Python 课程作业--对百度贴吧帖子进行爬虫

施文章 1120150631，王炜浩 1120150622

这篇文档 是**Python课程作业的说明文档合集**，您可以通过这些文档来了解程序的设计思路，如何使用这个程序，以及该程序最终运行结果。同时本文还列出了程序的待改进的地方。本文档使用markdown编辑器编辑

- **设计文档**

> 设计目标

```
1.对百度贴吧中的任意帖子内容进行爬虫抓取
2.考虑到帖子内容很多是楼主连载，设置了可以选择是否只抓取楼主发帖内容
3.为了方便离线阅读，我们这里将抓取到的内容分析并保存到文件，格式为TXT格式
```

1.URL格式的确定

首先，我们先观察一下百度贴吧的任意一个帖子。

比如：
`http://tieba.baidu.com/p/4138767607?pn=1`
`tieba.baidu.com`指的是百度二级域名，指向百度贴吧服务器
`/p/4138767607`是服务器某个资源，也是帖子的哦地址定位符，“帖子的神秘代码”
`pn`指的是页码
这里将*URL*两部分，一部分为基础部分，一部分为参数部分
上面例子`http://tieba.baidu.com/p/4138767607`是基础部分
`pn=1`是参数部分
2.页面抓取
用urllib2库来试着抓取页面内容。定义一个类名叫BDTB(百度贴吧)，一个初始化方法，一个获取页面的方法。其中，有些帖子我们想指定给程序是否要只看楼主，所以我们把只看楼主的参数初始化放在类的初始化上，即init方法。另外，获取页面的方法我们需要知道一个参数就是帖子页码，所以这个参数的指定我们放在该方法中

3.提取相关信息
在提取相关信息时候，主要分为三个部分，
1）提取帖子标题
2)提取帖子页数
3）提取正文内容
具体代码见主程序
4.将提取内容格式化便于输出阅读
在程序中定义了一个format格式函数
5.最后将获取的数据存入到文件中
`file = open("XX.txt","w")`
`file.writelines(obj)`
代码详见主程序

