**Scrapy的安装介绍**

Scrapy框架官方网址：[http://doc.scrapy.org/en/latest](http://doc.scrapy.org/en/latest/)

**Windows 安装方式**

* Python 2 / 3
* pio install Twisted-17.9.0-cp35-cp35m-win\_amd64.whl
* 升级pip版本：pip install --upgrade pip
* 通过pip 安装 Scrapy 框架pip install scrapy
* 通过pip 安装 pypiwin32

**Linux安装**

* Python 2 / 3
* 安装非Python的依赖 sudo apt-get install python-dev python-pip libxml2-dev libxslt1-dev zlib1g-dev libffi-dev libssl-dev
* 通过pip 安装 Scrapy 框架 sudo pip install scrapy

**Scrapy 框架**

* Scrapy是用纯Python实现一个为了爬取网站数据、提取结构性数据而编写的应用框架，用途非常广泛。
* 框架的力量，用户只需要定制开发几个模块就可以轻松的实现一个爬虫，用来抓取网页内容以及各种图片，非常之方便。
* Scrapy 使用了 Twisted['twɪstɪd](其主要对手是Tornado)异步网络框架来处理网络通讯，可以加快我们的下载速度，不用自己去实现异步框架，并且包含了各种中间件接口，可以灵活的完成各种需求。

**Scrapy架构****图(绿线是数据流向)：**

* Scrapy Engine(引擎): 负责Spider、ItemPipeline、Downloader、Scheduler中间的通讯，信号、数据传递等。
* Scheduler(调度器): 它负责接受引擎发送过来的Request请求，并按照一定的方式进行整理排列，入队（Queue），当引擎需要时，交还给引擎。
* Downloader（下载器）：负责下载Scrapy Engine(引擎)发送的所有Requests请求，并将其获取到的Responses交还给Scrapy Engine(引擎)，由引擎交给Spider来处理，
* Spider（爬虫）：它负责处理所有Responses,从中分析提取数据，获取Item字段需要的数据，并将需要跟进的URL提交给引擎，再次进入Scheduler(调度器)，
* Item Pipeline(管道)：它负责处理Spider中获取到的Item，并进行进行后期处理（详细分析、过滤、存储等）的地方.
* Downloader Middlewares（下载中间件）：你可以当作是一个可以自定义扩展下载功能的组件。
* Spider Middlewares（Spider中间件）：你可以理解为是一个可以自定扩展和操作引擎和Spider中间通信的功能组件（比如进入Spider的Responses;和从Spider出去的Requests）

**Scrapy的运作流程**

代码写好，程序开始运行...

* 引擎：Hi！Spider, 你要处理哪一个网站？
* Spider：老大要我处理xxxx.com。
* 引擎：你把第一个需要处理的URL给我吧。
* Spider：给你，第一个URL是xxxxxxx.com。
* 引擎：Hi！调度器，我这有request请求你帮我排序入队一下。
* 调度器：好的，正在处理你等一下。
* 引擎：Hi！调度器，把你处理好的request请求给我。
* 调度器：给你，这是我处理好的request
* 引擎：Hi！下载器，你按照老大的下载中间件的设置帮我下载一下这个request请求
* 下载器：好的！给你，这是下载好的东西。（如果失败：sorry，这个request下载失败了。然后引擎告诉调度器，这个request下载失败了，你记录一下，我们待会儿再下载）
* 引擎：Hi！Spider，这是下载好的东西，并且已经按照老大的下载中间件处理过了，你自己处理一下（注意！这儿responses默认是交给def parse()这个函数处理的）
* Spider：（处理完毕数据之后对于需要跟进的URL），Hi！引擎，我这里有两个结果，这个是我需要跟进的URL，还有这个是我获取到的Item数据。
* 引擎：Hi ！管道 我这儿有个item你帮我处理一下！调度器！这是需要跟进URL你帮我处理下。然后从第四步开始循环，直到获取完老大需要全部信息。
* 管道``调度器：好的，现在就做！

**注意！只有当调度器中不存在任何request了，整个程序才会停止，（也就是说，对于下载失败的URL，Scrapy也会重新下载。）**

**制作 Scrapy 爬虫 一共需要4步：**

* 新建项目 (scrapy startproject xxx)：新建一个新的爬虫项目
* 明确目标 （编写items.py）：明确你想要抓取的目标
* 制作爬虫 （spiders/xxspider.py）：制作爬虫开始爬取网页
* 存储内容 （pipelines.py）：设计管道存储爬取内容