**AID1912班的小伙伴们，考试即将开始**

**说明: 此次考试分为两部分，热身体和实操题，考试时间无限量！**

**注意：各位请把此次考试当作一次企业面试，内容均为爬虫前3天及之前所学内容，面试开始，剑指爬虫，祝你好运，加油！**

1. **热身题（每题10分，共30分）**

**1、使用requests模块向** [**http://www.abcdefg.com/**](http://www.abcdefg.com/) **（测试网站）发出请求，并获取响应内容html**

**前提：此网站的编码为 gb2312 ，解码时遇到不识别的字符不能抛出异常**

**2、现有URL地址格式如下：**

**url = ‘http://tencent.com/q?city=北京&keyword=爬虫&index=1**

**URL地址种存在中文，请使用2种方法，将其处理为一个合法的URL地址**

**3、请从如下的html代码结构种使用正则表达式匹配出 工作地点、工作类别、工作时间、工作职责，要求最终在终端打印如下格式：**

最终打印结果如下（必须严格按照如下格式输出）：

工作地点: 深圳

工作类别: 产品类

发布时间: 2020年03月06日

工作要求: 1）努力学习的人；2）多敲代码的人；3)面试表现不紧张的人；

**html = “””**

**<p data-v-58f864bf="" class="recruit-tips">**

**<span data-v-58f864bf="">深圳</span>**

**<span data-v-58f864bf="">|</span>**

**<span data-v-58f864bf="">产品类</span>**

**<span data-v-58f864bf="">|</span>**

**<span data-v-58f864bf="">2020年03月06日</span>**

**</p>**

**<ul>**

**<li data-v-58f864bf="" class="explain-item">**

**1）努力学习的人；**

**<br>**

**2）多敲代码的人；**

**<br>**

**3) 面试表现不紧张的人；**

**</li>**

**</ul>”””**

**二、实操题（上机题，每题35分，共70分）**

**题目1：抓取电影天堂中电影数据（二级页面数据抓取）**

**页面URL地址:** <https://www.dytt8.net/html/gndy/dyzz/>

**抓取页数：**共210页 – 因抓取速度较慢，故写出代码可以运行即可

**具体数据：**到每个电影的电影详情页，抓取 电影名称+下载链接

**要求:**

1. **一级页面：抓取电影详情页链接**
2. **二级页面：抓取电影名称、下载链接**
3. **所抓数据存入到MySQL数据库中**
4. **使用redis做成增量爬虫，新更新的电影在最上面出现**

**具体抓取数据如图：**





**题目2：纵横中文网小说全站抓取 – 请使用正则表达式 或者 xpath完成**

**URL地址:** [http://book.zongheng.com/store/c0/c0/b0/u0/p**{}**/v9/s9/t0/u0/i1/ALL.html](http://book.zongheng.com/store/c0/c0/b0/u0/p%7b%7d/v9/s9/t0/u0/i1/ALL.html)

**其中 {} 的值为 1 2 3 4 5 代表不同页**

**具体数据： 所有小说的小说名称、具体章节内容**

**要求：**

1. **一级页面：提取 小说链接**
2. **二级页面：提取 开始阅读对应的小说具体章节内容的链接**
3. **三级页面：提取 目录 对应的链接（链接中有此小说所有章节的明细及URL地址）**
4. **四级页面：提取 此小说所有章节的链接**
5. **五级页面：提取 具体的小说内容**