- **使用文档**

该程序可以在配有python安装环境下运行，如果没有，使用时候需要安装python
在命令窗口中运行该命令

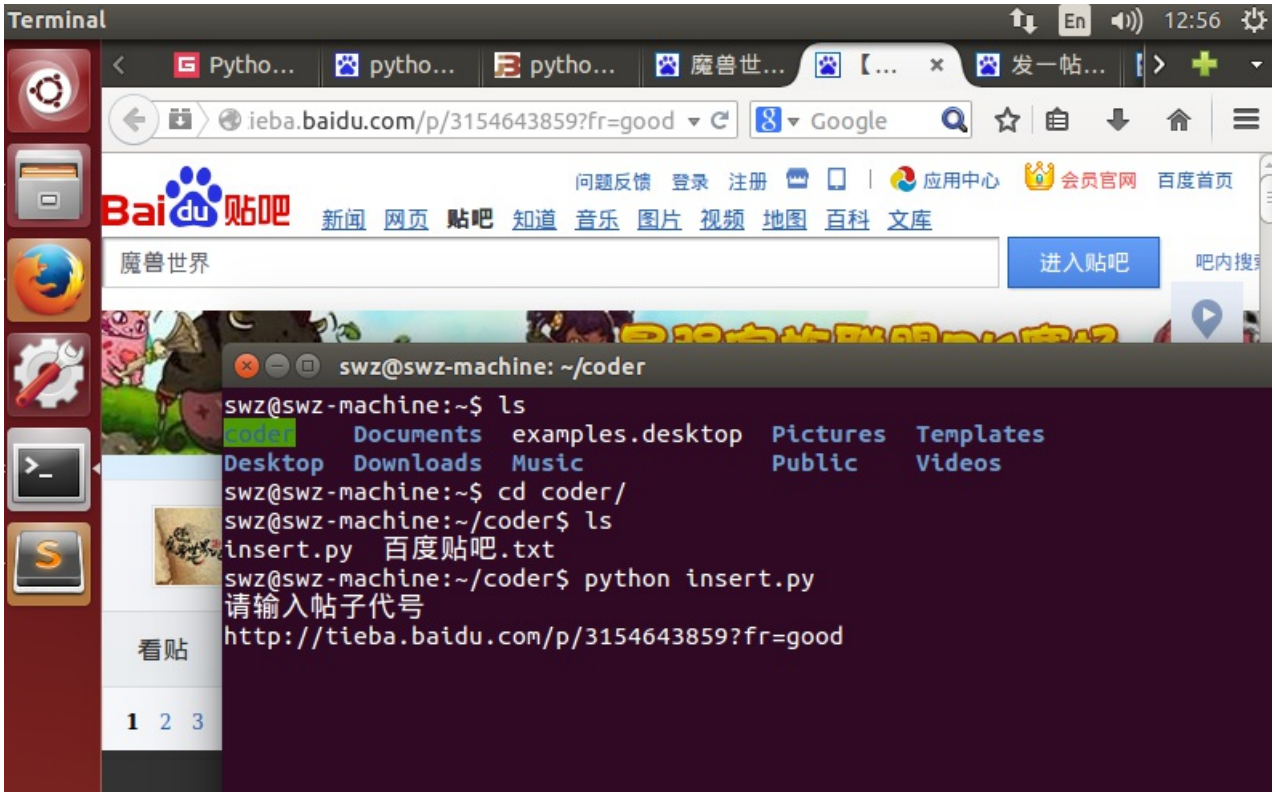
```
python insert.py
```

即可运行，按照出现的提示操作，确定 输入 1 表示 true,相反 输入 0 表示 false

- **测试文档**

我在现在ubuntu 14 环境下 运行结果如下图

选择的帖子的地址如上图

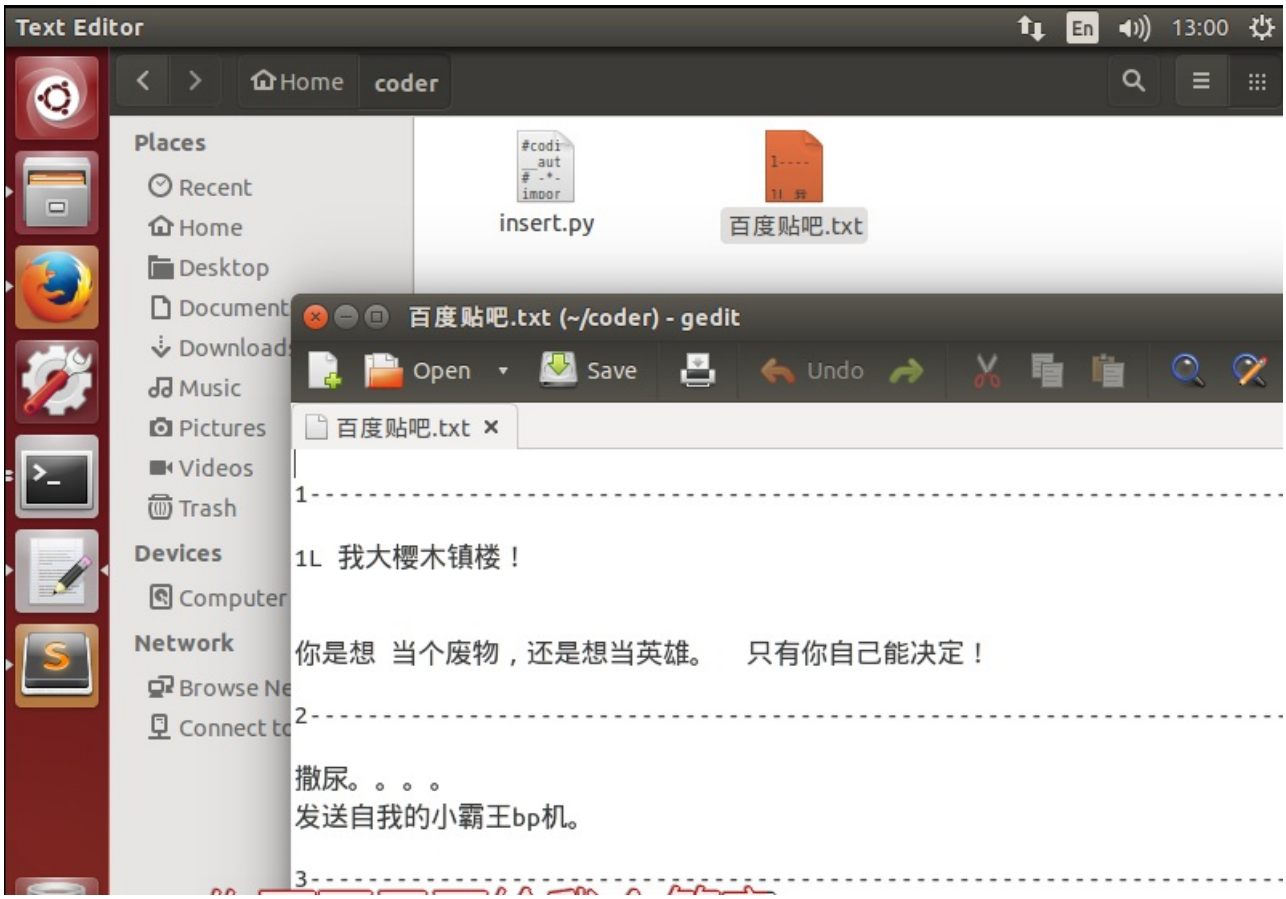


这里的帖子代号是：3154643859

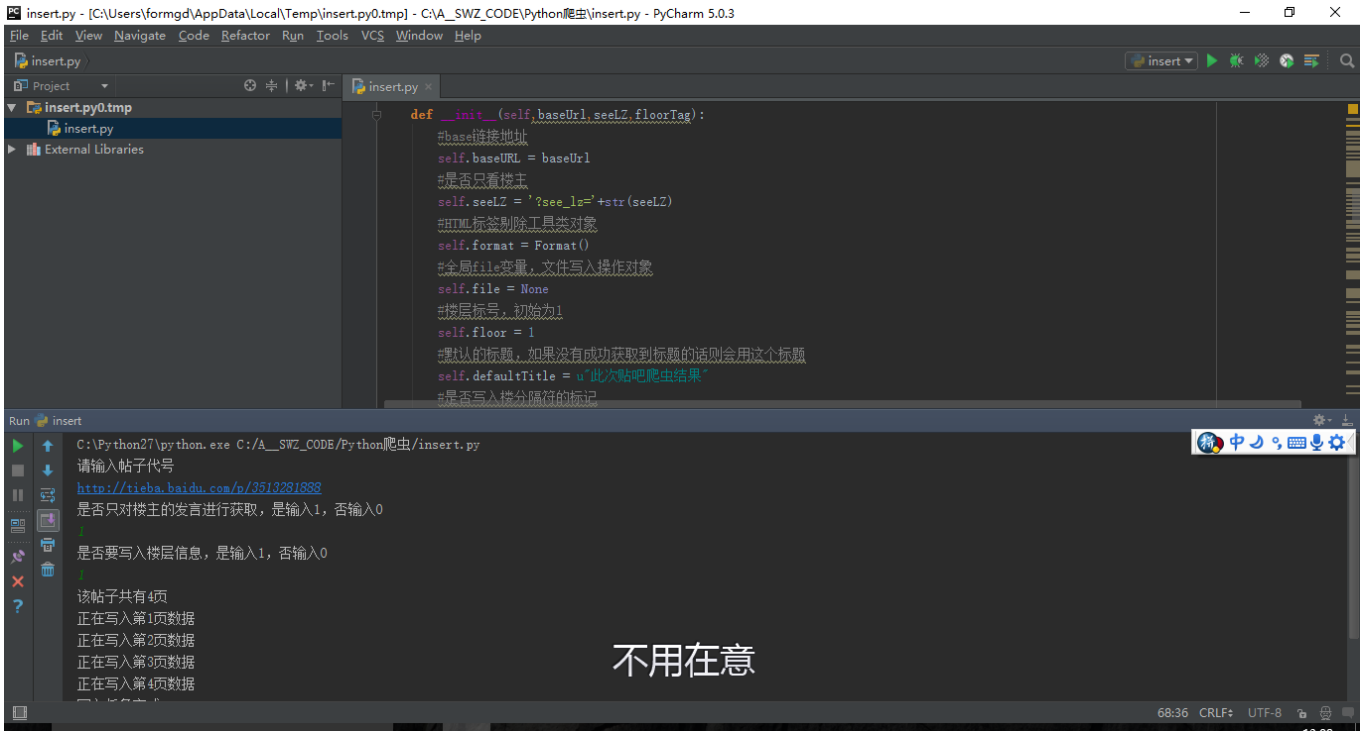
```
swz@swz-machine: ~/coder
swz@swz-machine:~$ ls
coder     Documents  examples.desktop  Pictures  Templates
Desktop   Downloads  Music             Public    Videos
swz@swz-machine:~$ cd coder/
swz@swz-machine:~/coder$ ls
insert.py  百度贴吧.txt
swz@swz-machine:~/coder$ python insert.py
请输入帖子代号
http://tieba.baidu.com/p/3154643859?fr=good
是否只获取楼主发言，是输入1，否输入0
1
是否写入楼层信息，是输入1，否输入0
1
该帖子共有83页
正在写入第1页数据
