Scrapy需求规格说明书

[V1.95]

|  |  |  |  |
| --- | --- | --- | --- |
| 编写 | 武丁泽宇 | 日期 | 2017年5月29 日 |
| 校对 | 郭炜锋 | 日期 | 2017年5月29日 |

北京航空航天大学计算机学院

二〇一七年三月二十九日

文档修改记录

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 版本号 | 日期 | 所修改章节 | 修改说明 | 修改人 | 审核人 |
| 1.00 | 2017.3.21 |  | 完成第一版 | 武丁泽宇 | 胡勇 |
| 1.10 | 2017.3.29 | 4 | 修改第4章内容 | 武丁泽宇 | 胡勇 |
| 1.11 | 2017.3.29 | 4 、5 | 修改4章内容,删除5章 | 武丁泽宇 | 胡勇 |
| 1.20 | 2017.4.4 | 1、2 、3、 4 | 按照修改建议进行修订 | 武丁泽宇 | 胡勇 |
| 1.30 | 2017.4.5 | 3 | 完善业务需求和用例图 | 武丁泽宇 | 胡勇 |
| 1.40 | 2017.4.9 | 全文 | 排版修改 | 胡勇 | 胡勇 |
| 1.50 | 2017.4.16 | 1.3、全文字体、全文格式 | 将1.3定义使用表格表述，将正文的格式统一 | 郭炜锋 | 胡勇 |
| 1.60 | 2017.4.23 | 全文修改 | 根据老师批注进行修改 | 武丁泽宇 | 胡勇 |
| 1.70 | 2017.4.26 | 全文修改 | 根据各组评审意见进行修改和调整 | 武丁泽宇 | 胡勇 |
| 1.75 | 2017.5.3 | 批注章节修改增加第5章 | 按照批注进行修改完善 | 武丁泽宇 | 胡勇 |
| 1.80 | 2017.5.4 | 全文 | 格式化文本格式 | 胡勇 | 胡勇 |
| 1.90 | 2017.5.11 | 全文 | 修改RUCM和用例图 | 武丁泽宇 | 胡勇 |
| 1.95 | 2017.5.29 | 第5章 | 增加改进需求 | 武丁泽宇 | 胡勇 |

目录

[1 引言 5](#_Toc1187029337)

[1.1 编写目的 5](#_Toc530941205)

[1.2 背景 5](#_Toc1141423573)

[1.3 定义 5](#_Toc1360639550)

[1.4 参考资料 7](#_Toc854469777)

[2 任务概述 7](#_Toc1950995715)

[2.1 目标 7](#_Toc1646537294)

[2.2 用户特点 7](#_Toc50781682)

[2.3 假定与约束 7](#_Toc371815511)

[3 需求与设计 7](#_Toc1231850004)

[3.1 需求 8](#_Toc210633330)

[3.1.1业务需求 8](#_Toc600795973)

[3.1.2功能需求 9](#_Toc236921976)

[3.1.3 非功能性需求 9](#_Toc1312335360)

[3.2 框架及组件概述 10](#_Toc701651546)

[3.3 用例图 12](#_Toc125941071)

[3.4 功能模块补充描述 17](#_Toc392979288)

[3.4.1 Scrapy Engine 18](#_Toc1275667253)

[3.4.2 Spiders 19](#_Toc1221825123)

[3.4.3设置（Settings） 19](#_Toc2108413675)

[3.4.4 Item Pipeline 20](#_Toc641604642)

[3.4.5 下载器中间件 20](#_Toc1845618088)

[3.4.6 Spider中间件 20](#_Toc106370052)

[3.5 扩展功能模块 20](#_Toc1319423790)

[3.5.1 日志（Logging） 20](#_Toc264466514)

[3.5.2 页面服务（Web Service） 21](#_Toc590611445)

[3.5.3 Item Exporters 21](#_Toc1407783469)

[3.5.4 自动限速(AutoThrottle)扩展 21](#_Toc342883506)

[3.5.5 自定义扩展 21](#_Toc2085795896)

