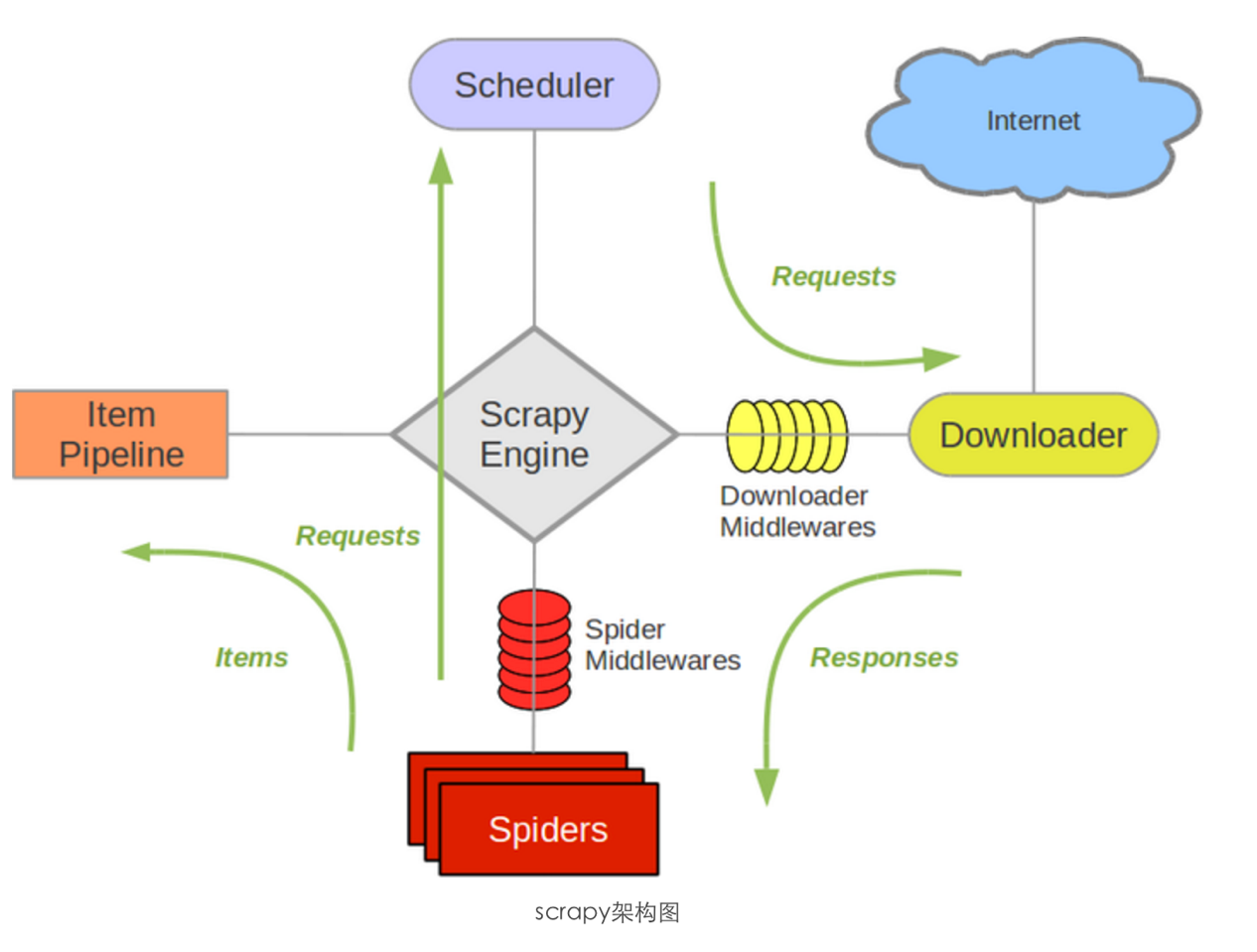
***Scrapy Processing Architecture***



1. Terminal:

$ ***cd /Users/ceciliaLee/Desktop***

$ ***scrapy startproject newScrapy ##*** Create a new Scrapy project

1. The content structure of  */ Users/ceciliaLee/Desktop/newScrapy:*

|---- newScrapy

| |---- newScrapy

| |---- \_\_init\_\_.py

| |---- items.py #用来存储爬下来的数据结构（字典形式）

| |---- pipelines.py #用来对爬出来的item进行后续处理，如存入数据库等

| |---- settings.py #爬虫配置文件

| |---- spiders #此目录用来存放创建的新爬虫文件（爬虫主体）

| |---- \_\_init\_\_.py

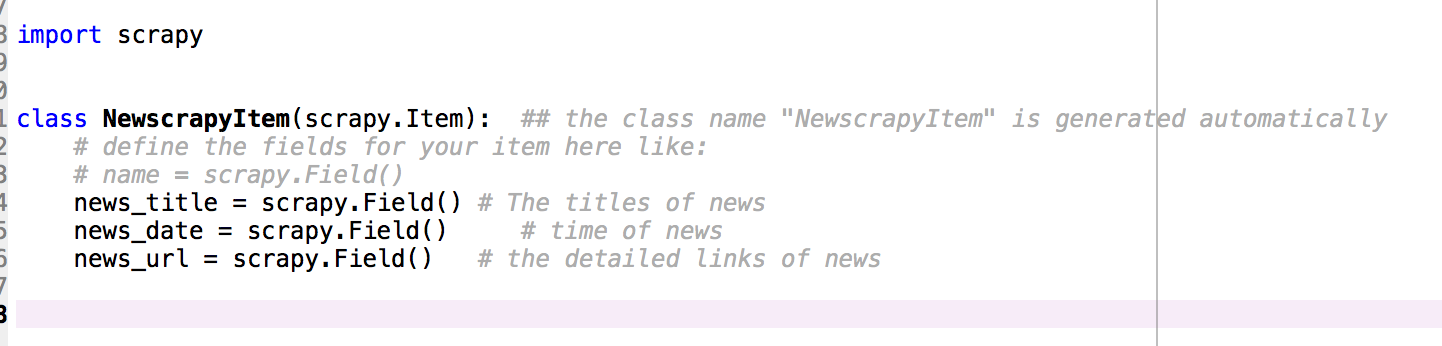
| |---- scrapy.cfg #项目配置文件

1. Setting ***items***

This tutorial is aiming at crawling news from <http://news.njupt.edu.cn/s/222/t/1100/p/1/c/6866/i/1/list.htm>

