Python爬虫实战一之使用Beautiful Soup抓...

开发需求：

1、爬取百度招聘下的招聘信息：岗位名称、工作地点、公司名称、薪资、发布时间

2、超链接招聘具体信息：职位描述

3、可以根据岗位名称包含主要字段进行分类存储

4、可以根据发布时间进行分类存储

5、可以根据新增水平进行分类存储

.....

开发思路

1、找到翻页的url规律

第一页：http://zhaopin.baidu.com/quanzhi?query=测试&sort_type=1&city=济南&detailmode=close&rn=20&pn=0

第二页：http://zhaopin.baidu.com/quanzhi?query=测试&sort_type=1&city=济南&detailmode=close&rn=20&pn=20

第三页：http://zhaopin.baidu.com/quanzhi?query=测试&sort_type=1&city=济南&detailmode=close&rn=20&pn=40

.....

2、找到超链接的url规律

http://zhaopin.baidu.com'+href

其中从招聘信息中获取href标签半个路径

3、如何获取需要的岗位名称、工作地点、公司名称、薪资、发布时间、职位描述等信息，并封装为一个字典对象中

4、生成excel文件，并把字典数据存储进去

```
1 import xlwt
```

```
import time,os
class StatisticsReport(object):
    t=time.strftime('%Y%m%d%H%M%S',time.localtime(time.time()))
    #设置单元格样式
    def set_style(self,name,height,bold=False):
        # 初始化样式
        style = xlwt.XFStyle()
        # 为样式创建字体
        font = xlwt.Font()
        font.name = name
        font.bold = bold
        font.color_index = 4
        font.height = height
        style.font = font
        return style
    def __createStatisticsReport__(self):
        RunNo=self.t
        reportname=RunNo+'.xls'
        self.__setreportname__(reportname)
        ReportFile=xlwt.Workbook()
        #创建1个测试报告sheet页名称
        ReportFile.add_sheet(u'测试类岗位',cell_overwrite_ok=True)
        #-------------写入按测试类岗位的信息表头
        #岗位名称     地区     公司名称     薪资     发布时间  岗位信息
        wr_tree = ReportFile.get_sheet(0)
        row0=[u'岗位名称',u'地区',u'公司名称',u'薪资',u'发布时间',u'岗位信息']
        #生成按测试类岗位的信息表头
        for i in range(0,len(row0)):
            wr_tree.write(0,i,row0[i],self.set_style('Times New
Roman',220,True))
        reportpath=os.path.abspath("..")+'\\'
        print reportpath+reportname
        ReportFile.save(reportpath+reportname)
    def __setreportname__(self,reportname):
        self.reportname=reportname
    def __getreportname__(self):
        return self.reportname
```

2、获取招聘信息并存储excel文件

```
import urllib2
from bs4 import BeautifulSoup
import xlrd,os
from xlutils.copy import copy
from StatisticsReport import StatisticsReport
def GetJobinfo():
    dict1={}
    n=1
    head = {}   #设置头
```

```
    head['User-Agent'] = 'Mozilla/5.0 (Windows NT 10.0; WOW64; rv:52.0) Gecko/20100101 Firefox/52.0'
    urlhead='http://zhaopin.baidu.com/quanzhi?query=测试&sort_type=1&city=济南&detailmode=close&rn=20&pn=';
    for i in range(0,n):
        #获取url全路径
        get_url=urlhead+str(2*i*10);
        #模拟浏览器，定制http请求头
        request=urllib2.Request(url=get_url,headers = head)
        #模拟浏览器，调用urlopen并传入Request对象，将返回一个相关请求response对象
        reponse=urllib2.urlopen(request)
        #这个应答对象如同一个文件对象，可以在Response中调用.read()读取这个文件信息
        zhaop_html=reponse.read().decode('utf-8')
        # UTF-8模式读取获取的页面信息标签和内容
        zhaop_htmltables=BeautifulSoup(zhaop_html,'lxml');
        #获取所有'a'标签以及内容
        get_linka_list=zhaop_htmltables.find_all('a')
        j=0
        for alink in get_linka_list:
            href=alink.get('href')
            if href.find('/szzw?detailidx=')==-1:
                pass
            else:
                get_joburl='http://zhaopin.baidu.com'+href;
                jobname=alink.span.get_text()
                jobarea=alink.find_all('p')[0].get_text().split('|')[0]
                jobcompany=alink.find_all('p')[0].get_text().split('|')[1]
                jobpay=alink.find_all('p')[1].get_text()
                jobdate=alink.find_all('p')[2].get_text()
                get_jobrequest=urllib2.Request(url=get_joburl,headers = head)
                #模拟浏览器，调用urlopen并传入Request对象，将返回一个相关请求response对象
                get_jobreponse=urllib2.urlopen(get_jobrequest)
                #这个应答对象如同一个文件对象，可以在Response中调用.read()读取这个文件信息
                get_job_html=get_jobreponse.read().decode('utf-8')
                # UTF-8模式读取获取的页面信息标签和内容
                zhaop_htmltables=BeautifulSoup(get_job_html,'lxml');
                job_tag=zhaop_htmltables.find_all(name='div',attrs={'class':'abs'});
                job_decs=job_tag[0].find_all('p')
                job_decinfo=''
                for job_dec in job_decs:
                    job_decinfo+=job_dec.get_text()+'\n'
                list1=[jobname,jobarea,jobcompany,jobpay,jobdate,job_decinfo]
```

```
                j+=1
                key='test'+str(j)
                dict2=dict.fromkeys([key], list1)
                dict1.update(dict2)
    return dict1
def GenerateReport(report,job_dict):
    #os.path.abspath("..") 用于返回上一级绝对路径
    reportpath=os.path.abspath("..")+'\\'
    reportname=report.__getreportname__()
    bk=xlrd.open_workbook(reportpath+reportname)
    wb=copy(bk)
    wa=wb.get_sheet(0)
    for i in range(0,len(job_dict)):
        for j in range(0,len(job_dict.values()[i])):
            wa.write(i+1,j,job_dict.values()[i][j])
    wb.save(reportpath+reportname)
if __name__ == '__main__':
   # 方法调用
    report=StatisticsReport()
    report.__createStatisticsReport__()
    job_dict=GetJobinfo()
    GenerateReport(report,job_dict)
```

