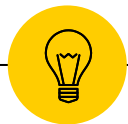


Lab2



《Web应用开发》 / 任课教师：罗志一

计算机科学与技术学院



概览 Overview

- 数据解析与持久化
- 中期作业提交
- 课程大作业分组



正则表达式是一种基于字符的语言

标记字符

- 元字符
- 量词

子表达式

- 嵌套子表达式

什么是基于字符的语言？

正则表达式像Python语言一样定义了一些具有特定语义的关键字，只不过这些关键字大多都是字符。

我们称这些字符为**标记字符 (marker character)**

即它们出现时并不代表字符本身，而是有固定的语义。

除了元字符和量词外的其他字符在正则表达式中出现时都代表自己本身。

```
import re
```



正则表达式语法要素：标记字符

元字符

每个元字符默认匹配一位字符串

.	匹配除换行符以外的任意字符
^	匹配字符串的开始
\$	匹配字符串的结尾
a b	匹配字符a或字符b
\w	匹配字母或数字或下划线
\W	匹配非字母或数字或下划线
\s	匹配任意空白符
\S	匹配非空白符
[aeiou]	匹配字符组中的字符（注：可由区间构造字符组，如[a-z0-9]）
[^XYZ]	匹配除了字符组中字符的所有字符
(匹配子表达式开始
)	匹配子表达式结束



正则表达式语法要素：标记字符

量词

量词修饰它的前一位字符；控制前面的元字符出现的次数

*	重复0次或更多次
+	重复1次或更多次
?	重复0次或1次（出现或不出现）
{n}	重复n次
{n,}	重复n次或更多
{n,m}	重复n次到m次

量词也可以修饰量词

- .*** 贪婪匹配（*表示尽可能多的去匹配）
- .*?** 惰性匹配（?修饰*，表示让*取一个可以匹配上的最小量值）



正则表达式语法要素：标记字符

量词

文本	这是第一堂课，这是第二堂课
正则表达式	匹配结果
.*课	这是第一堂课，这是第二堂课
.*?课	这是第一堂课

量词也可以修饰量词

.* 贪婪匹配（*表示尽可能多的去匹配）

.*? 惰性匹配（?修饰*，表示让*取一个可以匹配上的最小量值）

```
import re
text = "这是第一课堂"
```

```
pattern = re.compile(".*课")
m = pattern.search(text)
```

```
m = re.search(".*课", text)
```

```
if m is not None:
    print(m.group())
```



正则表达式语法要素：标记字符

转义字符

当需要匹配用于正则匹配模式的特殊字符（例如：`.`）时，就需要用到转义符了，即在要匹配的特殊字符前面加反斜线`\`转义一下即可（例如：`\.`）。

文本	1人民币=3585.67越南盾
正则表达式	re.search匹配结果
<code>\d.\d</code>	358
<code>\d\\.\d</code>	5.6



正则表达式语法要素：子表达式

嵌套子表达式

- 包裹在圆括号内的子表达式作为一个整体参与匹配
- 执行匹配后，除了总匹配结果，子表达式匹配结果也会被存下来
- 给子表达式(分组)起名字：(?P<分组名字>子正则表达式)



正则表达式语法要素：子表达式

嵌套子表达式

➤ 给子表达式(分组)起名字：(?P<分组名字>子正则表达式)

```
1.import re
2.s = """<div class="西游记"><span id="10010">中国联通</span></div>"""
3.# re.S这个flag的作用是：让元字符.能匹配换行符，怕断行
4.COMPILED_REGX = re.compile(r'<span id="( ?P<id>\d+)">( ?P<name>\w+)</span>', re.S)
5.result = COMPILED_REGX.search(s)
6.print(result.group())
7.# 获取id组的内容
8.print(result.group("id"))      # 10010
9.# 获取name组的内容
10.print(result.group("name"))    # 中国联通
```



字符串的检索、匹配和替换

➤ re.search & re.findall

re.search全文去检索匹配，找到一个结果就返回，返回match对象。re.search可以返回匹配正则表达式的第一个内容，但是如果想要获取匹配正则表达式的所有内容就要借re.findall方法了。

➤ re.match

从头开始匹配，找到一个结果就返回，返回match对象，好比默认在正则表达式最开始加了个^元字符。

➤ re.sub

可以借助re.sub来进行文本内容替换

例如把一串文本中的所有数字都去掉：`content = re.sub('\d+', '', content)`



编码问题

◎ 使用UTF-8编码

```
with open('xx.html', 'w', encoding='utf-8', errors='ignore') as outf:  
    outf.write(text)
```

```
with open('xx.html', 'r', encoding='utf-8') as f:  
    text = f.read()
```



网页解析

● 直接演示



数据持久化

- json
- pickle



中期作业提交

- 11.13日截止
- 最迟在11.14日周一上课结束前拷贝给我

课程大作业分组统计





开发工具Pycharm

用学校邮箱注册JetBrains账户并登录

account.jetbrains.com

Help

Welcome to JetBrains Account

- Access your purchases**
and view your order history
- Identify expired and outdated licenses,**
order new licenses and upgrades
- Manage your company licenses**
and distribute them to end users

Sign in with existing account

Email address or Username

Password

Sign In [Forgot password?](#)

Or sign in with:

Google

Not registered yet?

