Python工程师测试题（基础爬虫）

**提示：制作测试题前，请仔细阅读该文档**

**一、携带资料和工具**

1、自带电脑，以免环境不兼容，无法展示作品；

2、自带个人纸质简历一份；

**二、测试题目**

**1、开发环境：在Linux平台下**

**2、统一使用Python3写测试题**

**3、利用Python3开发爬虫，抓取指定网站下分类栏目页下的文章**

采集的网站网址/关键词：比特币https://www.5186.net/

采集内容及标准：

1）文章标题

2）文章页面的URL（网址）

3）文章正文内容：采集的文章必须**保持原文章格式及段落**

4）输出格式：爬虫采集到的内容，最终需导出为**Excel格式**（格式模板见附件“Excel样板”，Excel版本不限），如图1-1

5）采集数量：**1000篇**

6）非文章内容要清洗掉，如广告、特殊字符、文章转载于XXX、html代码等

7）去重功能：采集过程中，爬虫自动对文章的URL（网址）和文章标题做去重处理，无论两者哪个出现重复，爬虫都不再抓取该文章。

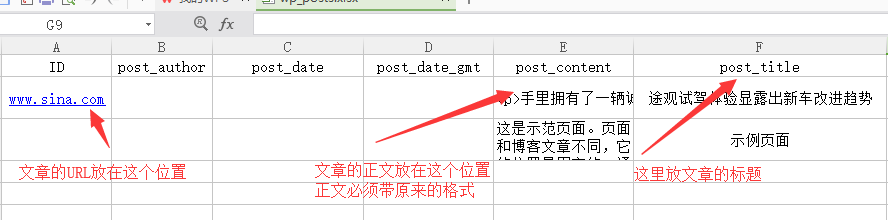


图1-1

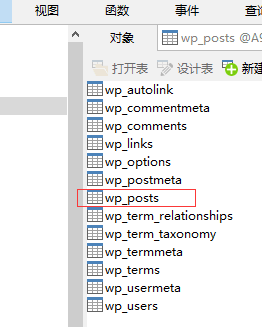
**4、将Python采集的文章内容，根据规定格式导入本地的数据库**

数据库类型：MySql

数据库表单名称：wp\_posts

表单内容及表头名称：见附件中的Excel表头名称

工作原理：Python爬虫抓取的文章，在完成采集后，自动导入本地MySql数据库内



请面试者在收到测试题之后，1个工作日完成上述内容，并提供**Python源码、MySql数据库**及**采集后的Excel表格**，发送至 [shanghaiseozhaopin@163.com](mailto:shanghaiseozhaopin@163.com)

邮件格式： 面试时间+姓名+系统平台（Lunix）

附件：文章采集（数据库）Excel样板