120181080503

软件1802

李海

爬取10万个网站的产品服务信息

root@localhost用户 密码为123456 数据库为testdb

库中表为temp\_icp\_web2，读取到的内容存放在content表中

1

geturl()函数

先从temp\_icp\_web2表中读取公司名称，id，url

识别url是否可用，将不可用的url存入数据库，标记内容为地址有误

将识别好的url和公司ID作为字典数组返回

2

getcontent（url，1）函数，根据编码不同，自行更改编码，读取一个页面下的有关产品服务的信息，当发现时，读取父亲节点下所有节点的text信息，发现链接时，进入链接再次调用函数getcontent（url，0），数字为0防止一直循环，将读取内容去除空格和换行后返回为一个字符串，当地址无法访问后，将返回地址无法访问，当没发现有关内容时，返回无相关内容。

3

MMain（）主函数设置进程池，会先把数据库清空，之后有8个进程进行爬取并存放内容置数据库中。

4

getip()，从http://www.xicidaili.com/nn/中获取随机ip，作为ip代理，但会减慢运行，故没有使用

Content.sql为空数据库

Content 前500个.sql为爬取其中约500个公司的数据，爬取的数据在数据库后半部分，前半部分为地址有误的公司。