Create JetBrains Account

Your email address

Sign Up



开发工具Pycharm

● 申请免费的学生或教师license

2. 第二年license到期，
申请新的license。

更早 (4)

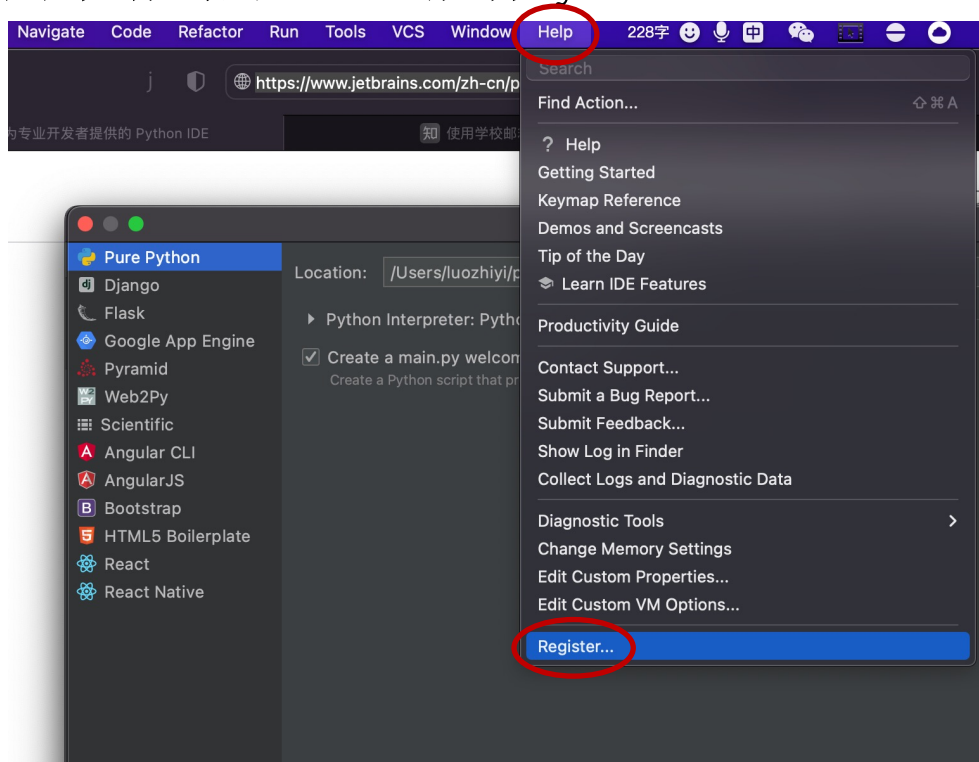
<input type="checkbox"/>	JetBrains Sales		[收件箱] License Certificate for your JetBrains Educational Pack / Order D375986090 ...Please use your JetBrains Account to access JetBrains ...	5月25日
<input type="checkbox"/>	JetBrains Account		[收件箱] JetBrains Educational Pack confirmation ...Kind Regards, The JetBrains team www.jetbrains.com ...	5月25日
<input type="checkbox"/>	JetBrains Sales		[收件箱] License Certificate for your JetBrains Educational Pack / Order D374028785 ...Kind Regards, The JetBrains team www.jetbrains.com ...	2021-05-26
<input type="checkbox"/>	JetBrains Account		[收件箱] JetBrains Educational Pack confirmation ...Kind Regards, The JetBrains team www.jetbrains.com ...	2021-05-26

