****许多自学爬虫(python)的小伙伴因为没有经历过面试所以在找工作之前难免有些抓不住重点，虽然自己有些技术****

****但是因为发挥不好而错失工作机会，本人经过n次面试以后特总结以下面试常见问题，为想要转爬虫的小****

****伙伴提供一些参考。****

一.项目问题：

    一般面试官的第一个问题八成都是问一下以前做过的项目，所以最好准备两个自己最近写的有些技术

含量的项目，当然一定要自己亲手写过的，在别的地方看的源码，就算看的再清楚，总归没有自己敲的

了解的多。以下是抽出的几点

1.你写爬虫的时候都遇到过什么反爬虫措施，你是怎么解决的

2.用的什么框架，为什么选择这个框架(我用的是scrapy框架，所以下面的问题也是针对scrapy)

二.框架问题（scrapy）可能会根据你说的框架问不同的问题，但是scrapy还是比较多的

1.scrapy的基本结构（五个部分都是什么，请求发出去的整个流程）

2.scrapy的去重原理 （指纹去重到底是什么原理）

3.scrapy中间件有几种类，你用过那些中间件，

4.scrapy中间件再哪里起的作用（面向切面编程）

三.代理问题

1.为什么会用到代理

2.代理怎么使用（具体代码，请求在什么时候添加的代理）

3.代理失效了怎么处理

四.验证码处理

1.登陆验证码处理

2.爬取速度过快出现的验证码处理

3.如何用机器识别验证码

五.模拟登陆问题

1.模拟登陆流程

2.cookie如何处理

3.如何处理网站传参加密的情况

六.分布式

1.分布式原理

2.分布式如何判断爬虫已经停止了

3.分布式去重原理

七.数据存储和数据库问题

1.关系型数据库和非关系型数据库的区别

2.爬下来数据你会选择什么存储方式，为什么

3.各种数据库支持的数据类型，和特点，比如：redis如何实现持久化，mongodb

是否支持事物等。。

八.python基础问题

# 基础问题非常多，但是因为爬虫性质，还是有些问的比较多的，下面是总结

1.python2和python3的区别，如何实现python2代码迁移到python3环境

2.python2和python3的编码方式有什么差别（工作中发现编码问题还是挺让人不爽的）

3.迭代器，生成器，装饰器

4.python的数据类型

九.协议问题

# 爬虫从网页上拿数据肯定需要模拟网络通信的协议

 1.http协议，请求由什么组成，每个字段分别有什么用,https和http有什么差距

2.证书问题

3.TCP,UDP各种相关问题

十.数据提取问题

1.主要使用什么样的结构化数据提取方式，可能会写一两个例子

2.正则的使用

3.动态加载的数据如何提取

4.json数据如何提取

十二.算法问题

# 这个实在不好总结，比较考验代码功力，大部分会让你写出时间复杂度比较低的

算法。小伙伴们要善用python的数据类型，对python的数据结构深入了解。

以上就是总结内容，欢迎小伙伴们共同探讨。每个公司各有特点，但是这些算是基础，也是常见问题。

祝大家能找到理想工作，不写bug