正在写入第2页数据
正在写入第3页数据
正在写入第4页数据
正在写入第5页数据
正在写入第6页数据
正在写入第7页数据
```

```
swz@swz-machine: ~/coder
正在写入第62页数据
正在写入第63页数据
正在写入第64页数据
正在写入第65页数据
正在写入第66页数据
正在写入第67页数据
正在写入第68页数据
正在写入第69页数据
正在写入第70页数据
正在写入第71页数据
正在写入第72页数据
正在写入第73页数据
正在写入第74页数据
正在写入第75页数据
正在写入第76页数据
正在写入第77页数据
正在写入第78页数据
正在写入第79页数据
正在写入第80页数据
正在写入第81页数据
正在写入第82页数据
正在写入第83页数据
写入任务完成
swz@swz-machine:~/coder$
```

结果如下图



又在windows 10 环境下 运行结果如下图



结果如下


Python爬虫
文件
主页
共享
查看
此电脑 › 本地磁盘 (C:) › A_SWZ_CODE › Python爬虫
搜索"Python爬虫"
名称
修改日期
类型
大小
insert.py 2016/1/5 11:35 PY 文件 6 KB
百度贴吧.txt 2016/1/5 11:02 文本文档 409 KB
此次贴吧爬虫结果.txt 2016/1/5 11:22 文本文档 18 KB
快速访问
桌面
下载
文档
图片
7941977RUL
Python爬虫
TDDownload
源图片
OneDrive
此电脑
视频
图片
文档
下载
音乐
桌面
本地磁盘 (C:)
DVD 驱动器 (D:)
网络
3 个项目

如果没有你 - 萧敬腾