1. 首次申请license，
有效期一年。



开发工具Pycharm

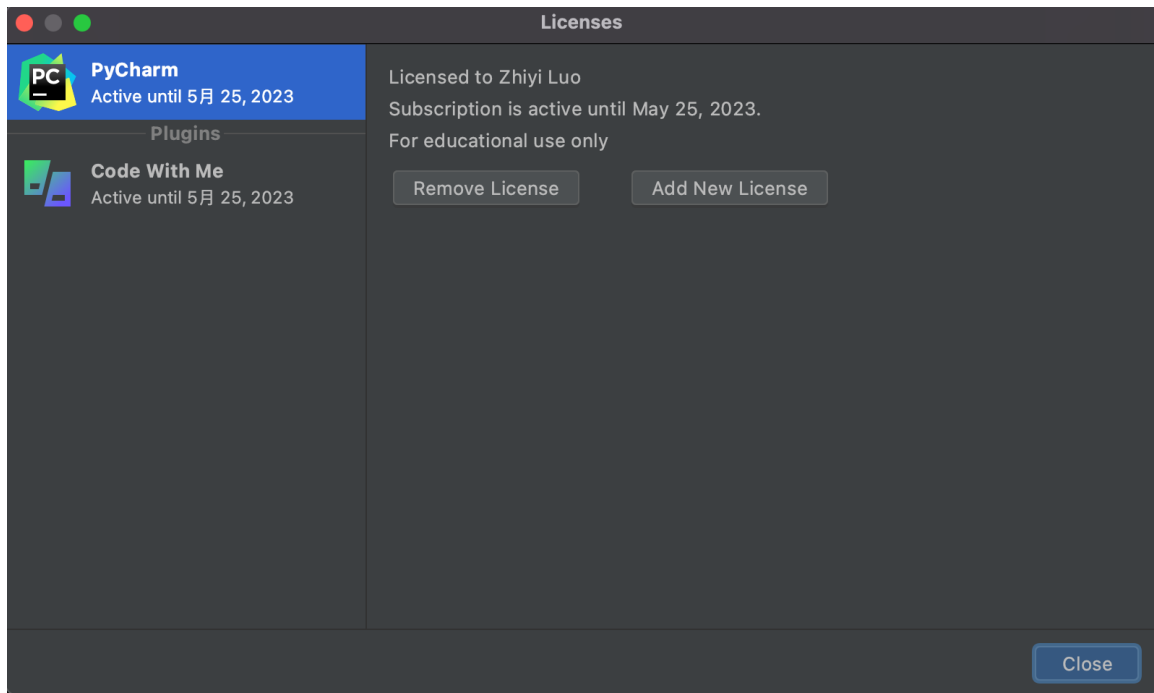
用申请到的license激活Pycharm





开发工具Pycharm

用申请到的license激活Pycharm





开发环境：Anaconda

🕒 下载Anaconda Installer

- 官方下载链接：

<https://www.anaconda.com/products/distribution>



ANACONDA

Products ▾

Pricing

Solutions ▾

Resources ▾

Partners ▾

Blog

Company ▾

Contact Sales

Individual Edition is now

ANACONDA DISTRIBUTION

The world's most popular open-source Python distribution platform

浏览器会根据当前使用的操作系统为你推荐安装文件，点击“Download”下载即可。





开发环境：Anaconda

◎ 从清华大学镜像下载Anaconda

- 如果安装包下载速度过慢，可以使用Anaconda国内源（例如：清华大学镜像）进行下载。
- Anaconda清华源的下载列表链接为：
<https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/>



开发环境：Anaconda

从清华大学镜像下载Anaconda

The screenshot shows the 'mirrors.tuna.tsinghua.edu.cn' website. The page title is '清华大学开源软件镜像站' (Tsinghua University Open Source Software Mirror). The navigation bar includes links for HOME, EVENTS, BLOG, RSS, PODCAST, and MIRRORS. The main content area displays the 'Index of /anaconda/archive/' with a table of files. The table has three columns: File Name, File Size, and Date. The files listed are various Anaconda installers for different operating systems and architectures, including Mac OS X, Windows, Linux, and ARM64. The last update date is 2022-09-01 05:51.

File Name ↓	File Size ↓	Date ↓
Parent directory/	—	—
Anaconda3-2022.05-MacOSX-arm64.sh	304.8 MiB	2022-06-08 01:42
Anaconda3-2022.05-MacOSX-arm64.pkg	316.4 MiB	2022-06-08 01:42
Anaconda3-2022.05-Windows-x86.exe	487.8 MiB	2022-05-11 02:36
Anaconda3-2022.05-Windows-x86_64.exe	593.9 MiB	2022-05-11 02:36
Anaconda3-2022.05-MacOSX-x86_64.sh	584.0 MiB	2022-05-11 02:36
Anaconda3-2022.05-MacOSX-x86_64.pkg	591.0 MiB	2022-05-11 02:36
Anaconda3-2022.05-Linux-x86_64.sh	658.8 MiB	2022-05-11 02:35
Anaconda3-2022.05-Linux-s390x.sh	279.8 MiB	2022-05-11 02:35
Anaconda3-2022.05-Linux-ppc64le.sh	367.3 MiB	2022-05-11 02:35
Anaconda3-2022.05-Linux-aarch64.sh	567.6 MiB	2022-05-11 02:35
Anaconda3-2021.11-Windows-x86_64.exe	510.3 MiB	2021-11-18 02:14



开发环境：Anaconda

- ◎ 从清华大学镜像下载Anaconda
 - 下载完成之后，双击安装包进行安装。



开发环境：Anaconda

◎ 从清华大学镜像下载Anaconda

○ 使用帮助文档在：

<https://mirrors.tuna.tsinghua.edu.cn/help/anaconda/>

- 在用户目录下配置`.condarc`文件
- macOS和Linux用户可直接在用户目录`/home/username`下创建`.condarc`文件，并写入配置内容。
- Windows用户无法直接创建名为`.condarc`的文件，可先执行`conda config --set show_channel_urls yes`命令生成该文件，再写入配置内容。



