项目需求：





实现思路：

研究网站，地点省市区由javascript动态生成，抓包可以看出查询操作是通过发post请求实现的，返回json字符串，里边包含了我们需要提取的起步价、元/公斤等信息。

Post请求的表单form\_data:

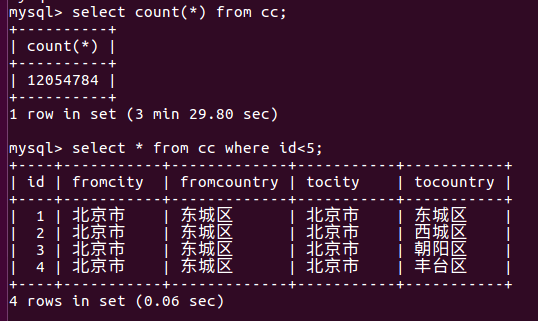
form\_data={'priceQueryVo.conCity':conCity,'priceQueryVo.conCounty':conCounty,

'priceQueryVo.shipperCity':shipperCity,'priceQueryVo.shipperCounty':shipperCounty}

所以我们就需要先获取到所有的出发地和接收地的市区组合。

1. 想办法获取所有的市区（如南京市江宁区，取下来总共3472个）
2. 想办法获取所有的市区的组合，形成需要提交的post表单（到这步，我们有3472\*3472=12054784个组合，也就是我们需要对改网站post爬取1200w次。而这个次数似乎没法通过查找共性来减少。） 到这里才发现这个项目对初学者并不简单，必须引入数据库和多线程甚至分布式才能解决。
3. 在mysql中存储1200w个市区组合
4. 从数据库读取这些市区组合包装成form\_data提交post请求到网站爬取json字符串，从json字符串中获取我们要的数据，insert到数据库中。

Mysql中，table cc存储所有的市区组合：



最终结果如下，我们获取对应的送货时间价格等信息到数据库中：



这个项目是在网上学了别人的爬取糗事百科和简单的爬取知乎后，自己在自由职业网站[www.upwork.com上找的python](http://www.upwork.com上找的python)爬虫类项目。此项目对新手来说数据量较大，需要爬取站点1200w次，每次又有相应的数据需要处理存贮，发现之前自己想的单线程爬取存储到文件里根本不能完成任务。必须要引入数据库和多线程，花了一早上研究了mysql和python多线程threading，因为上一份工作对oracle比较熟悉，发现mysql跟oracle学起来基本就是一个东西了。

参考文献：

Mysql自行百度

https://www.ibm.com/developerworks/aix/library/au-threadingpython/

代码说明：

主体分两部分：

get\_all\_cc\_into\_db.py

获取所有需要提交表单的市区组合，并插入的数据库中的table cc中

scraping\_threads.py

多线程爬取数据，最终数据插入到数据库table final\_data中。

两个类ThreadUrl来读取queue中的表单数据然后从站点爬取数据写入到out\_queue中

DatamineThread读取out\_queue中的数据，并处理数据，得到我们需要的数据insert到final\_data中。