window系统:

# 当系统同时安装Python2 和Python3 时, 使用pip3 install scrapy安装scrapy时，

# 报错：[Pip - Fatal error in launcher: Unable to create process using '"'](https://stackoverflow.com/questions/37220055/pip-fatal-error-in-launcher-unable-to-create-process-using)

# 解决方案：python3 –m pip install scrapy

# windows中自定 scrapy crawl jobbole 报错：No module named win32api

# 解决方案： pip install pypiwin32, linux 不会出现此问题

# 当节点的class属性包含多个值时，如下：

# <span data-post-id="112239" class=" btn-bluet-bigger href-style vote-post-up register-user-only "><i class="fa fa-thumbs-o-up"></i> <h10 id="112239votetotal">2</h10> 赞</span>

# xpath写法如下：

# xpath('//span[contains(@class," vote-post-up")]')

## 4-1 scrapy安装以及目录结构介绍

安装scrapy可以看我另外一篇博文：[Scrapy的安装--------Windows、linux、mac等操作平台](http://www.cnblogs.com/jinxiao-pu/p/6679802.html)，现在是在虚拟环境中安装可能有不同。

### 1.创建有python3的虚拟环境

mkvirtualenv --python=C:\Users\admin\AppData\Local\Programs\Python\Python35\python3.exe py3scrapy

### 2.安装scrapy

进入环境py3scrapy，**pip install -i**[**https://pypi.douban.com/simple/**](https://pypi.douban.com/simple/)**scrapy** 豆瓣源安装非常快。

### 3.补充

进入虚拟环境： workon py3scrapy

创建项目： scrapy startproject ArticleSpider

建立spider： scrapy genspider jobbole blog.jobbple.com

[回到顶部](http://www.cnblogs.com/jinxiao-pu/p/6713333.html#_labelTop)

## 4-2 pycharm 调试scrapy 执行流程

### 1.运行爬虫文件

建立一个main.py文件，在ArticleSpider文件目录下

from scrapy.cmdline import execute

import sys,os

sys.path.append(os.path.dirname(os.path.abspath(\_\_file\_\_)))

execute(['scrapy','crawl','jobbole'])

os.path.abspath(\_\_file\_\_) --------main.py目录

dirname() --------main.py父目录

### 2.要学会用断点和DEBUG

在实战中操作

[回到顶部](http://www.cnblogs.com/jinxiao-pu/p/6713333.html#_labelTop)

## 4-3~5 xpath的用法

### 1.xpath简介

* xpath使用路径表达式在xml和html中进行导航。
* xpath包含标准函数库。
* xpath是一个w3c的标准。

### 2.xpath节点关系

* 父节点
* 子节点
* 同胞节点
* 先辈节点
* 后代节点

### 3.xpath语法

  
  


### 4.补充

**为什么有时候自己写的xpath明明对的，却获取不到数据？**

原因：F12产生的源码，不同于网页源代码，前者可能是js加载完的源代码。response.xpath()是根据网页源代码来提取信息的。

.

**问题：No modle named ‘win32api’**

解决： pip install -i <https://pypi.douban.com/simple/> pypiwin32

.

**contains()用法**

response.xpath("//span[contains(@class, 'bookmark-btn')]/text()").extract()[0]

表示在span标签中class属性中含有 bookmark-btn 即为符合

.

**正文保留html标签，以便后续研究**

.

scrapy shell testurl 调试xpath

如果在py3中就都显示中文了

http://images2015.cnblogs.com/blog/1129740/201704/1129740-20170415104230689-1278208795.png

.

re.match(reg,html).group() #正则匹配

.

tag\_list=['职场','2 评论','今昔']

[element for element in tag\_list if not element.strip().endswith('评论')]

#结果['职场', '今昔']

[回到顶部](http://www.cnblogs.com/jinxiao-pu/p/6713333.html#_labelTop)

## 4-6~7 css选择器实现字段解析

  
  