开发环境：Anaconda

● 创建Python虚拟环境

- `conda create -n <环境名称>`
- 例如，可以使用如下命令将新建环境命名为webcourse，并指定该环境的Python版本为3.9: `conda create -n webcourse python=3.9`

● 查看Anaconda中的所有虚拟环境

- `conda info --envs`

● 激活指定环境

- `conda activate <环境名称>`

● 退出当前激活的环境

- `conda deactivate`

如果.condarc中配置了清华源
此处无需翻墙即可成功。

```
-> conda create -n webcourse python=3.9
Collecting package metadata (current_repodata.json): done
Solving environment: done
```

```
## Package Plan ##
```

```
environment location: /Users/luozhiyi/local/anaconda3/envs/webcourse
```

```
added / updated specs:
```

```
- python=3.9
```

```
The following packages will be downloaded:
```

package	build		
ca-certificates-2022.07.19	hecd8cb5_0	124 KB	defaults
certifi-2022.6.15	py39hecd8cb5_0	154 KB	defaults
ncurses-6.3	hca72f7f_3	857 KB	defaults
openssl-1.1.1q	hca72f7f_0	2.2 MB	defaults
pip-22.1.2	py39hecd8cb5_0	2.4 MB	defaults
python-3.9.12	hdfd78df_1	10.3 MB	defaults
readline-8.1.2	hca72f7f_1	321 KB	defaults
setuptools-63.4.1	py39hecd8cb5_0	1.1 MB	defaults
sqlite-3.39.2	h707629a_0	1.2 MB	defaults
tk-8.6.12	h5d9f67b_0	3.1 MB	defaults
tzdata-2022a	hda174b7_0	109 KB	defaults
xz-5.2.5	hca72f7f_1	244 KB	defaults
zlib-1.2.12	h4dc903c_2	94 KB	defaults
Total:		22.1 MB	

```
The following NEW packages will be INSTALLED:
```

ca-certificates	anaconda/pkgs/main/osx-64::ca-certificates-2022.07.19-hecd8cb5_0
certifi	anaconda/pkgs/main/osx-64::certifi-2022.6.15-py39hecd8cb5_0
libcxx	anaconda/pkgs/main/osx-64::libcxx-12.0.0-h2f01273_0
libffi	anaconda/pkgs/main/osx-64::libffi-3.3-hb1e8313_2
ncurses	anaconda/pkgs/main/osx-64::ncurses-6.3-hca72f7f_3
openssl	anaconda/pkgs/main/osx-64::openssl-1.1.1q-hca72f7f_0
pip	anaconda/pkgs/main/osx-64::pip-22.1.2-py39hecd8cb5_0
python	anaconda/pkgs/main/osx-64::python-3.9.12-hdfd78df_1
readline	anaconda/pkgs/main/osx-64::readline-8.1.2-hca72f7f_1
setuptools	anaconda/pkgs/main/osx-64::setuptools-63.4.1-py39hecd8cb5_0
sqlite	anaconda/pkgs/main/osx-64::sqlite-3.39.2-h707629a_0
tk	anaconda/pkgs/main/osx-64::tk-8.6.12-h5d9f67b_0
tzdata	anaconda/pkgs/main/noarch::tzdata-2022a-hda174b7_0
wheel	anaconda/pkgs/main/noarch::wheel-0.37.1-pyhd3eb1b0_0
xz	anaconda/pkgs/main/osx-64::xz-5.2.5-hca72f7f_1
zlib	anaconda/pkgs/main/osx-64::zlib-1.2.12-h4dc903c_2

```
Proceed ([y]/n)? y
```

Proceed ([y]/n)? y

Downloading and Extracting Packages

sqlite-3.39.2	1.2 MB	#####	100%
readline-8.1.2	321 KB	#####	100%
xz-5.2.5	244 KB	#####	100%
setuptools-63.4.1	1.1 MB	#####	100%
pip-22.1.2	2.4 MB	#####	100%
certifi-2022.6.15	154 KB	#####	100%
ncurses-6.3	857 KB	#####	100%
tzdata-2022a	109 KB	#####	100%
zlib-1.2.12	94 KB	#####	100%
tk-8.6.12	3.1 MB	#####	100%
openssl-1.1.1q	2.2 MB	#####	100%
ca-certificates-2022	124 KB	#####	100%
python-3.9.12	10.3 MB	#####	100%

Preparing transaction: done

Verifying transaction: done

Executing transaction: done

```
#
# To activate this environment, use
#
#     $ conda activate webcourse
#
# To deactivate an active environment, use
#
#     $ conda deactivate
#
(base)
10:29:11 with luozhiyi in ~ via 🍷base took 2m 7s
→ conda activate webcourse
(webcourse)
10:50:37 with luozhiyi in ~ via 🍷webcourse
→ which python
/Users/luozhiyi/local/anaconda3/envs/webcourse/bin/python
(webcourse)
10:50:59 with luozhiyi in ~ via 🍷webcourse
→ conda deactivate
(base)
```



