## 实验八: python爬虫

### 一 实验目的

1)robots协议

2)tomcat的安装

3)搭建单机开发环境:最简单的网页

4)创建自己的robots.txt

### 二 实验设备

电脑主机: CPU i7 6700 ;内存 8G ; 硬盘 2TB ; Intel 芯片主板;

显示器: 20-英寸 LCD.

软件环境: windwos7,python3.5,pycharm或IDLE

### 三 实验内容

**1 打开豆瓣网页：**<https://www.douban.com/>

1 读取 robots.txt文件，截图填入报告

2 从禁止爬虫爬取的子目录里面选择一个填入报告，例如：

User-agent: \*

Disallow: /subject\_search

表示对于任何一种爬虫，豆瓣网站上/subject\_search路径下的资源不允许爬取

**2 安装apache tomcat**

完成安装后打开:http://127.0.0.1:8080,截图,填入实验报告

### 3 编写一个简单的网页

在tomcat的webapps目录下新建python目录,在里面创建index.html文件,输入以下内容:



写好以后,使用浏览器打开: <http://127.0.0.1:8080/python>

调整浏览器窗口大小,把浏览器截图填入实验报告,注意必须截取浏览器显示网址的部分

### 4 仿照豆瓣的爬虫协议，创建自己的robots协议，使用浏览器访问，截图，填入报告。

注意思考：robots.txt文件应该放在哪一个文件夹下

### 四 实验报告要求

**提交电子档**