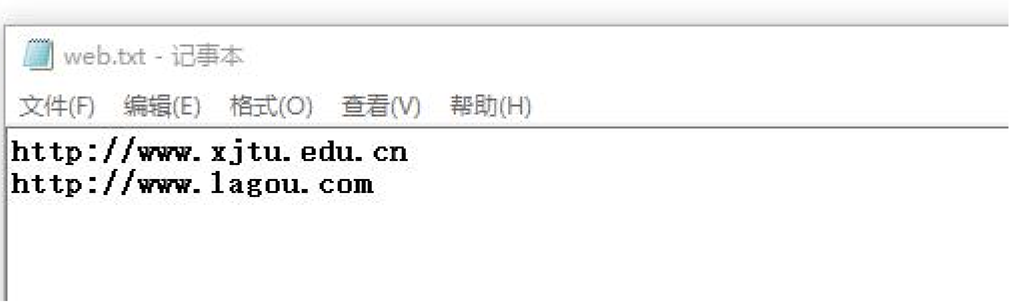
使用说明

要想使用 newspider，只需要了解 http 文件夹中四个文件的作用。

①web. txt：将所要爬取的网站写入此文件中，网站域名以“http://”开头。如果想爬取更多的网站，只需要依次将域名写入文件中，如下图所示。删除本文件，爬取将不会开始。



```
http://www.xjtu.edu.cn
http://www.lagou.com
```

②queue. txt:此文件为队列文件，整个爬取过程就是队列的先进先出过程，删除本文件会使得爬取过程从头开始。

③point. txt:此文件为断点文件，断点续传功能要依靠此文件的参与。删除此文件的同时，也要删除 queue. txt 文件。

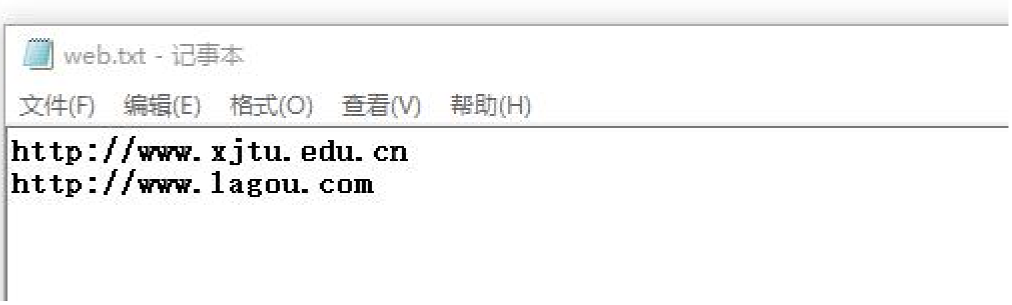
④url. txt:queue. txt 的参照文件，无实际作用，可随意删除。

⑤你可以同时删除 queue. txt、point. txt、url. txt 这三个文件，运行程序时这三个文件又会重新生成。

Instructions

To use newspider, you only need to understand the role of the four files in the HTTP folder.

①web.txt：Write the site name in this file,the website domain name begins with “http://”.If you want to crawl more sites,only need to write the domain name in the file,as shown in the following figure.Deleting this file, crawling will not start.



```
http://www.xjtu.edu.cn
http://www.lagou.com
```

② queue.txt:This file is a queue file,the crawling process is the First-In-First-Out process of the queue,deleting this file will make the crawling process over again.

③point.txt:This file is a breakpoint file, the resume function depends on this file.If you delete this file, you also should delete the queue.txt.

④url.txt:queue.txt reference file,it has no actual effect, so can be arbitrarily deleted.

⑤You can delete the queue.txt, point.txt, url.txt at the same time,and these will be generated when running the program.