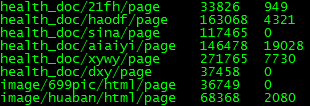
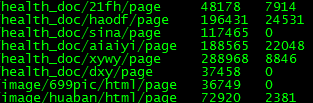
**工作内容：**根据上周检查出的爬虫慢的原因，找到合适的解决方案来提高网页抓取速度。

**产出：**

1.调研url查重方式，有多种方式：缓存方式（将url保存成文件或者数据库）、哈希表方式、 Bloom Filter方式。由于我们的历史url数量非常多，缓存方式查询会非常慢，单一的哈希表发生冲突的概率太高了，而Bloom Filter方式非常适合大数据查询，缺点是有非常小的概率会误判，但是对于我们的问题，对于网址的判断允许小概率的误判（有相关文献指出其错误率约为0.0000889，也就是说对于10万个url，可能存在9个url误判，错误率非常低），所以这种方式非常适合解决我们的问题。

2.将Bloom Filter应用到21fh、haodf、aiaiyi爬虫代码中，以下图是效果的对比。可以看到21fh、haodf网页爬去速度有了非常大的提升，说明该方法还是十分可靠的。

**之前 加入Bloom Filter后**