# 项目环境

1、系统：windows 10。

2、浏览器 ：firefox，版本：62.03-64位 (注：浏览器的版本需要与浏览器的driver文件相匹配，因此安装时设置不检查更新)。

3、项目语 ：python 版本：3.6.5 Anaconda（注：python 2.7，3.5版本的python中有些语法，方法的参数与3.6.5不匹配，可能无法执行。Anaconda中自带许多第三方包，推荐使用。）。

4、IDE : pycharm。

5、需额外安装的第三方库：selenium，BeautifulSoup。

6、执行电脑配置 ：4核8线程，16g内存。

# 执行步骤

1. 在pycharm 中打开get\_p2peye\_data 文件夹。

2、configure 文件设置 ：打开configure 文件，设置cpu\_num, thread\_num\_per\_cpu 与执行电脑相符。

3、执行爬取程序，程序名根据抓取的数据名来命名。保存在DataDone文件夹中。

4、执行collection程序：因某些数据的数据量较大，因此设计为分段执行以规避网站的反爬机制，同时也可减少内存占用。分段后的文件分别写在DataDone文件下的对应目录下，执行collection程序可将文件合并，并写在该目录下。注：已执行，再次执行时须将已经合并的大文件删除。

# 爬取建议

1. 网站不定期更新。
2. 网站高峰期爬取容易触发该网站的反爬机制，建议深夜与凌晨爬取。
3. 若执行电脑内存容量不足，建议调小cpu\_num或thread\_num\_per\_cpu以降低并发数。