[3.6 故障处理要求 22](#_Toc889854407)

[3.7 其他专门要求 23](#_Toc1616454850)

[4 运行环境规定 23](#_Toc1125341585)

[4.1 设备 23](#_Toc1420795613)

[4.2 支持软件 23](#_Toc610394775)

[4.3 接口 23](#_Toc338497487)

[4.3.1 硬件接口 23](#_Toc127781742)

[4.3.2 软件接口 23](#_Toc413906842)

[4.3.3 通信接口 24](#_Toc1985034782)

[4.3.4 用户接口 24](#_Toc178563424)

[5拓展和改进 24](#_Toc785722353)

[5.1 方案思路 24](#_Toc1069401138)

[5.2 详细设计 25](#_Toc389196755)

[5.2.1搭建测试环境 25](#_Toc1386518326)

[5.2.2反爬改进 25](#_Toc1306323114)

[5.2.3 下载器改进 26](#_Toc1701532115)

[5.2.4 分布式爬取 26](#_Toc2088169872)

[5.2.5 分布式爬取两级过滤器 26](#_Toc1432264185)

[5.2.6 爬取前端界面显示 27](#_Toc2094511404)

[5.2.7 爬取链接设置 27](#_Toc1216353477)

1 引言

本软件Scrapy是开源爬虫框架。本文详细描述了Scrapy的功能、非功能性、业务、拓展和改进需求。

1.1 编写目的

本软件需求规格说明书，是为软件设计、软件测试人员和用户编写的。

本软件需求规格说明书的适用读者，包括参加能力验证的开发测试人员、Scrapy技术人员，以及项目的其他相关人员。

1.2 背景

软件名称：Scrapy

项目的组织机构：Scrapy开源项目开发组

项目的实施机构：github站点上239位贡献者

项目背景：本项目是用于开发一个高速并发的网络爬虫的框架，用于爬取网站的数据信息并导出其数据结构。

1.3 定义

|  |  |
| --- | --- |
| 术语 | 说明 |
| 爬虫 | 具有抓取网页内容功能的软件 |
| 爬虫引擎(Scrapy Engine) | 引擎负责控制数据流在系统中所有组件中流动，并在相应动作发生时触发事件。 |
| 调度器(Scheduler) | 调度器从引擎接受request并将他们入队，以便之后引擎请求他们时提供给引擎。 |
| 下载器(Downloader) | 下载器负责获取页面数据并提供给引擎，而后提供给spider。 |
| Spiders | Spider是Scrapy用户编写用于分析response并提取item(即获取到的item)或额外跟进的URL的类。 |
| Item Pipeline | Item Pipeline负责处理被spider提取出来的item。 |
| 下载器中间件(Downloader middlewares) | 下载器中间件是在引擎及下载器之间的特定钩子(specific hook)，处理Downloader传递给引擎的response。 |
| Spider中间件(Spider middlewares) | Spider中间件是在引擎及Spider之间的特定钩子(specific hook)，处理spider的输入(response)和输出(items及requests)。 |
| 数据流(Data flow) | Scrapy中的数据流。 |
| 事件驱动网络(Event-driven networking) | Scrapy基于事件驱动网络框架 Twisted 编写。 |
| XML | 可扩展[标记语言](http://baike.baidu.com/item/%E6%A0%87%E8%AE%B0%E8%AF%AD%E8%A8%80" \t "/home/wudy/文档\\x/_blank)，[标准通用标记语言](http://baike.baidu.com/item/标准通用标记语言" \t "/home/wudy/文档\x/_blank)的子集，是一种用于标记电子文件使其具有结构性的[标记语言](http://baike.baidu.com/item/标记语言" \t "/home/wudy/文档\x/_blank)。 |
| JSON | 存储和交换文本信息的语法。类似XML |
| CSV | 逗号分隔值，文件以纯文本形式存储表格数据（数字和文本）。纯文本意味着该文件是一个[字符](http://baike.baidu.com/item/%E5%AD%97%E7%AC%A6" \t "/home/wudy/文档\\x/_blank)序列，不含必须像二进制数字那样被解读的数据。 |
| XPath | 即XML路径语言（XML Path Language），它是一种用来确定XML文档中某部分位置的语言。XPath基于XML的树状结构，提供在数据结构树中找寻节点的能力。 |
| CSS | 层叠样式表，又称串样式列表、级联样式表、串接样式表、层叠样式表、階層式樣式表，一种用来为结构化文档（如HTML 文档或XML 应用）添加样式（字体、间距和颜色等）的计算机语言，由W3C 定义和维护。 |
| Telnet | Telnet协议是一种应用层协议，使用于互联网及局域网中，使用虚拟终端机的形式，提供双向、以文字字符串为主的交互功能。 |
| cookies | 指某些网站为了辨别用户身份而储存在用户本地终端上的数据（通常经过加密）。 |
| session | 会话（session）是一种持久网络协议，在用户（或用户代理）端和服务器端之间建立关联，从而起到交换数据包的作用机制。 |
| DNS | 互联网的一项服务。 它作为将域名和IP地址相互映射的一个分布式数据库，能够使人更方便地访问互联网。 |