开发环境：Flask

- 激活webcourse环境
 - `conda activate webcourse`
- 在当前环境中安装flask
 - `conda install -c anaconda flask`

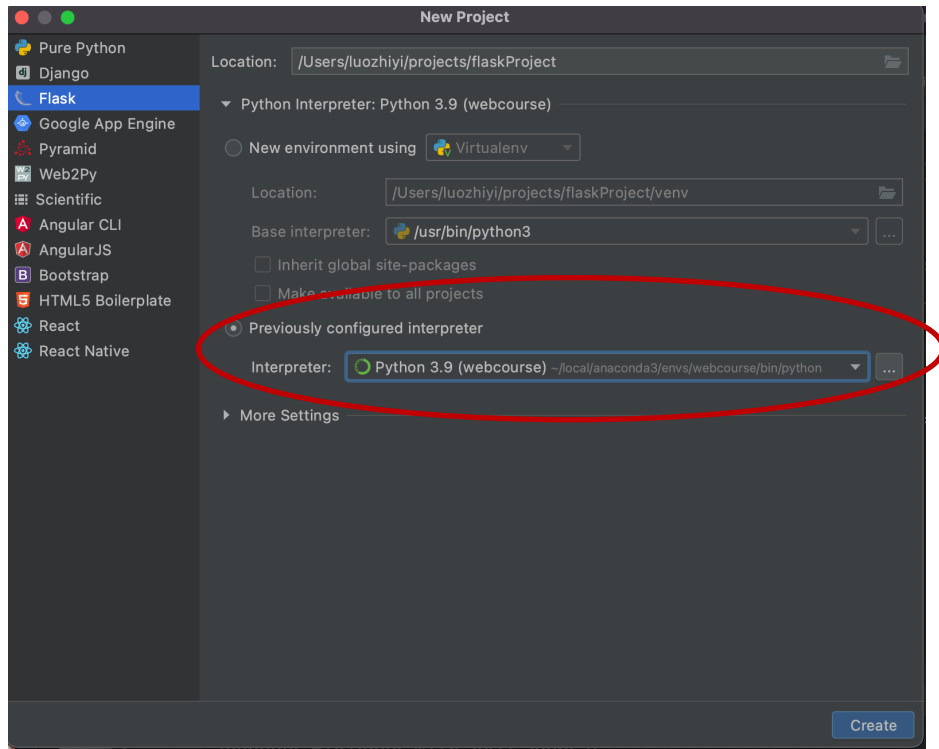
```
→ python
Python 3.9.12 (main, Jun 1 2022, 06:36:29)
[Clang 12.0.0] :: Anaconda, Inc. on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import flask
>>> █
```

