# 作业2的总体要求：

写一个或者几个脚本，获取附件内的部分信息

1. 要求使用python3.5
2. 不要求使用任何爬虫“框架”，不要求使用任何自动化测试“框架”。我们的目的是学会获取web数据的技能和技术，不是要让大家提交一个工程
3. 在windows上运行即可，不反对在linux上运行
4. 要求提交相关的数据结果，但不需要全，每个题目运行一下有千把个示例数据即可（选择7题的可以提交百来个即可）。
5. 按组提交，在考试周的周一的14天前提交，过期不候
6. 请发邮件至: [xiaoqiong.zh@hotmail.com](mailto:xiaoqiong.zh@hotmail.com)，邮件中需标注小组名（你随便起，但是起好不能变），小组成员（名字，学号，email）及其分工，并标记组长大人。
7. 最后的包可以以大附件的形式发送至 [haidescs@foxmail.com](mailto:haidescs@foxmail.com) 邮箱。
8. 数据可以放在一个类似csv的文件里，也可以存数据库，存数据库的限mysql（不要用mongodb），并提供建库sql
9. 每个小组或者选择1-6题，或者只选择7题，当然，如果有英雄人物做完了所有7个题目，在整体大作业中都会有较大比例的加分。

# 附件1，复杂js的操纵

中国债券信息网

<https://www.chinabond.com.cn/>

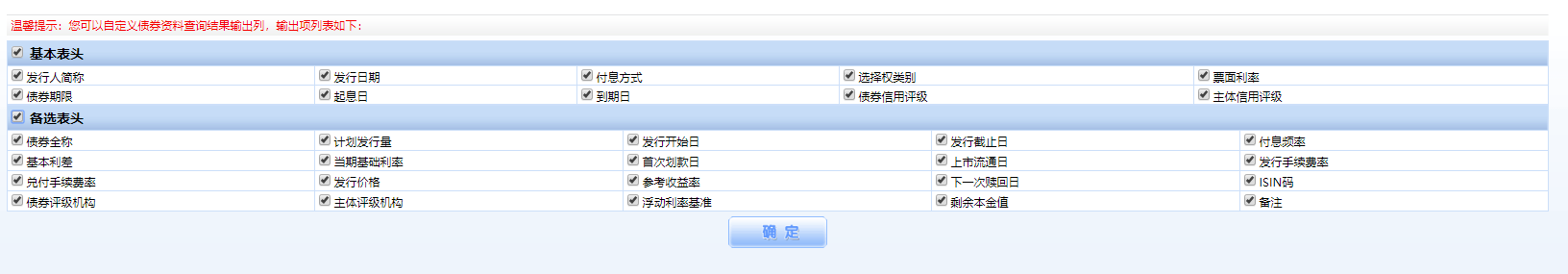
要求内容是这个查询结果：



这里的所有信息



需要注意的是，要求包括所有信息，也就说所有列



这个网站的数据集非常小，只有23000条数据，这里锻炼的是对javascript动态加载数据的精细操作，而且，这个网站的数据有一个陷阱。仔细分析查询结果页面url可以看出，其数据是动态、异步、分段加载的。比较复杂。

<https://www.chinabond.com.cn/jsp/include/EJB/queryResult.jsp?pageNumber=1&queryType=0&sType=2&zqdm=&zqjc=&zqxz=00&eYear2=0000&bigPageNumber=1&bigPageLines=500&zqdmOrder=1&fxrqOrder=1&hkrOrder=1&qxrOrder=1&dqrOrder=1&ssltrOrder=1&zqqxOrder=1&fxfsOrder=1&xOrder=12345678&qxStart=0&qxEnd=0&sWhere=&wsYear=&weYear=&eWhere=&sEnd=0&fxksr=--&fxjsr=--&fxStart=--&fxEnd=--&dfStart=-00-00&dfEnd=-00-00&start=0&zqfxr=&fuxfs=&faxfs=00&zqxs=00&bzbh=&sYear=&sMonth=00&sDay=00&eYear=&eMonth=00&eDay=00&fxStartYear=&fxStartMonth=00&fxStartDay=00&fxEndYear=&fxEndMonth=00&fxEndDay=00&dfStartYear=&dfStartMonth=00&dfStartDay=00&dfEndYear=&dfEndMonth=00&dfEndDay=00&col=28%2C2%2C5%2C33%2C7%2C21%2C11%2C12%2C23%2C25%2C0%2C3%2C31%2C32%2C6%2C8%2C9%2C10%2C13%2C14%2C15%2C16%2C17%2C19%2C22%2C24%2C26%2C27%2C29%2C20>

