Urllib2 的使用

1) 最简单的爬虫
网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。python 的 urllib\urllib2 等模块很容易实现这一功能，下面的例子实现的是对 baidu 首页的下载。具体代码如下：

```
import urllib2
page=urllib2.urlopen("http://www.baidu.com")
print page.read().decode('utf-8')
```

2) 提交表单数据
1. 用 GET 方法提交数据
提交表单的 GET 方法是把表单数据编码至 URL。在给出请示的页面后，加上问号，接着是表单的元素。 如在百度中搜索“马伊琍”得到 url 为： http://www.baidu.com/s?wd=%E9%A9%AC%E4%BC%8A%E7%90%8D&pn=100&rn=20&ie=utf-8&usm=4&rsv_page=1 其中? 后面为表单元素。wd=%E9%A9%AC%E4%BC%8A%E7%90%8D 表示搜索的词是“马伊琍”，pn 表示从第 100 条信息所在页开始显示（感觉是这样，我试了几次，当写 100 时，从其所在页显示，但如果写 10，就是从第 1 页 显示），rn=20 表示每页显示 20 条，ie=utf-8 表示编码格式，usm=4 没明白是什么意思，换了 1、2、3 试了下， 没发现什么变化，rsv_page=1 表示第几页。如果要下载以上页面比较简单的方法是直接用上面的网址进行提取。 如代码：

```
#coding:utf-8
import urllib2
keyword=urllib2.quote('马伊琍')
page=urllib2.urlopen("http://www.baidu.com/s?wd="+keyword+"&pn=100&rn=20&ie=utf-8&usm=4 &rsv_page=1")
print page.read()
```