## 实习项目：

采集相关网址的指定内容，经过采集和抽取，并将数据存入mysql数据库。

相关网址：<http://www.cbdio.com>

项目采用python编写。

之前已经完成的内容为将爬取的内容存为本地的json文件，现在将数据存入mysql数据库。

项目文件为‘0105\_cbdio.py’，实现了爬取网站的一级和二级页面内容，并将指定的内容存入了数据库，同时，在代码中加入了uuid模块，并根据url生成GUID；另外还使用了scrapy框架同样实现了上述功能，但是目前使用scrapy之爬取了一级页面的内容，爬取二级的功能有待增加。

目前项目存在的问题还有：

1、Windows下装scrapy没有成功（需要安装VS），在linux下成功装上。

2、因为代码中使用了uuid模块来生成GUID，在使用python2.7运行的时候出现了如下问题：

File "0105\_cbdio.py", line 58, in SpideDataToMysql

guid = uuid.uuid3(uuid.NAMESPACE\_DNS, second\_url)

File "/usr/lib64/python2.7/uuid.py", line 528, in uuid3

hash = md5(namespace.bytes + name).digest()

UnicodeDecodeError: 'utf8' codec can't decode byte 0xa7 in position 1: invalid start byte

经过分析发现问题的原因是python的版本，uuid模块在python35和python27下的代码存在区别：

hash = md5(namespace.bytes + bytes(name, "utf-8")).digest()

hash = md5(namespace.bytes + name).digest()

可以看到，python3里面先将name重新进行了解码和编码，因此没有错误。

最后，在linux下安装了python3，然后使用Python3执行了0105\_cbdio.py，显示成功。

3、因为scrapy安装在python27上面，因此在使用scrapy工具爬取网站的时候，如果使用uuid来生成GUID，会出现错误，无法爬取成功，因此，目前暂且使用‘TIME’来作为‘GUID’存入数据库，后面再解决这个问题。

PS：之前文档里面提到的在linux下将数据存为json文件出现的编码问题已经解决。