这里的bigPageLines=500的意思是一次加载500条数据，但是每一页只显示19条，所以在第27页的时候会重新加载下一个500条数据…… 这里是特别需要注意的地方。

# 附件2，微信公众号的操纵

获取微信公众号“什么值得买”的近100天的全部文章。公众号如下。



完成这项任务比较简单，但是有以下注意点：

1. 老师要求的是“近100天”的文章，可不是全部文章，所以如何获取每一篇文章的meta-data就是关键（当然，这里老师只要求获取文章发布时间这一个元数据）
2. 这个题目的目的不是要考察大家搭建模拟器获取移动数据的能力（请不要这样做）。而是研究问题的能力，和研究数据的能力。如果通过一种通用的可行方法，爬取微信公众号这种封闭空间的内容信息。好好研究，一定有收获。
3. 页面中有3种数据，第一种是文章url，文章内容，文章作者，文章发布时间这些固定的内容；第二种是阅读数，点赞数，再看数，留言这些时刻变化，但是也特别有用的内容；第三种是广告等时刻变化，而且也没啥用的内容。
4. 不需要保存图片和视频信息，文章不需要内容，只需要除了内容以外的“元数据”，例如题目、链接、作者、时间（注意对“今天”的处理）、由多少阅读量，多少赞

# 附件3，基于搜索的操纵

目标是通过百度，搜索所有我国目前的5家国有商业银行及12家股份制商业银行跟贸易金融和供应链金融相关的网页。

1. 首先要弄清楚这17家银行都有哪些，这个就不用大家去查了，17个银行的名字写死就行了
2. 搜索词有2个“贸易金融”和“供应链金融”
3. 要求必须是17家银行官网的内容
4. 要求必须是通过百度获取，而不是先跳转到各个官网，然后用各自官网的内建搜索来获取（虽然这样精准一些，但是不符合本题目的立意）。
5. 学有余力的同学可以尝试将结果数据按照产品介绍（就是详情页）和新闻进行一个简单分类
6. 更加厉害的同学可以尝试自动的梳理产品介绍页面，因为大多数情况下某个金融产品有多层（也就是多个）网页进行介绍。

