练习参考

IP地址

匹配合法的IP地址

```
192.168.1.150
0.0.0.0
255.255.255.255
17.16.52.100
172.16.0.100
400.400.999.888
001.022.003.000
257.257.255.256
```

`(\d{1,3}\.){3}\d{1,3}`

`(?:\d{1,3}\.){3}\d{1,3} # 400.400.999.888`

对于ip地址验证的问题

- 可以把数据提出来后，交给IP地址解析库 处理，如果解析异常，就说明有问题，正则的验证只是一个初步的筛选，把明显错误过滤掉。
- 可以使用复杂的正则表达式验证地址正确性
- 前导0是可以的

```
import socket
nw = socket.inet_aton('192.168.05.001') # 错了抛异常
print(nw, socket.inet_ntoa(nw))
```

```
分析：
每一段上可以写的数字有1、01、001、000、23、023、230、100，也就说1位就是0-9,2位每一位也是0-9,3位第一位只能0-2，其余2位都可以0-9
(?:([0-2]\d{2}|\d{1,2})\.){3}([0-2]\d{2}|\d{1,2}) # 解决超出200的问题，但是256呢？
200是特殊的，要再单独分情况处理
25[0-5]|2[0-4]\d|[01]?\d\d? 这就是每一段的逻辑
(?:(25[0-5]|2[0-4]\d|[01]?\d\d?)\.){3}(25[0-5]|2[0-4]\d|[01]?\d\d?)
```

提取文件名

选出含有ftp的链接，且文件类型是gz或者xz的文件名

```
ftp://ftp.astron.com/pub/file/file-5.14.tar.gz
ftp://ftp.gmplib.org/pub/gmp-5.1.2/gmp-5.1.2.tar.xz
ftp://ftp.vim.org/pub/vim/unix/vim-7.3.tar.bz2
http://anduin.linuxfromscratch.org/sources/LFS/lfs-packages/conglomeration//iana-etc/iana-etc-2.30.tar.bz2
http://anduin.linuxfromscratch.org/sources/other/udev-lfs-205-1.tar.bz2
http://download.savannah.gnu.org/releases/libpipeline/libpipeline-1.2.4.tar.gz
http://download.savannah.gnu.org/releases/man-db/man-db-2.6.5.tar.xz
http://download.savannah.gnu.org/releases/sysvinit/sysvinit-2.88dsf.tar.bz2
http://ftp.altlinux.org/pub/people/legion/kbd/kbd-1.15.5.tar.gz
http://mirror.hust.edu.cn/gnu/autoconf/autoconf-2.69.tar.xz
http://mirror.hust.edu.cn/gnu/automake/automake-1.14.tar.xz
```

```
.*ftp.*\.(?:gz|xz)
.*ftp.*/([^/]+(?:.xz|.gz))$    # 后面的分组就是文件名
(?<=.*ftp.*/)[^/]*\.(?:gz|xz) # 断言文件名前一定还有ftp，python的re模块不支持断言中使用*、+或{3,4}等不固定长度的匹配
```

匹配邮箱地址

```
test@hot-mail.com
v-ip@magedu.com
web.manager@magedu.com.cn
super.user@google.com
a@w-a-com
```

确定规则，邮箱名为字母、数字、下划线、减号、点号，域名一样

```
\w[\w.-]*@([\w-]+\.)+\w+
```

@前至少一个字符开头，可以有字母、数字、下划线、减号、点号

@后至少有一个字符后跟点，重复至少1次，最后以1个字符串结尾

匹配html标记

提取href中的链接url，提取文字“马哥教育”

```
<a href='http://www.magedu.com/index.html' target='_blank'>马哥教育</a>
```

有的时候可以这样写`<a href=http://www.magedu.com/index.html target='_blank'>马哥教育</a>`，属性attr不标准的写法不写单双引号，这个一般交给浏览器可以正常处理。

```
href=['"]?([^'"\s]*)
```

提取内容时，要考虑一下情况

```
<a href='http://www.magedu.com/index.html' target='_blank'>马哥教育</a><a>马哥教育1</a>
```

```
<a.*?>[^<>]*
```

a标签要到第一个右尖括号结束，不能贪婪。内容是其之后第一个左尖括号之前

匹配URL

```
http://www.magedu.com/index.html
https://login.magedu.com
file:///ect/sysconfig/network
```

```
(\w+)://(\S*)
```

匹配二代中国身份证ID

```
321105700101003
321105197001010030
11210020170101054x
17位数字+1位校验码组成
前6位地址码，8位出生年月，3位数字，1位校验位（0-9或x）
```

```
身份证验证
身份证验证需要使用公式计算，最严格的应该实名验证。
\d{17}[0-9xX]|\d{15}
```