上海市网络测量实验报告

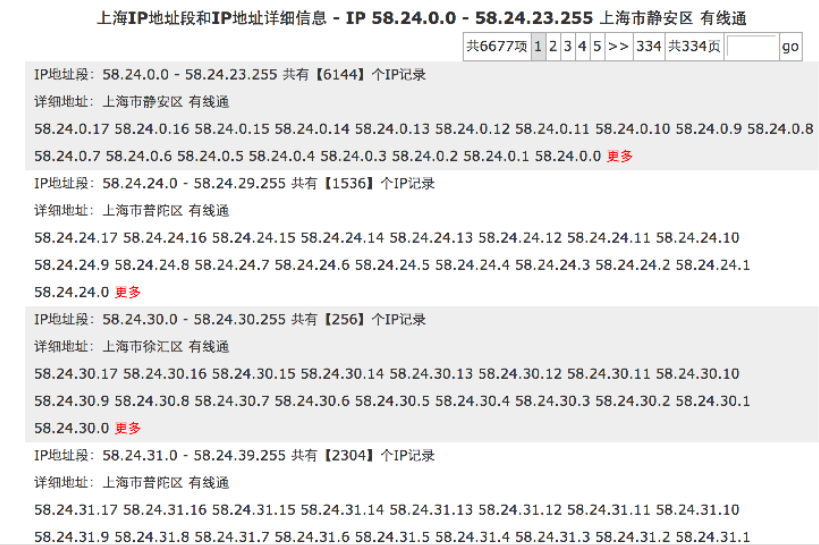
计35 钱迪晨；朱新瑞；吴之璟

1. **实验介绍**

本实验使用主动测量的方式，通过向网络中发送数据，观察结果，来研究上海市各区的IP分布、不同时段的使用率及延迟等。

1. **IP收集**
   1. **获取数据源**

通过网络获取上海市IP数据，经多方面对比，选取了数据最为详细全面的数据源如下图所示：

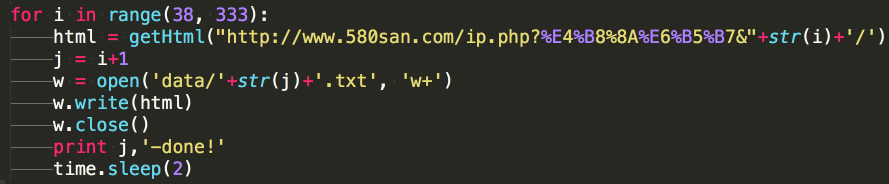


数据包含上海市每个子网的IP范围、其所处的详细区块地址以及网络类型。

* 1. **爬取数据**

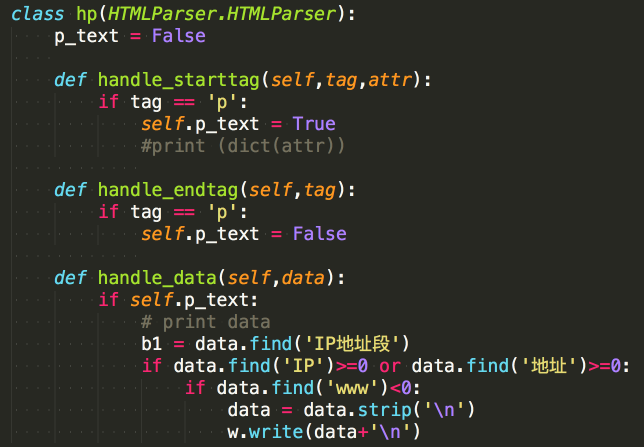
由于IP信息存储在共334个网页上，因此使用python来读取这些网页并存储，便于后续使用。考虑到网络及网站安全限制等问题，每爬取一个网页程序就休眠2秒钟：