恭喜！安装成功！



开发第一个Web应用

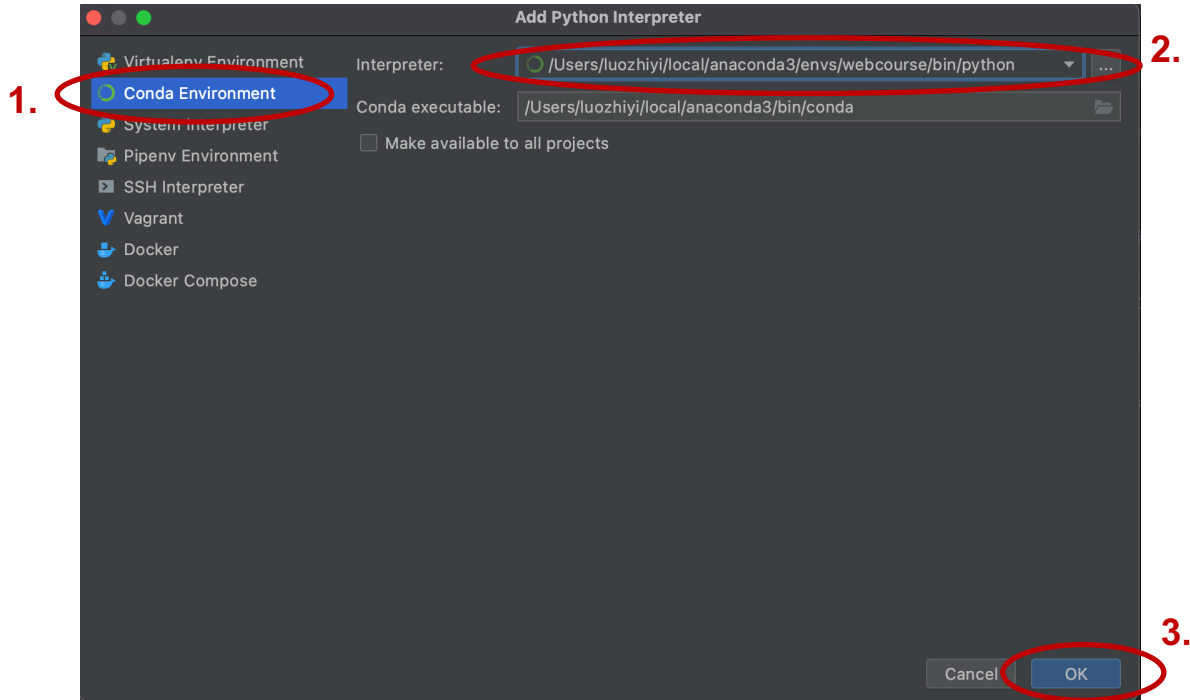
在Pycharm中创建一个Flask项目





开发第一个Web应用

在Pycharm中创建一个Flask项目





开发第一个Web应用

- 在Pycharm中创建一个Flask项目

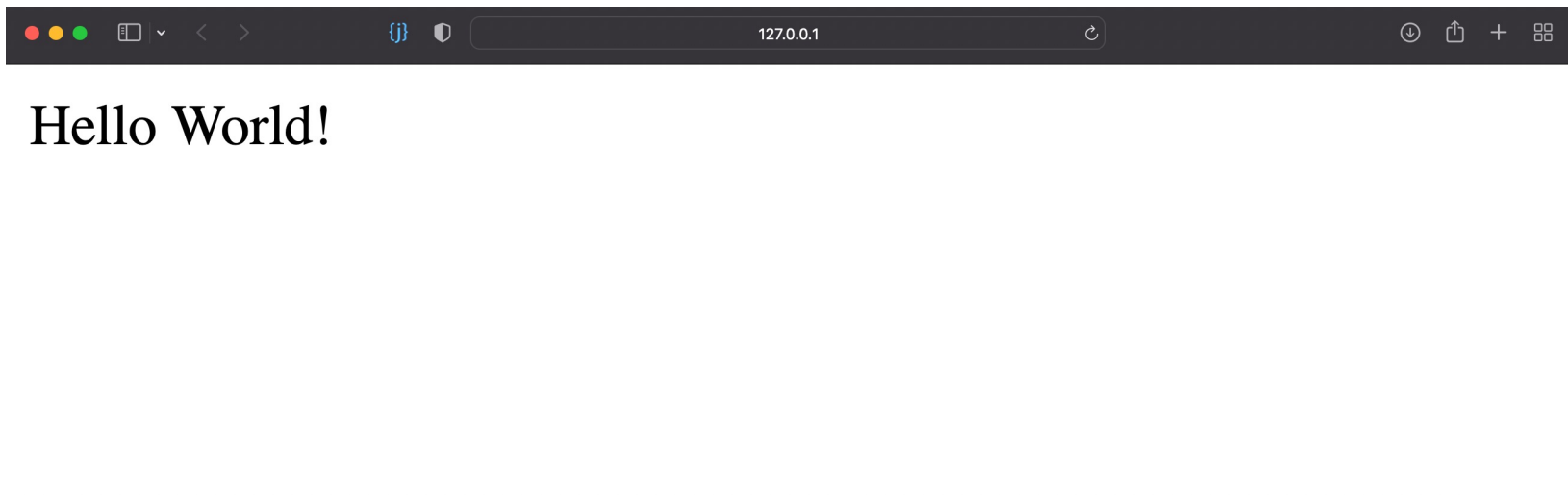
```
1  from flask import Flask
2
3  app = Flask(__name__)
4
5
6  @app.route("/")
7  def hello_world():
8      return "Hello World!"
9
10
11  if __name__ == "__main__":
12      app.run(debug=True)
```

