新人练习题2(多线程抓取)

试编写多线程程序，要求同时对多个url抓取网页内容，并可以指定最多线程数。

待抓取的url是写在硬盘上的输入文件中的，每行一个，总数不一定，抓回的网页内容要写在硬盘上的指定目录下，每个网页一个文件、以其url作为文件名。

url位置：（ 下载地址：ftp://szjjh-ps-dyndi99.szjjh01.baidu.com:/home/work/hadoop-mnt-w/user/liantiao/AUTO-liantiao/rank760/20130423/url.list ）

要求:

1. 时间3天
2. 用c或c++完成, 抓取部分不可使用现成的http库
3. 先提交详细设计
4. 利用单测来保证code的质量
5. 可先完成一个简单的版本，时间充足的情况下，现实一到两个优化改进版本
6. 提较code到自己的svn目录，打上tag，提交codeview

bonus:

1. 处理302
2. 处理重复url
3. 实现一到二个，增加抓取速度的优化方案

reference:

http://pswiki.baidu.com/twiki/bin/view/Main/Psex_net1
http://en.wikipedia.org/wiki/List_of_HTTP_status_codes
http://web-sniffer.net/ （ 了解http协议 ）