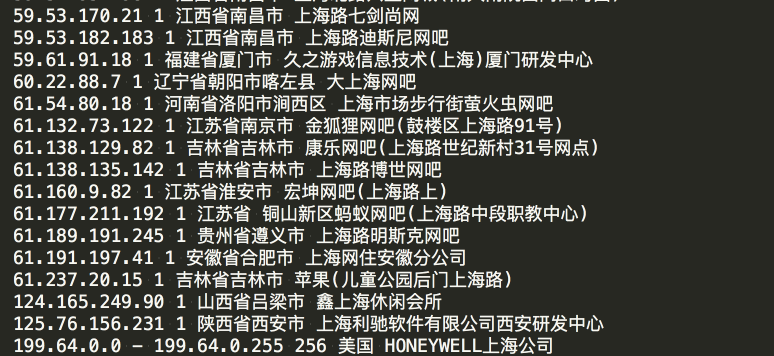
* 1. **提取IP信息**

存储下来的网页信息比较杂乱，因此使用HTMLParser来解析网页，从中提取出每个子网的IP区段、详细地址及网络类型等信息存储在文件中：



* 1. **过滤IP**

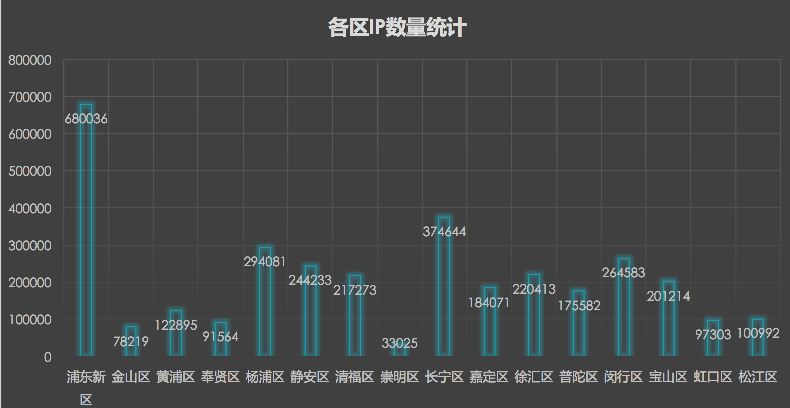
观察提取出的IP发现，网站上的IP信息有些问题，如下图所示，很多并不在上海的IP区段也出现在其中：



观察发现由于在地址中出现了“上海”两个字，比如“江西省南昌市 上海路迪士尼网吧”，虽然不在上海，但地址中出现了上海，因此也被统计入了上海IP中。所以在这一实验阶段，又对提取出的IP进行了过滤。

* 1. **进行简单的统计**

在过滤完保证每个IP的有效性之后对IP地址按照详细地址信息进行区的划分，上海市每个区内的IP地址数量见下表：



由图可以看出，像浦东新区这种商业区IP数量会比较多，而闵行区这样的高校集中区域也有较多的IP数量。相反的，像崇明县这样稍显落后的区域IP数量就会比较少。

1. **实验流程**
   1. **总流程概述**

对于每个区得到的IP进行整理，整理出统一格式

选取一个合适的时间段进行网络测量

整理好实验测量数据，进行结果评估

* 1. **时间与测试机选取**

由于本小组进行实验时，没有较好的实验室资源支持，便选用组员朱新瑞的苹果电脑进行网络测量，由于实验计划是进行24小时不间断测量，每1小时进行一次主动测量。最后选取了星期五晚上开始测量，测量到星期天早上结束，实际测量时长为36小时

* 1. **IP筛选**

注意到，由于整理的IP数量巨大，如果要全部运行一次会非常耗费时间，并且我们还需要跑数十次，于是对每个区域的IP都进行了筛选。

具体的筛选方法是，根据该组IP的总数量，确定一个选取系数P，然后在该组中随机选取总数量/P数量的IP进行测量，作为该区域的结果。

本实验进行了两次，第一次根据运行时间调整了P，第二次整理了数据。但最后在运行时还是出现了时间偏差，推测是python内部的内存泄漏导致操作系统建立虚拟内存带来的额外时间。

1. **一些失败的尝试**
   1. **网络拓补测量**

曾经想进行网络拓补测量，在查阅资料后，得到两个方法

一是课上讲的SNMP方法，但经短时间研究表明，该方法要求两段都必须有该软件客户端才可进行。

二是traceroute命令，它可以进行主动发包，留下沿途所有服务器的信息并返回。看起来很适合，但在实际操作中，traceroute有以下几个问题：1是运行时间过慢，一个IP就要运行1分钟至5分钟不等；二是沿途的服务器有些会进行信息保护，在运行时不会提供IP地址。

* 1. **网络带宽测量**

网络带宽曾经也是我们想测量的数据，但是同样也考虑到需要两端客户端的问题就放弃了

另外，网络带宽测量还有两个可能需要注意的地方，一是这个测量结果受我们本地网络带宽限制，可能不一定准确；另一方面该测量可能会导致网络拥堵，带来一些不安全因素。

1. **数据分析**
   1. **数据规律**

根据我们获得的所有数据，我们发现了如下规律。延迟和ping的成功率有大致上的反比例关系，具体来说，14个区域的数据显示了这种关系，只有3个区域的数据没有非常明显的显示这种关系。

* 1. **时间规律**

在实验最开始的时候我们会猜测，ping的成功率应该符合人们的生活作息。然而实际数据显示的结果和我们的猜测有一定的差距。

首先居住区和工作区的ping的成功率的确和时间有关系，一半在半夜1点钟开始减少，早上10点左右开始增多，我们认为这是工作族上班睡觉导致的原因，当然也有一些长时间不管的设备也会一直占用ip对我们的分析产生一定的影响。

其次对于居住区的ping的成功率和时间的关系，没有出现***非常***明显的半夜低，工作时间高，这使我们感到好奇，不过联想到一般家里人都用的路由器很可能是长时间不关的以及其他的因素。给我们最大的体会是网络的复杂性不是可以预计的。

* 1. **奇怪的规律**

我们测试了上海大学城，发现了大家在周六晚上的时候，ping的成功率非常高，但是到了周日晚上，ping的成功率和周六相比少了非常之多，我们估计可能上海的大学也熄灯吧。

1. **实验总结**

本次实验我们学习到了数据的采集，代码的编写，以及网络测量的基本实践。我们也感受到了真实数据和实验数据当中的区别。

这次实验也让我们学习了网络测量的基本方式，特便是一些工具的使用。如何加快测试的数据我们也做了尝试。总体上来说我们从这次实验里面学到了很多东西。