nsd1906_devops_day01

多进程

- windows不支持多进程
- os.fork返回值在父子进程中是不一样的
  - 父进程中，值非 0 （子进程的pid）
  - 子进程程中，值是 0
- 多进程编程的思路
  - 父进程只负责生成子进程
  - 子进程做具体的工作
  - 子进程工作结束后，应该exit退出

多线程

- 各操作系统均支持
- 程序，就是在磁盘上存储的可执行文件
- 当程运行时，就会出现进程。可以说，进程就是加载内存中的一系列指令。每个进程都拥有自己独立的运行空间。
- 进程还可以拥有多个线程。同一进程内的所有线程，共享进程的运行空间。
- 多线程没有僵尸进程的问题
- 多线程的编程思路：
  - 主线程（类似于父进程），负责生成工作线程
  - 工作线程（类似于子进程），做具体的工作

urllib模块

- 实现http客户端功能
- 包括4个子模块
  - request：最常用的模块
  - error：定义错误，实现异常
  - parse：用来解析和处理URL
  - robotparse：解析页面的robots.txt文件

```
>>> from urllib import request
>>> url = 'http://www.163.com'
>>> html = request.urlopen(url)
>>> html.readline()
b' <!DOCTYPE HTML>\n'
>>> html.read(10)
b'<!--[if IE'
>>> html.readlines()
```

修改请求头

```
>>> url = 'http://www.jianshu.com'
>>> html = request.urlopen(url)  # 403: Forbidden
# 简书它会做基本检查，发现请求不是正常的人为行为，将会拒绝
# 改变头部信息，骗过简书服务器，把客户端浏览器改为火狐
>>> url = 'http://www.jianshu.com'
>>> heads = {'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64; rv:52.0) Gecko/20100101
Firefox/52.0'}
>>> r = request.Request(url, headers=heads)
>>> html = request.urlopen(r)
>>> hrml.read()
```

url编码

- url只允许一部分ascii字符，其他字符需要编码

```
>>> url = 'https://www.sogou.com/web?query=元旦'
>>> html = request.urlopen(url)  # 报错，因为汉字是不允许的

>>> url = 'https://www.sogou.com/web?query=' + request.quote('元旦放假')
>>> url
'https://www.sogou.com/web?query=%E5%85%83%E6%97%A6%E6%94%BE%E5%81%87'
```

使用wget模块

```
(nsd1906) [root@room8pc16 day01]# pip install wget
>>> import wget
>>> url = 'https://img03.sogoucdn.com/app/a/100520021/6d573d4ca8f01112c416672f6b34ac49'
>>> wget.download(url, '/tmp/sea.jpg')
```

下载网易首页上所有的图片

- 找到所有图片的url
- 下载图片

1. 下载网易首页，保存为一个文件
2. 从网易文件的每一行中查找图片网址，收集网址到一个列表中
3. 遍历列表，下载文件