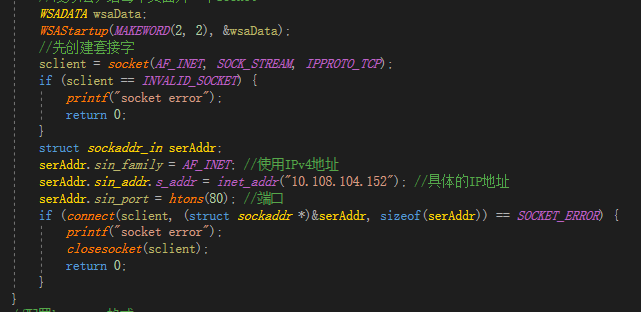
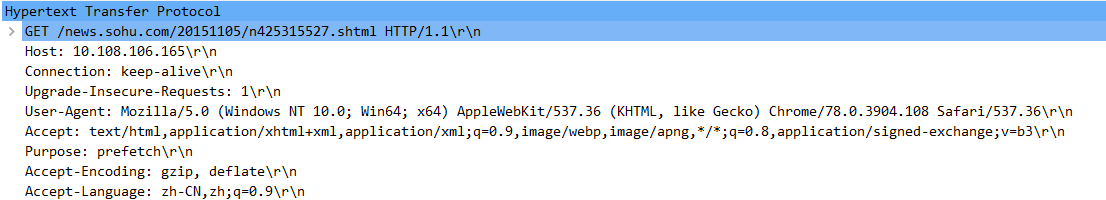
实验四——windows下实验指导书

这个实验用windows开发的都是壮士。不说了,我登我的centos去了。

1. 使用socket套接字建立TCP连接（建立长连接将极大的加速程序运行，图中没有建立，setsockopt去配置就行）



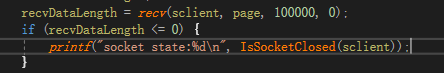
1. 然后拼接形成http协议包头，通过send api向服务器进行发送(这里设置keepalive)



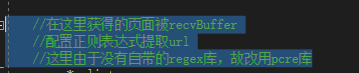


1. 通过recv api获取返回的网页，这里如果返回值<=0 很大可能性是链接超时了

(这里有点奇怪的，我自己配置长连接的时候日常超时断连，我还自己写了心跳包还是不行。Ps:win下心跳包要自己写。后来退而求其次用来多线程重复connect)

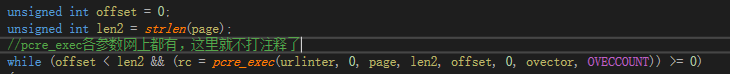


1. 随后对返回的页面进行解析



Win下C99没有正则表达式库，因此用了开源正则库pcre

Pcre需要自己编译安装，当然也可以用vcpkg一劳永逸，安装过程按下不表

注意pcre\_exec只匹配出现的第一个位置，需要创建迭代器进行循环

1. 最后将爬取的url送入队列并循环至队列空。
2. 去重操作建议直接37叉trie树(26+10+1叉)，毕竟布隆过滤器对网页关联取id无能为力
3. 多线程爱用不用，反正单线程2h也就跑完了。

Win下C99实现线程池超麻烦的，实在做不出来报告里写一下就行，不准动刀啊...

8）要是真的真的真的真的有困难，晚点我会给一个完整的网页关联列表，先做pagerank算法，然后在提交的作业文件夹里说明。明确说，会扣分，但不至于挂。要是想坑一下聪明可爱帅气(误)的我，你猜我能怎么办。