1.4 参考资料

GJB 438B 国军标开发通用文档。

2 任务概述

2.1 目标

Scrapy的出现是为了构建一整套便捷高效的爬取框架，便于进行站点数据爬取，提取结构性数据。可以应用在包括数据挖掘，信息处理或存储历史数据等一系列的程序中。任何人都可以根据需求方便地修改框架内容。并且框架提供了多种类型爬虫的基类，同时使用Twisted这个异步网络库来处理网络通讯。构建的架构清晰，包含了各种中间件接口，可以灵活的完成各种需求。使用Python语言，以便达到简介高效的目的。

2.2 用户特点

熟悉网页抓取和Python的程序开发用户。

2.3 假定与约束

无

3 需求与设计

3.1 需求

3.1.1业务需求

典型业务需求场景：页面数据抓取

对爬虫最典型业务场景——页面数据抓取来说，在需要大量下载网络数据以便后续分析的时候，往往通过手动下载是不能满足需求的，就需要有一个自动化的下载方法。一般的高级语言都会提供相关网络数据下载的功能函数，通过调用就可完成简单的下载。但是在被抓取对象的结构比较复杂时，就需要开发人员自己根据需要编写相应的抓取程序，而抓取程序很大程度上都有重复公用的代码，每次都重新开发编写显然非常不合适。因此，就需要有一个抽象的开发爬虫的框架方便开发人员进行开发，避免了重复造轮子的工作。

Scrapy就是在这种需求背景下开发出来的。它将可重用的代码部分抽象出来成为模块调用，使得开发人员只需关注程序要实现的具体功能对象而不用再关心基础模块的搭建，为快速开发爬虫软件提供便捷、节约时间。另外，在爬虫代码编写中最耗时的是反爬虫的问题。使用scrapy框架在开始写代码之前可以利用内置的shell去请求你抓取网站的数据页面，检测网站的反爬虫能力，再用框架提供的模块编写相应程序，就会使得编写代码的复用率很高，更换爬取对象时一般只需改改正则表达式和爬取队列即可，从而大大的提高了工作的效率。

为了使得框架使用起来具有一定的灵活性，又具有一定的便捷性，因而将框架设计成多个模块，简单的爬取只需下载器去下载页面数据。对于多任务同时进行的要求就需要引入并行的spider去并发执行，而这些并发任务就需要一个调度器去合理的分配任务执行顺序。对于数据也要有相应的处理模块，每个模块的任务分配和整个程序的数据流控制就需要有一个核心的控制引擎去负责。

一个典型的案例就是亚马逊公司，亚马逊利用爬虫技术，在scrapy框架的基础上进行扩展，实现了其商品广告接口功能，通过灵活的API接口，为客户提供其商品的实时广告链接数据。对抓取的商品交易信息、价格信息、折扣信息等数据进行分析处理后自动化地实时更新商品广告，从而省去很多的人力成本。

