周报-李诚

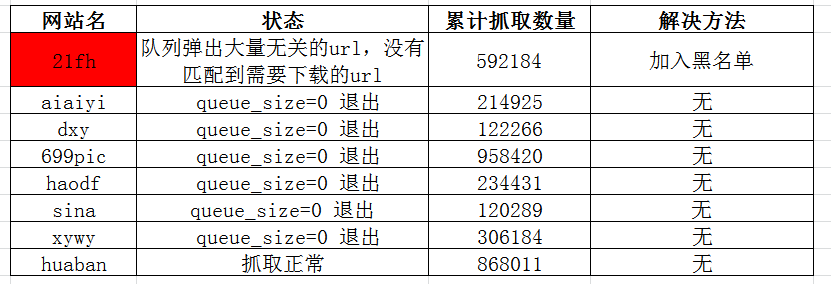
**工作内容：**

1. 爬虫的维护
2. 完成基于LTP的主谓常见搭配程序
3. 解析出网站中的疾病与部位

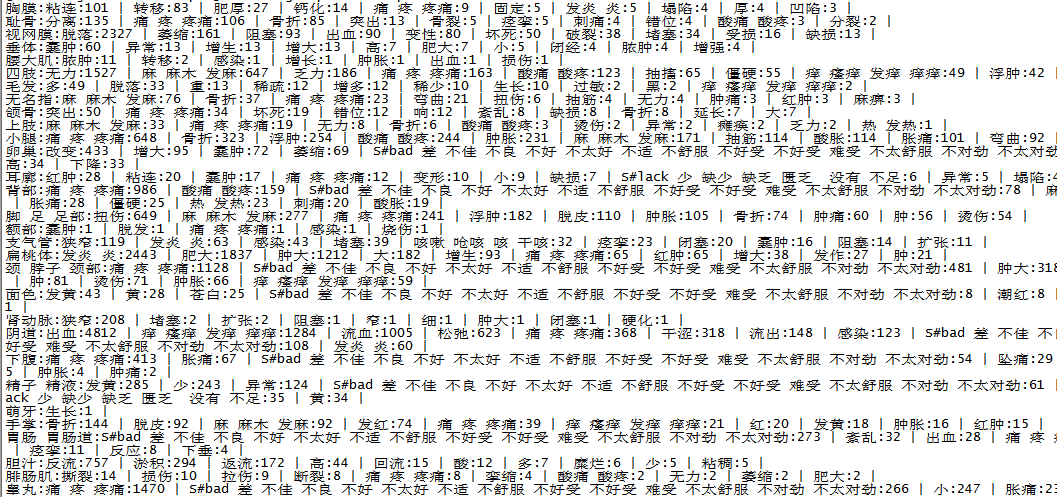
**产出：**

1. 爬虫的维护：

以下为本周一时统计的状态，对于21fh，将无用的url加入黑名单。目前剩下的两个网站21fh、huaban抓取效果较好，可以保持每天至少保存10000个网址。

2. 完成基于LTP的主谓常见搭配程序

用SBV关系对主谓词进行进一步筛选，拿organ\_nlp和problem\_nlp测试的结果较好，如下图，测试结果在209的服务器上，目录：/data/user/licheng/project/app/hospital\_guide\_robot/backends/nlp\_licheng/based\_ltp/res\_sub\_pre。但是程序运行非常慢，已经将程序改为后台运行，等周一来的时候看看效果。



3. 解析出网站中的疾病与部位

解析出三个网站(100ask,21fh,xywy)的疾病与部位，将重复的去除了，但是由于每个网址对某些疾病的名称取名的差异、部位分类的差异，仍有大量的疾病是重复的没有被去除。最终结果已经上传至github，相对路径为：project/app/kg/extract/disease\_location