第三章 Python 语言与网络爬虫

提交日期：5 月 16 日
提交形式：作业以 word 或 pdf 形式，提交到助教邮箱，命名方式为“姓名”+“学号”
提交邮箱：邮箱地址 cwt@bupt.edu.cn ，邮件主题需包含 "信息表达"这四个字作为关键（以方便自动归档），提交成功后会有自动回复。

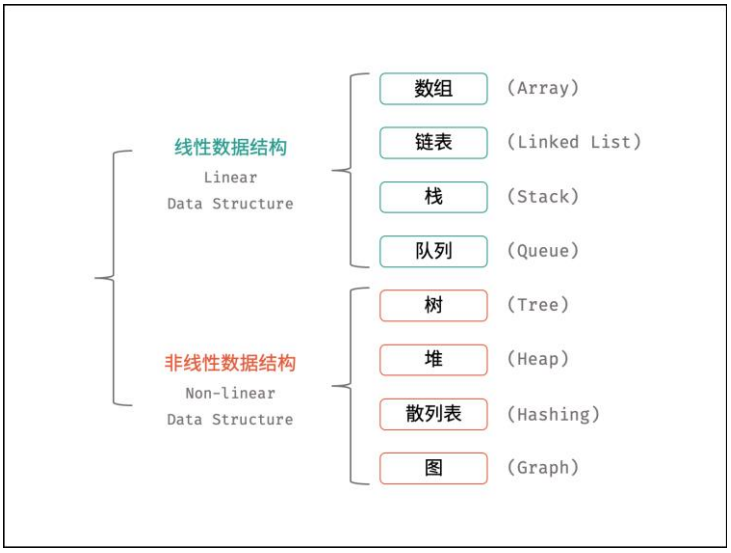
第一部分 python 基础

1、【综合题】请简要解释 python 中“==“和”is“的区别，并判断以下结果（True 或 False）：

请对下表空白处进行填空（True 或 False）

语句一	语句二	判断 a == b	判断 a is b
a = -256	b = -256		
a = "fdfff"	b = "fdfff"		
a = "fdfff"	b = a		
a = -256	b = a		
a = 5	b = 5		

2、【编程题】从以下八种数据结构中，**任选 3 种**数据结构给出其 Python3 语言的初始化与构建方法。



示例：
1、 可变数组

```
# 初始化可变数组
array = []

# 向尾部添加元素
array.append(2)
```

3、【编程计算】给出以下程序的输出结果：

```
a=20
b=-3
print(a+b)
print(a-b)
print(a*b)
print(a/b)
print(a%b)
print(a**b)
print(a//b)
print(abs(b))
print(int("1010",2))
print(float("3.14"))

c =complex(a,b)
print(c)
print(c.conjugate())

print(divmod(a, b))
```

4、【编程展示】打印九九乘法表

```
1*1=1
1*2=2  2*2=4
1*3=3  2*3=6  3*3=9
1*4=4  2*4=8  3*4=12 4*4=16
1*5=5  2*5=10 3*5=15 4*5=20 5*5=25
1*6=6  2*6=12 3*6=18 4*6=24 5*6=30 6*6=36
1*7=7  2*7=14 3*7=21 4*7=28 5*7=35 6*7=42 7*7=49
1*8=8  2*8=16 3*8=24 4*8=32 5*8=40 6*8=48 7*8=56 8*8=64
1*9=9  2*9=18 3*9=27 4*9=36 5*9=45 6*9=54 7*9=63 8*9=72 9*9=81
```

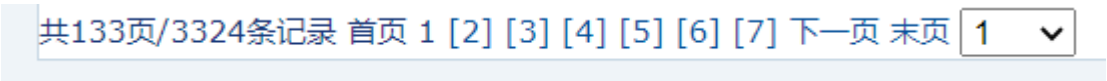
第二部分 网络爬虫

5、【实践题】从 https://www.dytt8.net/html/gndy/china/index.html 网站上爬取<尽可能多>的<国内电影>的电影标题、时间以及描述信息，并按照下图格式写入 csv 文件。

title	year	content
《乜代宗师》	2020	点击：0◎译 名 乜代宗师/The Grand Grandmaster ◎片 名 乜代宗師 ◎年 代 202
《杀手蝴蝶梦》	1989	点击：0◎译 名 My Heart Is That Eternal Rose ◎片 名 杀手蝴蝶梦 ◎年 代 1989
《造梦游戏》	2018	点击：0◎片 名 造梦游戏 ◎年 代 2018 ◎产 地 中国大陆 ◎类 别 喜剧 / 爱情 /
《诛仙 I /诛仙电影版》	2019	点击：0◎译 名 Jade Dynasty ◎片 名 诛仙 I /诛仙/诛仙电影版 ◎年 代 2019 ◎
《狐踪谍影》	2020	点击：0◎译 名 Fox Hunting ◎片 名 狐踪谍影 ◎年 代 2020 ◎产 地 中国大陆 ◎

方法提示：

1. 找到规律，判断每一页的 URL



2. 通过正则匹配，匹配时间，可能出现的文本描述类型如下
   - ◎年 代 2020
   - ⊙年 代 2018
   - <不存在关于时间的描述>
   - 【出品年代】2014
   - ◎时 间 2013
   - 上映日期: 2013-11-26

   （考虑到可能没有讲到正则匹配，这一步可以不做，只将电影的描述信息存入即可）
3. 爬取内容越多越好，最少爬取 10 页内容
4. 不要提交爬取的文件（太大了），截一张小图并提交源码（源码文件标注姓名， 写详细注释）