创建app.py文件



部署Web应用

- 运行app.py
 - python app.py
 - 访问<http://127.0.0.1:5000/>





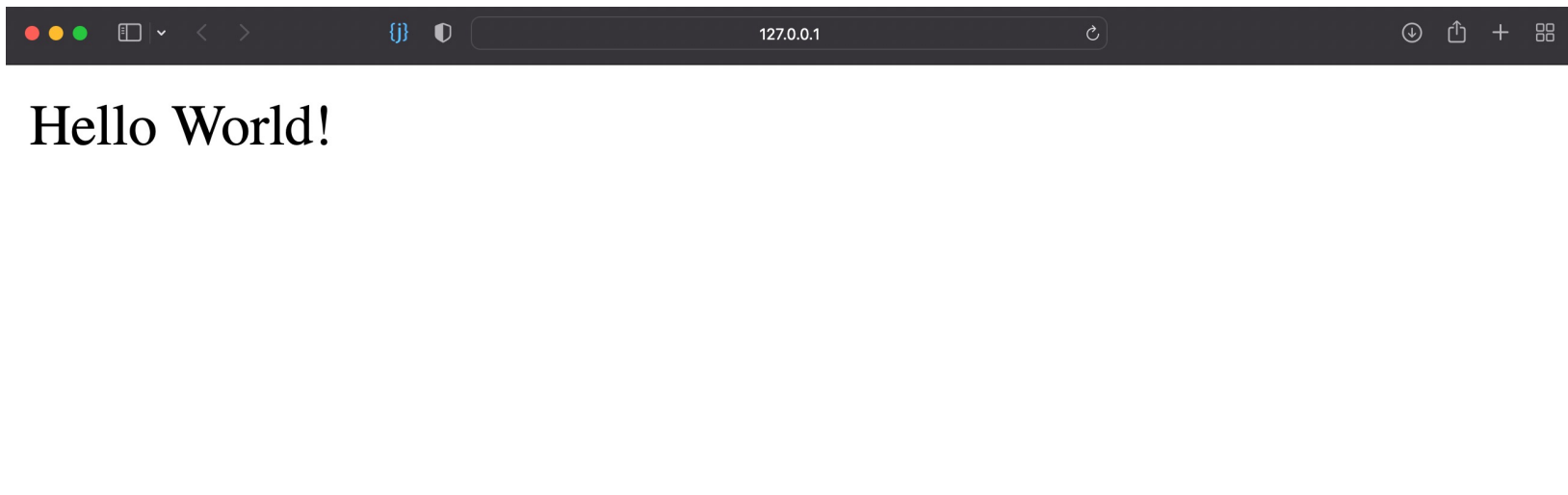
部署Web应用

运行app.py

- python app.py

- 访问 <http://127.0.0.1:5000/>

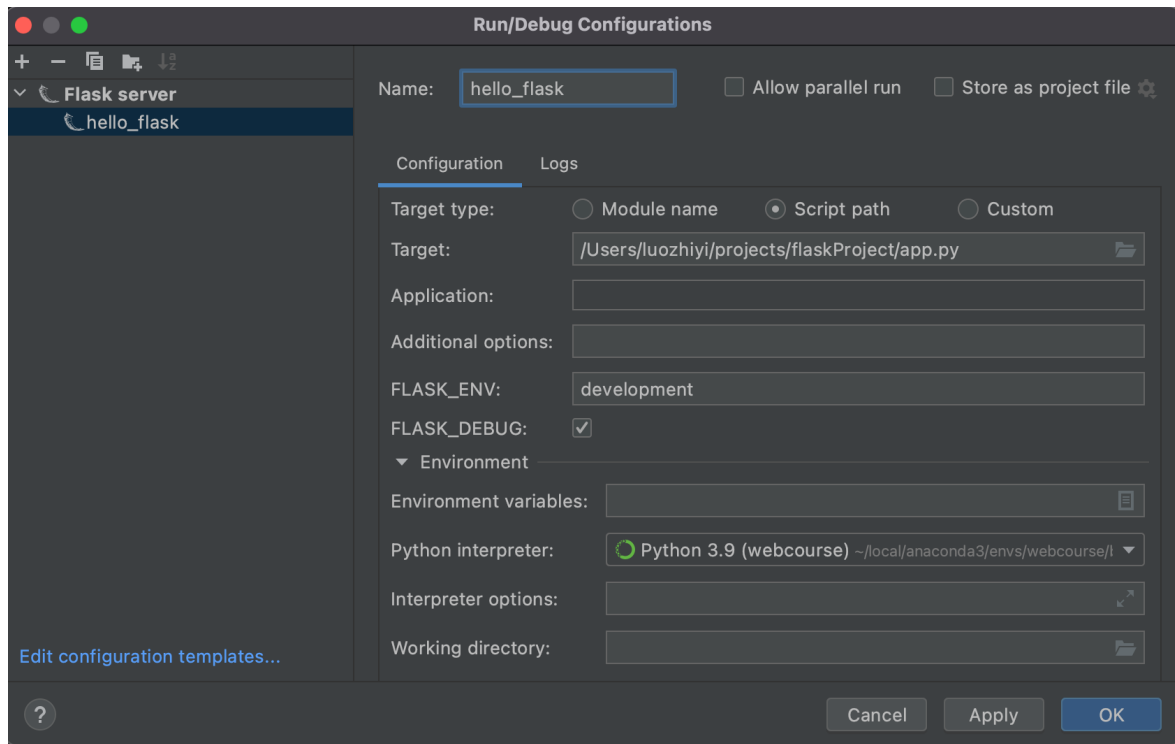
URL





开启Debug模式

方法一：如图配置Run Configurations, 勾选FLASK_DEBUG





开启Debug模式

- 方法二：传入debug=True参数

```
if __name__ == "__main__":  
    app.run(debug=True)
```

