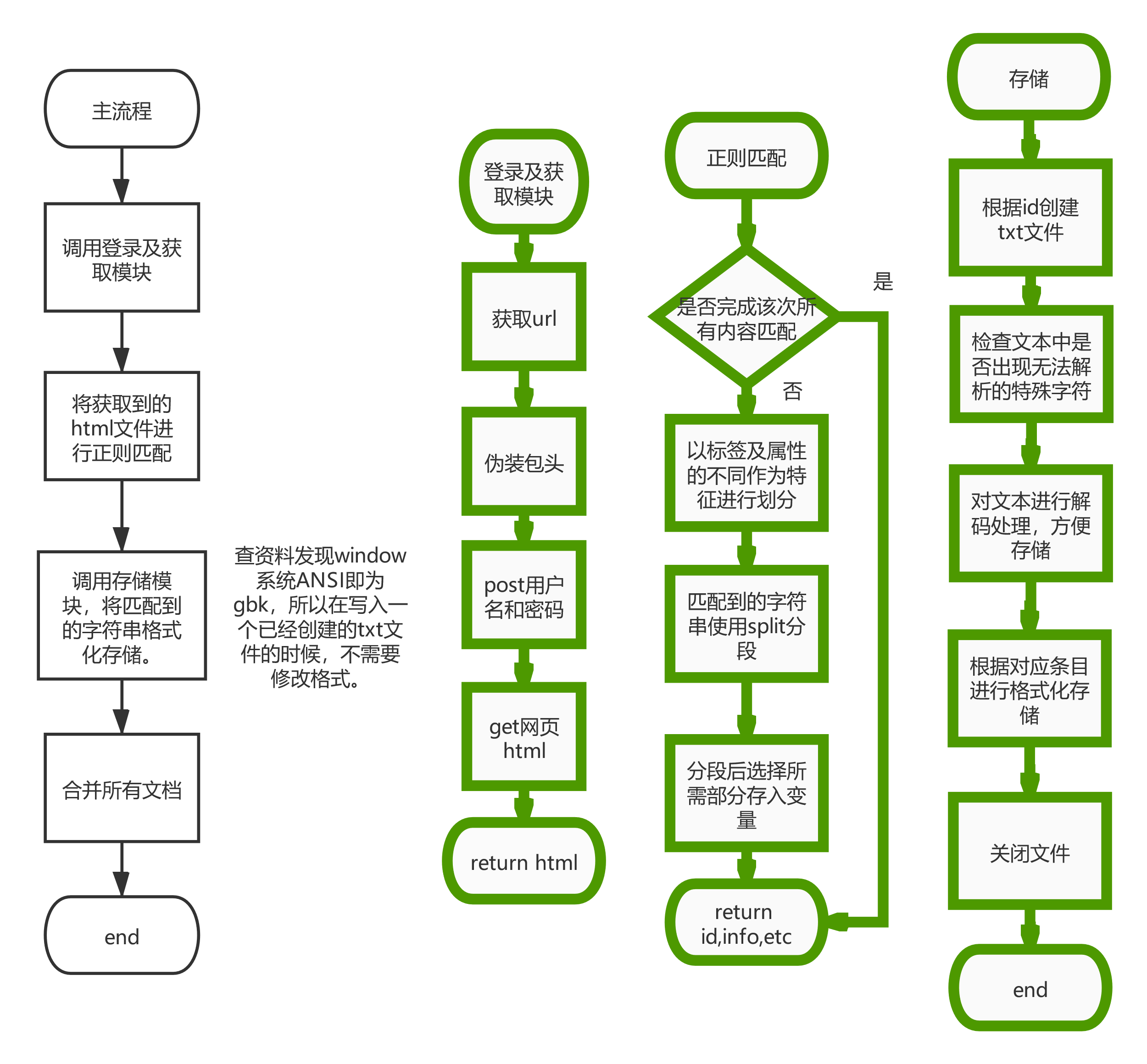
|  |
| --- |
| 北京邮电大学 |
| Web搜索技术 |
| 信息的精确获取 |

|  |
| --- |
| 田宇佳  2017211124班  2017210421 |

1.整体思路和流程图



2.反思与总结：

1.网络问题:

一开始我试图使用webVPN的方式解决登陆问题。通过python登录vpn再发送请求在学校的vpn链接后面增加大创网站的网址进行访问的方式。这种方式在浏览器上是可以成功访问网站的，但是在爬虫访问的时候似乎就出现了一些我不可以解决的小问题，尽管登录返回了200ok，但再次请求的时候报了很多ssl的错。所以经过几个小时的挣扎之后，我最后还是使用了global protect进行全局VPN的配置。

2.正则匹配的设置问题。

一开始我全部使用的是，I和M模式，后来发现这样的话无法匹配info部分的长文本，因为那个部分的文本前后与div标签之间有很长的换行字符，而且该div标签在工具中匹配时可以成功的在py中不可以，在这一部分更改了多次。所以最后借鉴了一些别人的写法，并将模式更改使用到了s模式，这样就可以匹配多行字符。

3.编码格式问题

在这个实验中我也发现，很多司空见惯的字符，其实在一些编码格式里是可行的，一些编码格式里则根本不存在，比如说，位于行中间的点这样的字符。

除此之外不同格式的字符串匹配。我发现对于一些情况，是可以进行直接转换的，另外一些情况需要先encode（）再decode（B）。虽然这个问题可能主要原因是我自己不知道哪些部分获取到的是有编码字符串，哪些部分是我组装出来的没有编码格式的字符串。但是确实可以算是一个玄学问题。