The information being stored are: the title, time, and detailed links of news in the website

Writing codes in ***/newScrapy/ items.py***



1. Editing ***spider***

***Spider*** is the main part of crawler, which is response for processing ***requset response*** and ***ur****l*, and deliver the results to***pipelines.py*** for subsequent processing.

Create a new python file, named ***newscrapySpider.py*** in ***/newScrapy/spiders***

Codes in ***/newScrapy/spiders*/*newscrapySpider.py****:*



**Notes:**

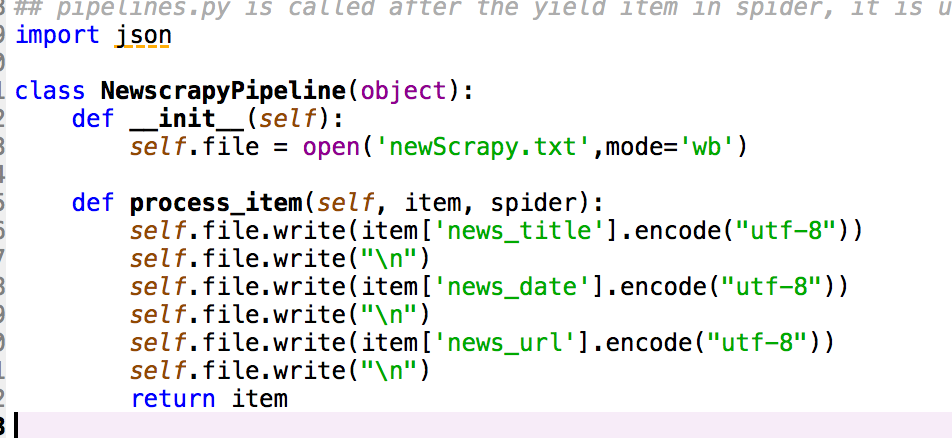
1. name为爬虫名称，在后面启动爬虫的命令当中会用到。
2. allowed\_domains为允许爬虫爬取的域名范围（如果连接到范围以外的就不爬取）
3. start\_urls表明爬虫首次启动之后访问的第一个Url，其结果会被自动返回给parse函数。4.
4. parse函数为scrapy框架中定义的置函数，用来处理请求start\_urls之后返回的response，由我们实现
5. news\_page\_num = 14和page\_num = 386别表示每页的新闻数目，和一共有多少页，本来也可以通过xpath爬取下来的，但是我实在是对我们学校的网站制作无语了，html各种混合，于是我就偷懒手动输入了。
6. 之后通过item = NewscrapyItem ()来使用我们之前定义的item，用来存储新闻的url、标题、日期。（这里面有一个小技巧就是通过|来接连xpath可以一次返回多个想要抓取的xpath）
7. 通过yield item来将存储下来的item交由后续的pipelines处理
8. 之后通过生成next\_page\_url来通过scrapy.Request抓取下一页的新闻信息
9. scrapy.Request的两个参数，一个是请求的URL另外一个是回调函数用于处理这个request的response，这里我们的回调函数是parse\_news
10. parse\_news里面的步骤和parse差不多，当然你也可以改造一下parse然后直接将其当做回调函数，这样的话一个函数就ok了
11. Editing ***pipelines.py*** under ***/newScrapy***

***pipelines.py*** is used for further processing of data, such as type conversion, restore in databases, write to local machine, and so on.

***pipelines.py*** is called after the yield item in spider, it is used to process each single item.

Don't forget to add your pipeline to the ITEM\_PIPELINES setting in ***/newScrapy/settings.py***

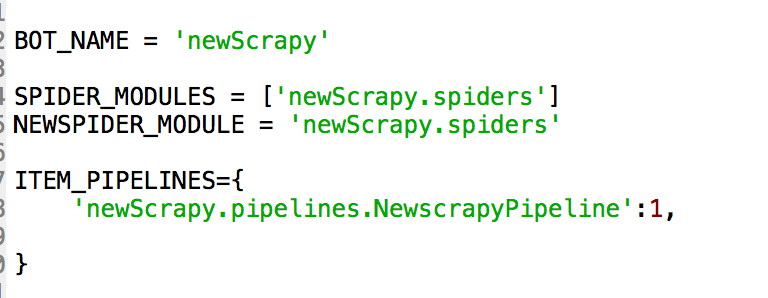
Codes in ***/newScrapy/pipelines.py:***



1. Editing ***/newScrapy/settings.py***

***settings.py*** is used to restore the configuration information of the crawler.

Codes in ***/newScrapy/settings.py (***the most simplest configuration info is adding ITEM\_PIPELINES as follows***):***



1. Executing the crawler

Programming in Terminal:

1. Change the direction to ***/newScrapy/spiders:***

cd /Users/ceciliaLee/Desktop/newScrapy/newScrapy/spiders

1. Execute the crawler:

scrapy crawl newScrapy

# newScrapy is the name of Folder where the crawler was created