1.

配置：

- scrapy 由于在windows系统中安装较为繁琐，建议在linux系统(ubuntu等)中安装。

代码中运用啦logger模块，或许在某些电脑中无法使用（未测试），可以将相关代码注释。

* mysql 其中账号密码可以在db\_handler.py中设定。

2.

运行：

scrapy crawl ZSpider -s JOBDIR=crawls/ZSpider-1

而非

scrapy crawl ZSpider

其中JOBDIR指定待爬网页的存放地点。可以通过CTRL+C暂停爬取，运行以上命令恢复爬取。

Setting.py中的DOWNLOAD\_DELAY为每个网页的爬取时间间隔，可以任意调节。

ZSpider.py中，可以将自己注册的账号对应的cookie添加到cookienameList中。

3.

实现过程：

通过查看网页源码，我们发现啦后台数据存放的地址。且对应地址中的内容直接为JSON格式存储，方便了我们的爬取。并且访问后台只需要有相关账号对应的cookie即可。

只需要变动地址中相应的ID和page\_num即可爬取对应条目和页码的数据。

我们每一次访问都采用随机的cookie（目前只有两个cookie），随机的user-agent，以防止其封号。

之后爬取时发现，若以3s的间隔还是有可能被封号。目前能想到的对应策略就是增加注册的账号，以扩大可选cookie的数量，相对延长了每个账号爬取的时间。

此外，由于数学题中含有大量的公式，以图片形式存储。爬取图片的地址在另一个域名下

（knowbox.cn），在该域名下爬取图片似乎不限制次数。之后我们可以考虑将图片单独爬取，以更快的频率。后续可以考虑进一步的实现。