结果：

	A	B	C	D	E	F
1	岗位名称	地区	公司名称	薪资	发布时间	岗位信息
2	济南良成月	济南市-历城	济南良成	面议	昨天发布	职位类型：其它 发布时间：2017-07-24 有效日期：2018-07-24 岗位职责: 无
3	LTE网优/测	济南-历城区	山东闻远	4000-8000	今天发布	职位类型：其他技术职位发布时间：2017-07-25有效日期：2017-0
4	软件测试(转	济南-不限	软通动力	5000-8000	2017-07-23	职位类型：IT/互联网发布时间：2017-07-23有效日期：2017-09-02
5	自动化测试	济南	银客未来	12000-180	昨天发布	职位类型：技术发布时间：2017-07-24有效日期：2017-09-241.负
6	软件测试(应	济南	淄博智博	4000-8000	今天发布	职位类型：软件测试发布时间：2017-07-25有效日期：2017-08-24
7	LTE网优LT	济南-不限	山东闻远	3000-5000	昨天发布	职位类型：电子/通信发布时间：2017-07-24有效日期：2017-09-03
8	物流软件测	济南-历下区	山东海文	5000-8000	昨天发布	职位类型：采购/贸易/物流/交通运输发布时间：2017-07-24有效日
9	资深测试工	济南-历下区	车满满（	10000-180	今天发布	职位类型：自动化测试发布时间：2017-07-25有效日期：2017-08-
10	测试工程师	济南-历城区	山东思臣	4000-8000	今天发布	职位类型：测试经理发布时间：2017-07-25有效日期：2017-08-24
11	APP测试专	济南-历下区	北京东方	3000-6000	今天发布	职位类型：移动端测试发布时间：2017-07-25有效日期：2017-08-
12	软件测试（	济南-历下区	北京锐融	5000-1000	今天发布	职位类型：软件测试发布时间：2017-07-25有效日期：2017-08-24
13	8279_软件	济南	深圳市金	7000-1200	今天发布	职位类型：技术发布时间：2017-07-25有效日期：2017-09-25岗位
14	5千起聘测	济南-历城区	北京才秀	5000-8000	昨天发布	职位类型：计算机/互联网/通信发布时间：2017-07-24有效日期：2
15	测试工程师	济南	深圳市金	8000-1200	今天发布	职位类型：技术发布时间：2017-07-25有效日期：2017-09-251、计
16	功能测试(急	济南-历下区	山东秦商	3000-4000	今天发布	职位类型：功能测试发布时间：2017-07-25有效日期：2017-08-24
17	软件测试工	济南-不限	中软国际	5000-8000	昨天发布	职位类型：电子/通信发布时间：2017-07-24有效日期：2017-09-12
18	测试工程师	济南	北京辰森	6000-1200	今天发布	职位类型：技术发布时间：2017-07-25有效日期：2017-09-251、制
19	功能测试(急	济南-历下区	山东秦商	3000-4000	今天发布	职位类型：功能测试发布时间：2017-07-25有效日期：2017-08-24
20	测试主管(t	济南-不限	山东世纪	5000-8000	2017-07-23	职位类型：IT/互联网发布时间：2017-07-23有效日期：2017-09-17
21	5K聘机械乡	济南-历下区	山东海文	3000-5000	昨天发布	职位类型：生产制造/造纸印刷/服装服饰发布时间：2017-07-24有效
22						
23						
24						
25						
26						
27						
28						
29						
30						