3.1.2功能需求

当用户需要开发爬虫系统时，可以调用本框架，通过本框架提供的组件，搭建好一套高效的爬虫系统。用户花费更多的时间在逻辑实现上，而不必关心底层的实现细节。基于以上业务需求的分析与相应的业务要求,我们认为软件有如下的功能需求。

1. 爬虫引擎——负责控制数据流在系统中所有组件中流动，并在相应动作发生时触发事件
2. 调度功能——可调度的爬取请求队列
3. 下载功能——获取页面数据并提供给引擎，而后提供给 spider
4. 处理数据功能——从抓取到的网页中（HTML源码）提取数据
5. 日志功能——事件日志记录
6. 扩展机制——允许开发者自定义
7. 设置选项——用户可选运行模式
8. 异常处理——应急、错误处理

3.1.3 非功能性需求

Scrapy 需要支持以下非功能性需求以便更好了使用框架编写爬虫。

1. 高适用性。可以用scrapy crawl 来启动Scrapy，也可以使用 API 在脚本中启动Scrapy。默认情况下，执行 scrapy crawl 时，Scrapy每个进程运行一个spider。 Scrapy通过内部(internal)API 也支持单进程多个spider。除了支持单机爬取，也可以进行分布式爬取，支持启动多个Scrapyd，并分配到不同机器上。
2. 高可靠性。有些网站实现了特定的机制，以一定规则来避免被爬虫爬取。 框架需要能够实现避免被禁止(ban)，并且可以设置全局并发以便同时处理多个request。
3. 高便捷性。提供交互式shell终端 , 为测试CSS及XPath表达式，编写和调试爬虫提供了极大的方便。提供数据导出功能，提供对多格式(JSON、CSV、XML)、多存储后端(FTP、S3、本地文件系统)的支持。提供了一系列在spider之间共享的可复用的过滤器(即 Item Loaders)，对智能处理爬取数据提供了内置支持。
4. 高通用性。针对非英语语系中不标准或者错误的编码声明, 提供了自动检测以及健壮的编码支持。
5. 高扩展性。通过使用信号、设计好的API(中间件, extensions, pipelines)来定制实现功能。内置的中间件及扩展为下列功能提供支持: cookies and session 处理、HTTP 压缩 、HTTP 认证 、 HTTP 缓存 、user-agent模拟、爬取深度限制。

3.2 框架及组件概述



图 1 Scrapy框架图

Scrapy框架如图1所示（来源：http://www.jianshu.com/p/078ad2067419），各部分组件介绍如下。

1、Scrapy Engine

引擎负责控制数据流在系统中所有组件中流动，并在相应动作发生时触发事件。

2、调度器(Scheduler)

调度器从引擎接受request并将他们入队，以便之后引擎请求他们时提供给引擎。

3、下载器(Downloader)

下载器负责获取页面数据并提供给引擎，而后提供给spider。

4、Spiders

每个spider负责处理一个特定(或一些)网站。

5、Item Pipeline

Item Pipeline负责处理被spider提取出来的item。典型的处理有清理、 验证及持久化(例如存取到数据库中)。

6、下载器中间件(Downloader middlewares)

下载器中间件是在引擎及下载器之间的特定钩子(specific hook)，处理Downloader传递给引擎的response。 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能 。功能

7、Spider中间件(Spider middlewares)

Spider中间件是在引擎及Spider之间的特定钩子(specific hook)，处理spider的输入(response)和输出(items及requests)。 其提供了一个简便的机制，通过插入自定义代码来扩展Scrapy功能。

