全代码

```
import requests
url = "https://item.jd.com/2967929.html"
try:
    r = requests.get(url)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[:1000])
except:
    print("爬取失败")
```

爬取京东商品，r.text[:1000]表示取前 1000 行，防止数据过多影响

```
>>> import requests
>>> r = requests.get("https://www.amazon.cn/gp/product/B01M8L5Z3Y")
>>> r.status_code
503
>>> r.encoding
'ISO-8859-1'
>>> r.encoding = r.apparent_encoding
>>> r.text
'<!--\n        To discuss automated access to Amazon data please cont
act api-services-support@amazon.com.\n        For information about m
igrating to our APIs refer to our Marketplace APIs at https://develop
er.amazonservices.com.cn/index.html/ref=rm_5_sv, or our Product Adver
tising API at https://associates.amazon.cn/gp/advertising/api/detail/
main.html/ref=rm_5_ac for advertising use cases.\n-->\n<html><head><m
eta http-equiv="Content-Type" content="text/html;charset=utf-8"><titl
e>亚马逊</title><body style="text-align:center;"><br><div style="width
:600px;margin:0 auto;text-align:left;"><h2>意外错误</h2></div><br><div
style="width:500px;margin:0 auto;text-align:left;"><font color="red">
报歉，由于程序执行时，遇到意外错误，您刚刚操作没有执行成功，请稍后重试。或将
此错误报告给我们的客服中心：<a href="mailto:service_bj@cs.amazon.cn">ser
```

404 是因为 url 错误导致的吧，像来源审查导致的错误是 503

```
>>> r.request.headers
{'User-Agent': 'python-requests/2.11.1', 'Accept': '*/*', 'Accept-Encoding'
: 'gzip, deflate', 'Connection': 'keep-alive'}
>>> kv = {'user-agent':'Mozilla/5.0'}
>>> url = "https://www.amazon.cn/gp/product/B01M8L5Z3Y"
>>> r = requests.get(url, headers = {'user-agent':'Mozilla/5.0'})
>>> r.status_code
200
>>> r.request.headers
{'user-agent': 'Mozilla/5.0', 'Accept': '*/*', 'Accept-Encoding': 'gzip, de
flate', 'Connection': 'keep-alive'}
>>> r.text[:1000]
'\n\n\n  \n  \n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n
\n\n\n\n\n    \n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n    \n\n\n\n\n\n\n\n\n\n\n\n\
n\n\n\n    \n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n\n    <!doctype html><html cla
```

User-agent 表示为 python-request，亚马逊就拒绝爬取，这时要模拟浏览器类型

全代码

```python
import requests
url = "https://www.amazon.cn/gp/product/B01M8L5Z3Y"
try:
    kv = {'user-agent':'Mozilla/5.0'}
    r = requests.get(url,headers=kv)
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[1000:2000])
except:
    print("爬取失败")
```

模拟浏览器 mozillla/5.0 是浏览器标准的身份识别字段，可能是火狐或其他浏览器，然后网站也会根据请求头（request-header）的其它字段来检验，比如 cookies，referer 等等

搜索引擎关键词提交接口

百度的关键词接口：

http://www.baidu.com/s?wd=keyword

360的关键词接口：

http://www.so.com/s?q=keyword

```
>>> import requests
>>> kv = {'wd':'Python'}
>>> r = requests.get("http://www.baidu.com/s", params=kv)
>>> r.status_code
200
>>> r.request.url
'http://www.baidu.com/s?wd=Python'
>>> len(r.text)
302829
```

：是键值对的意思，=是赋值的意思，python 是要搜索的关键字

百度搜索全代码

```
import requests
keyword = "Python"
try:
    kv = {'wd':keyword}
    r = requests.get("http://www.baidu.com/s",params=kv)
    print(r.request.url)
    r.raise_for_status()
    print(len(r.text))
except:
    print("爬取失败")
```

```
>>> import requests
>>> kv = {'q':'Python'}
>>> r = requests.get('http://www.so.com/s',params=kv)
>>> r.status_code
200
>>> r.request.url
'https://www.so.com/s?q=Python'
>>> len(r.text)
228253
```

360搜索全代码

```
import requests
keyword = "Python"
try:
    kv = {'q':keyword}
    r = requests.get("http://www.so.com/s",params=kv)
    print(r.request.url)
    r.raise_for_status()
    print(len(r.text))
except:
    print("爬取失败")
```

网络图片的爬取

网络图片链接的格式：

http://www.example.com/picture.jpg

国家地理：http://www.nationalgeographic.com.cn/

选择一个图片Web页面：

http://www.nationalgeographic.com.cn/photography/photo_of_the_day/3921.html

图片爬取全代码

```
import requests
import os
url = "http://image.nationalgeographic.com.cn/2017/0211/20170211061910157.jpg"
root = "D://pics//"
path = root + url.split('/')[-1]
try:
    if not os.path.exists(root):
        os.mkdir(root)
    if not os.path.exists(path):
        r = requests.get(url)
        with open(path, 'wb') as f:
            f.write(r.content)
            f.close()
            print("文件保存成功")
    else:
        print("文件已存在")
except:
    print("爬取失败")
```

第五行代码是根目录加上以‘/’为分隔符的最后一个地址即 20170211.....jpg，先判断当前目录是否存在，如果不存在则创建这个目录 root（os.mkdir() 方法用于以数字权限模式创建目录。）将文件打开并定义为文件标识符 f，然后 r.content 表示返回内容的二进制形式，将返回的二进制形式写入文件中

os --- 操作系统接口模块

wb	**以二进制格式打开一个文件只用于写入。如果该文件已存在则打开文件，并从开头开始编辑，即原有内容会被删除。如果该文件不存在，创建新文件。一般用于非文本文件如图片等。**

http://m.ip138.com/ip.asp?ip=ipaddress

IP地址查询全代码

```
import requests
url = "http://m.ip138.com/ip.asp?ip="
try:
    r = requests.get(url+'202.204.80.112')
    r.raise_for_status()
    r.encoding = r.apparent_encoding
    print(r.text[-500:])
except:
    print("爬取失败")
```

取最后 500 行