* 输入输出
* 输入：astat数据：

格式：

access数据分隔符为空格，示例：

60.190.148.234 js.3tip.com - [10/Mar/2015:23:57:01 +0800] "GET /re/re.php?src=t6401&t=&ci=1019122922&r= HTTP/1.1" 200 0 "http://my.tv.sohu.com/us/232423049/78815806.shtml?lcode=AAAARpiVK1t7\_PU3RD8OneDxnhRwUEL9FGNRevctYNKyRZEfjfSInIqsUW7VydQm3iHXcVhlZ2DMoVDIyn8Aaqy\_PFVOstv6KKyko6..d4y&lqd=14042" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)" "uaid=57081b7e5c92dd938fc2ca91004c5845030"

Table 2 字段格式

|  |  |  |
| --- | --- | --- |
| **序号** | **字段名称** | **名称解释** |
| 1 | proxy\_ip | 地区代号 |
| 2 | domain | 域名 |
| 3 | time | 时间 |
| 4 | GET | GET信息 |
| 5 | user\_agent | user\_agent |
| 9 | cookie | 用户标识符 |

* 输出：推广数据目前提取两种数据

1. 私服数据，类似这种数据：

http://www.45chuanqi45.com/ http://www.sogou.com/web?query=%B4%AB%C6%E6%CB%BD%B7%FE%B7%A2%B2%BC%CD%F8&insite=www.45chuanqi45.com&freecode=1 62 61

http://www.52wzcq.com/ http://www.sogou.com/web?query=%B4%AB%C6%E6%CB%BD%B7%FE&insite=www.52wzcq.com&freecode=1 51 46

http://www.1213sf.com/ http://www.sogou.com/web?query=%B4%AB%C6%E6%CB%BD%B7%FE&insite=www.1213sf.com&freecode=1 30 28

每一行的格式：

url refer ips uv

url: 私服的url

refer:该url的refer

ips: 该url访问的不同ip数

uv:该url 访问的不同用户数

每一天形成一个文件，每一行的格式如上所示，文件名格式是：sifu\_<date>.txt

例如：sifu\_20150406.txt

1. 游戏弹窗数据

格式如下所示：

http://d3d.3dwwwgame.com/p/redirect.php?u=17452&p=MTc0NTIsNDEsMCwzN2NzNTY0NzgwMjU3OTMsMCwyNDM0Myw2ZGU3MzVmZjI2ZjJjZDdmMjU2YmU3OWIxMmYwZWRiZA%3D%3D

http://d3d.3dwwwgame.com/p/redirect.php?u=3788&p=Mzc4OCwxMjEsMCwzN2NzNTQ1NDk5ODU4MDEsMCw1MDQ4LDZkZTczNWZmMjZmMmNkN2YyNTZiZTc5YjEyZjBlZGJk

http://g.29m4.com/1/1695.html?uid=11650

http://g.29m4.com/1/1695.html?uid=5102

<http://g.29m4.com/1/1768.html?uid=21791>

每一行都是一个游戏弹窗的url

每一天形成一个文件，每一行的格式如上所示，文件名格式是：game\_<date>.txt

例如：game\_20150412.txt

* 数据提取过程
* 私服数据：在192.168.10.138 的/home/clickwise/lq/sifu下执行，以clickwise用户身份
* 1）执行sifu\_host.sh脚本：

执行示例：

sh sifu\_host.sh 20150427 > log.txt 2>&1 &

输入是hive中的astat表，输出在host\_20150427 文件夹下，输出格式为：

Host\001uv

可以使用下面的脚本对上面格式进行转换，以方便处理

cat 000000\_0 | awk -F'\001' '{if($2>5){print "http://"$1}}' >> sifu\_host\_20150427.txt

用gedit打开sifu\_host\_20150427.txt：

然后逐个点击url，保留是私服链接的url,不是的去掉：



选出的url仍在sifu\_host\_20150427.txt中，格式为：

http://www.45chuanqi45.com

http://www.1213sf.com

http://www.sf999.com

http://www.17123.com

http://www.sf138.com

http://119uc.com

http://33ok.com

http://www.zhao5f.com

http://www.sf78.com

http://sf.9000pk.com

<http://www.10pk.com>

可以将该文件改名为sifu.txt

* 2）执行sifu\_look.sh脚本

执行示例：

sh sifu\_look.sh 20150427 > log.txt 2>&1 &

输入是hive中的astat表，输出在host\_20150427 文件夹下，输出格式为：

url\001refer\001ips\001uv

url: 私服的url

refer:该url的refer

ips: 该url访问的不同ip数

uv:该url 访问的不同用户数

上面的refer为空的话是NA，所以还要转换一下格式将NA变成空，可以通过下面的脚本实现：

cat 000000\_0 | sed -e 's/\x1NA\x1/\x1\x1/g' | sed -e 's/\x1/\t/g ' >> sifu\_url\_20150427.txt

* 3) 合并前两步的结果

通过cn.clickwise.liqi.str.app. SIFUAddField完成合并

通过该类的方法

void addFiled(String sel\_file,String field\_file,String output\_file)实现

sel\_file:步骤1)生成的结果，即每一行一个sifu\_url

field\_file:步骤2)生成的结果，每一行的格式为 url\trefer\tips\tuv

output\_file:最终输出结果，每一行的格式也是url\trefer\tips\tuv，文件名sifu\_<date>.txt

例如：sifu\_20150427.txt

* 游戏弹窗数据: 在192.168.10.138 的/home/clickwise/lq/utr下执行，以clickwise用户身份
* 1）执行utr\_info.sh脚本：

执行示例：

sh utr\_info.sh 20150427 > log.txt 2>&1 &

输入也是astat数据，输出格式如下：

url\001title\001refer\001uv\001ips

url:网页url

title：网页的title

refer:该url的refer

uv:访问网页不同用户数

ips:访问网页的不同ip数

通过下面的命令选择出一些小说的url:

cat 000000\_0 | awk -F'\001' '{if($2~/小说/){print $1}}' >> xiaoshuo\_url.txt



在小说站点上手工模型一些点击，有时会弹窗游戏的窗口：

将弹窗游戏窗口的url记录下来：

例如：

<http://bf.qq.com/cp/a20150416ldy/index.htm?ADTAG=media.buy.dianjiunion.cpm_tanchuang=51913&track=699|2960>

逐个浏览小说站点，进行上述操作，最后捕获的游戏弹窗链接形成文件

game\_20150427.txt

* 数据导入界面：

登陆到110.96.34.211机器，进入到/datadisk/spread\_data文件夹：

将上面提取的两个文件sifu\_20150427.txt和game\_20150427.txt上传到该目录下。

结果如下：



然后执行 sh import\_all.sh 将两个文件导入界面或直接发给李老师，每周一完成该工作。

界面链接：

110.96.34.210:9000/clickeye/login.php

用户名:admin

密码:datamining2013