上海图书馆爬虫教程——————包简介

1. Urllib和urllib2

1. 引入包方法：import urllib import urllib2

2. 主要用于传入网页url，并通过html = urllib.urlopen(url)来返回一个html对象。Urllib2与urllib用法类似，但多了很多功能，如在使用urllib2.urlopen时可加入timeout参数使得当url在设定时间内打不开时自动跳过，以避免传入错误的url浪费时间。例如可配合try、except，

try:

urllib2.urlopen(whole\_img,timeout=1)

urllib.urlretrieve(whole\_img,path)

except :

print ’error’

1. beautifulsoup
2. 引入包方法：from bs4 import BeautifulSoup
3. 主要作用：Beautiful Soup将复杂HTML文档转换成一个复杂的树形结构，以获取html中的具体tag内容及string内容。如context = sub\_sp.find\_all('div',attrs={'class':'textpage\_itm'}) 作用为提取所有<div class="textpage\_itm">中的内容，具体用法见<https://www.crummy.com/software/BeautifulSoup/bs4/doc/index.zh.html#tag>
4. Re正则表达式
5. 引入包方法: import re
6. 主要作用：处理beautifulsoup中处理不了的内容，例如去除内容中的符号。html\_label = re.compile('<.\*?>') all\_context = html\_label.sub("",html\_context)
7. 详细正则表达式语法见http://www.runoob.com/regexp/regexp-syntax.html
8. pandas
9. 引入包方法:import pandas as pd
10. 主要作用:将爬虫结果转化成DataFrame，之后再保存成csv格式，如df1 = pd.DataFrame(all\_context,columns=["标题","网址","时间","发布单位","作者","标签","文章内容城市","摘要","正文","正文html"])
11. selenium
12. 引入包方法：from selenium import webdriver from selenium.webdriver.common.keys import Keys
13. 主要作用：对于网页中的js脚本使用selenium自动化工具来模拟人的操作，特别是内容的翻页通常都是使用js脚本来写。driver = webdriver.Chrome(chrome\_options=chrome\_options) driver.get(visitedUrl)
14. 具体进阶用法可以参考http://selenium-python.readthedocs.io/