# 附件4，基于查询的操纵

在金融界网站 http://trust.jrj.com.cn/product/data.shtml

获取信托期限在2年以上，且包含关键词REITs的信托产品，及其详细信息。如下：

得到的搜索结果页面如下



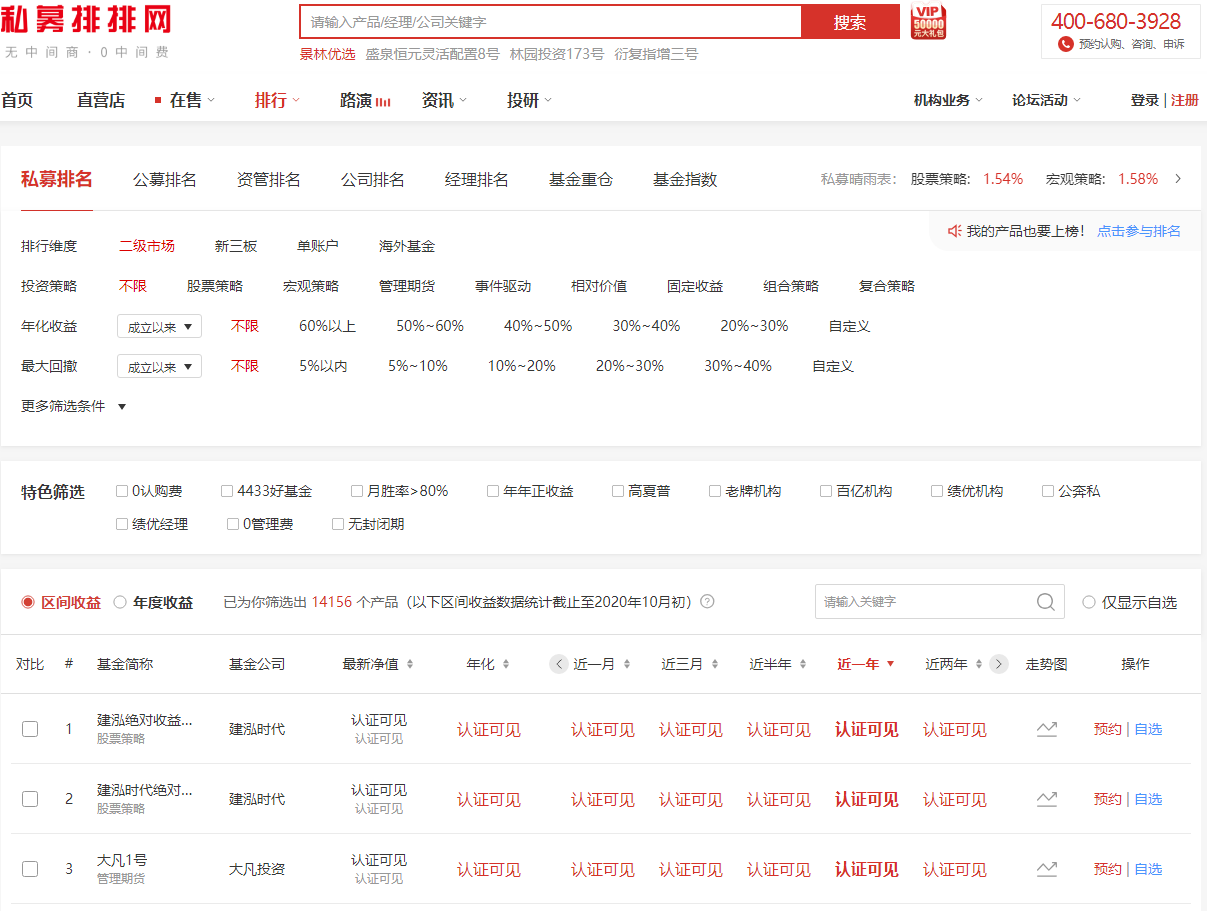
基于搜索结果的测试在自动化测试中是非常常见，也是非常重要的测试领域。因为根据用户和时间的不同，搜索结果也是“不确定”的。同时，这里除了考察操纵搜索功能之外，根据页面进行深度dig也是自动化测试和爬虫的基本技能。这里老师提出几个要求：

1. 数据结果放在2个不同的文件夹（数据库表）中，一个文件夹（数据库表）放信托产品，另一个文件夹（数据库表）中放这些信托产品所对应的信托公司。
2. 2年以上，是包括“2-3年”，“3-5年”，“5年以上”三个搜索结果的集合，所以要进行三次搜索。
3. 如果觉得困难，可以只填期限，不填产品名称reits，降低难度，不会扣分。
4. 搜索结果页面不包括“所有必须的”detail info，需要深挖到详情页面。
5. 如何去重，如何深挖，组织获取到的数据是本题的难点，需要良好的组织
6. 其实同学们有一个后门，就是 <http://trust.jrj.com.cn/list/companylist.shtml> 这个页面有所有注册的信托公司。也就是说正向的或者负向的都可以得到相同的结果
7. 包括两个维度的数据，信托产品详情，相关的信托公司详情

# 附件5，基于树形结构的操纵

私募排排网站 <https://dc.simuwang.com/>

获取所有私募基金的详情，以及所有基金公司



1. 这个不能基于查询，而是通过列表页一层一层挖下去。列表页有一部分基本信息和实时行情。但是我们不需要实时行情，我们需要的是各个基金产品的详情页，这就是一个树形结构
2. 第一层是



1. 每个点进去以后可以爬到每个基金的基本信息



1. 之后就是真正的详情，以及相关公司的情况。



1. 这里的基金公司详情理论上需要去重，但是如果嫌麻烦也可以不去重。

# 附件6，基于网状结构的操纵

请各位同获取哔哩哔哩所有认证帐号（包括个人认证和企业团体认证）的帐号的详细信息，包括他们的投稿和频道的部分基本信息就可以了。

这个只有一个要求：

1. 从某一个帐号出发，在他的关注人列表中寻找认证帐号，再从这些已经获得的认证帐号的关注人列表中寻找新的认证账号。

# 附件7，来一点机器学习和NLP的东西

请大家找出微信公众号、微博、知乎、bilibili、喜马拉雅、豆瓣中有对应关系的账号及其url

这里我给一个思路，

1. 以其中1-2个网站为root
2. 获取root网站中的账号（获取一定的数量，而且多平台运营的基本都是认证的大V，哈哈）
3. 用这些账号在其他网站中搜索
4. 对搜索结果的前几页进行NLP判断