# 爬网站脚本偶尔出现乱码的说明

1、关于工行主页 www.icbc.com.cn 无法爬到的原因是因为，这个页面只是一个分发页面，代码见 **工商银行反爬虫页面代码.html**

真实的页面地址其实是: [**http://www.icbc.com.cn/icbc**](http://www.icbc.com.cn/icbc)代码见，

**工商银行真实页面地址**[**www.icbc.com.cn斜杠icbc.html**](http://www.icbc.com.cn斜杠icbc.html)

工行这种银行的网站太注重安全了，所以故意不让爬取。

这点你们在进工行主页 <http://www.icbc.com.cn/> 看下跳转地址就知道了。

2、在input.txt中，网址之间不要出现空行。否则会出现乱码，现在加了这个空行判断了。

3、[www.google.com.hk](http://www.google.com.hk)的反爬虫。。。谷歌本身就是搞爬虫的。他的页面中就是一堆JS脚本。访问一旦频繁，就发现是爬虫。就拒绝访问了。所以谷歌首页没有数据很正常， 详见：

**谷歌首页反爬虫页面.html**

4、脚本我稍微改写了一下了，对乱码加了多重判断处理，详见测试数据：

**重新爬取新测试数据产生的结果.xls**

**2012年8月24日**