数据爬取






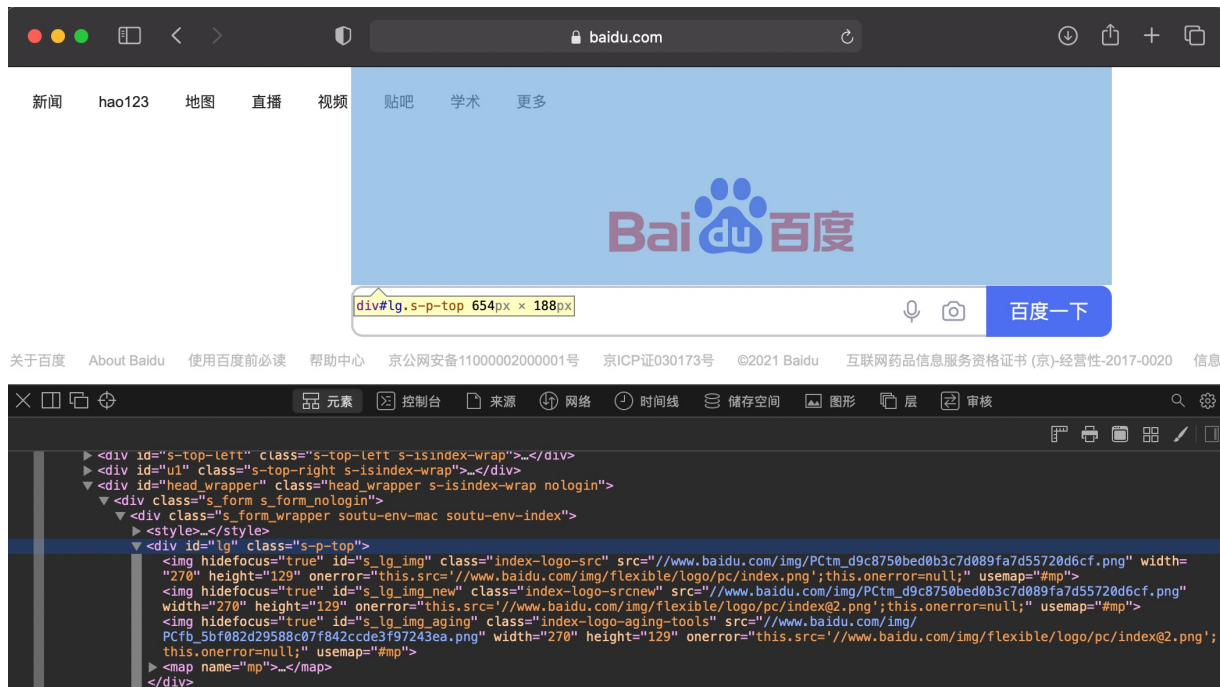
超文本标记语言HTML

- 一种用于创建网页的标准标记语言。
- 在互联网上，数以亿计的网页被保存和传播，制定一套通用标准语言来编写网页以整合各种资源就变得十分重要。就像说闽南语的福建人和说吴越语的浙江人之间是难以通过方言交流的，大家都需要说普通话来传递中文信息。**HTML**就是计算机在互联网上相互传递网页内容所使用的“普通话”，它由专门的组织制定并维护。
- **HTML**通过设计一系列的标签来标记各种资源，比如**img**标签标记图片，**p**标签标记文本段落等等。浏览器可以解析**HTML**编码的网页，再将解析内容展示给用户。例如，百度主页**HTML**源码中基于**img**标签将**Logo**图片嵌入到网页中，浏览器对其进行解析后展示给用户。



超文本标记语言HTML

- 例如，百度主页HTML源码中基于标签将Logo图片嵌入到网页中，浏览器对其进行解析后展示给用户。





请求与响应

- 我们知道爬虫需要模拟浏览器向服务器发送请求来抓取页面，那么具体该如何实现呢？Python强大的计算生态提供了功能齐全类库来帮助我们。下面我们介绍如何使用Python的请求库requests来完成这些请求，实现页面抓取。
- 激活环境： `conda activate pycourse`
- 安装requests库： `conda install requests`



示例：爬取百度主页

```
import requests
resp = requests.get('https://www.baidu.com', timeout=10)
resp.encoding = resp.apparent_encoding
print(resp.text[:500])
```


中期作业





爬取百度知道问答页面

- 每人需要爬取约**4000**左右的百度知道问答页面
- 请使用校园网访问：<http://10.11.195.12:1110/2crawl/>
- 根据学号姓名找到自己需要爬取的URL页面
- 数据格式说明：url[\t]关键词1[\t]关键词2



作业提交说明

- 新建一个目录**html/**，在该目录下存放各**html**页面。即爬取每个URL对应的网页源代码，并存成名为**<id>.html**的文件，放到**html**目录下。
 - 例如：将<https://zhidao.baidu.com/question/103948262.html>页面存储为**html/**目录下的**103948262.html**文件。
- 如果在给定的URL中存在一些无效URL页面，则将这些无效URL按照每行一个URL的格式写到一个名为**invalid.txt**的文本文件中。
- 最后，将**html**目录文件和**invalid.txt**文件一起打包，实验课上U盘拷贝提交。压缩包命名规则：姓名_学号[.zip|.7z|.rar|.tar.gz]
- **截止日期：2022年11月14日实验课上**



中期作业要求

- 中期作业需要在2022年11月14日前提交数据
- 评分标准：准确率 * 实际爬取页面数 / 作业分发的页面数
 - 准确率计算方法为：从提交的数据中随机抽取**100**个页面做验证，用通过验证的页面数目除以**100**即为准确率。