新人练习题1(脚本)

脚本的精髓在于快速开发, 用尽量简单的代码，现成的工具, 命令来完成这一部分工作。

参考资料：linux shell GUIDEhttp://wiki.baidu.com/download/attachments/53402266/shell_program_guide.pdf?version=1&modificationDate=1420696602000&api=v2

要求:
1天之内完成。可使用bash, sed, awk, perl, python等一切你觉得好用的脚本语言（建议使用shell）。
不准使用c,c++,java等高级语言。

题目:
a. 站点统计
定义：主域名，如下：
news.sina.com.cn 主域为sina.com.cn
www.baidu.com 主域为baidu.com
apple.club.sohu.com 主域为sohu.com

输入：url.list，格式是每行一条url，下载地址：cq01-ps-dev193.cq01.baidu.com:/home/users/pengweihua/fresh_training/exercise_2/url.list
要求计算：
（1）所有这些站点所在的主域列表；
（2）主域上的站点数和url数量；
（3）站点上url数量超过所在主域上站点平均url数量的站点集合

b. 日志排序
请用脚本语言SHELL完成日志排序：有用户日志文件，每行记录了一个用户查询串，长度为1-255字节，共1千万行，请排出查询最多的前100条。
日志文件下载地址：cq01-ps-dev193.cq01.baidu.com:/home/users/pengweihua/fresh_training/exercise_2/query_1000w.txt

bonus: 前100的query占总查询次数的比例是多少？把query查询次数的分布画出来，有什么规律吗?

c. 抓取搜索结果（建议使用python）
给定一批query, 得到这批query在baidu对应的前十结果
query从这里获取, 随机取1000query即可
query文件下载地址： cq01-ps-dev193.cq01.baidu.com:/home/users/pengweihua/fresh_training/exercise_2/query_1k.txt