简单网络爬虫

> 上次说到http协议 然后我们现在用`python`标准库`urllib2`来实现简单的网络爬虫

urllib2定义了以下方法：

urllib2.urlopen(URL, Data, timeout)

> url参数：网页URL，可接受request对象。

> Data参数：POST数据提交（例如：账号密码发送给服务器判断登陆）

> 返回一个类似于`open`文件对象 从中读取网页数据

urllib2.Request(URL, Data=None, headers={})

- 注意R大写
- Data为None时，发送的是GET请求，反之POST

urllib2小案例

GET请求python官网获取下载链接

```
import urllib
import urllib2
import re

if __name__ == '__main__':
  url = "https://www.python.org/downloads/"
  # GET请求官网下载地址    返回 对象.read()取出网页数据
  res = urllib.urlopen(url).read()
  # 编译为Pattern模式 匹配 取出列表中第一个数据
  r = re.compile(r"Download the latest version for Windows[\s\S]+?</a>[\
  s\S]+?</a>").findall(res)[0]
  # 进行数据清洗
  li = re.compile(r'a class="button" href="(.*?)">(.*?)<').findall(r)
  # 进行输出
  py3x = li[0]
  py2x = li[1]
  print py3x[1]+": "+py3x[0]+"\n"+py2x[1]+": "+py2x[0]
```

> 获得以下数据：

> Download Python 3.6.0：https://www.python.org/ftp/python/3.6.0/python-3.6.0.exe

> Download Python 2.7.13：https://www.python.org/ftp/python/2.7.13/python-2.7.13.msi

下次我们来玩点有意思的东西

还有一些很好的技术文章尽情戳知了课堂官方QQ：2156600937