3.3 用例图

在实际环境中，开发者和使用者的身份一般重叠，但划分为两类，使得用例图更清晰。Scrapy框架的用例图如图2所示。

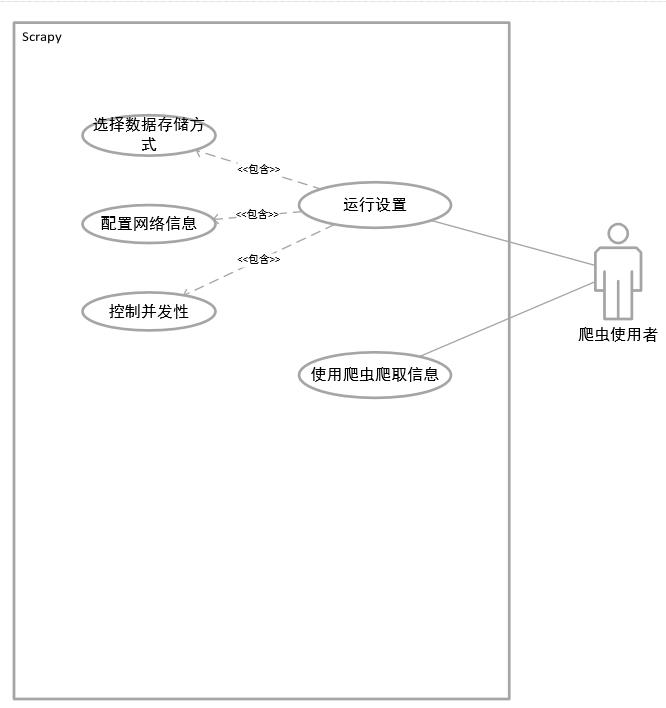
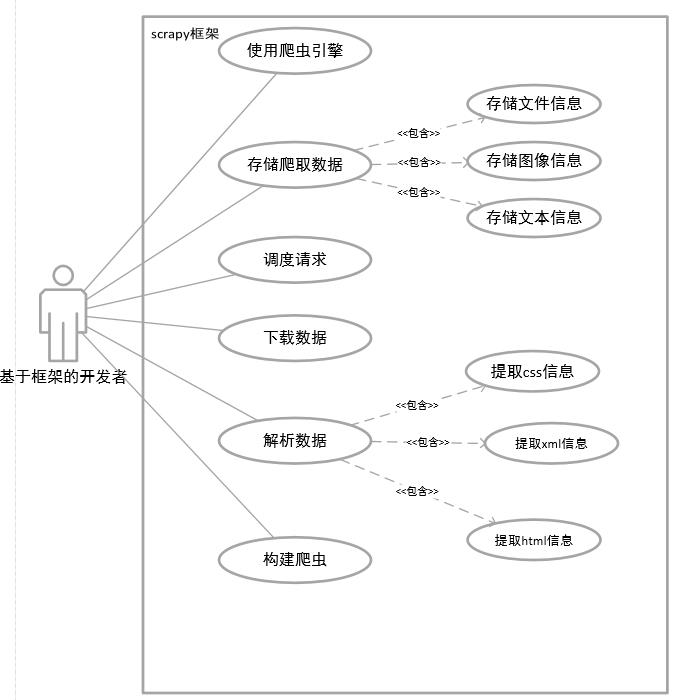


图2 Scrapy用例图

在主要的业务场景-页面数据抓取中，使用到的scrapy核心功能模块有：spider、设置、数据存储、下载器、调度器和scrapy engine。所以我们对这些模块画出RUCM图依次如下：Spider模块 RUCM图如图3所示，设置模块RUCM如图4所示，数据存储模块RUCM如图5所示，下载器模块RUCM如图6所示，调度器模块RUCM如图7所示, scrapy engine模块RUCM如图8所示。



图3 Spider模块 RUCM图



图4 设置模块RUCM图



图 5数据存储模块RUCM图



图6 下载器模块RUCM图



图7 调度器 模块RUCM图



图8 scrapy engine模块RUCM图

3.4 功能模块补充描述

对应于前面画出的各个Scrapy 核心功能模块RUCM图，我们给出scrapy engine和spider的更加详细补充描述，以表述RUCM未能展现的部分。由于在实际使用下载器、spider和数据存储时，需要经过它们的中间件传输数据，所以我们也将各个中间件的功能用文字描述出来。他们作为各个模块的一部分存在，因而没有使用RUCM图描述。

3.4.1 Scrapy Engine

Scrapy中的数据流由执行引擎Scrapy Engine控制，引擎控制的数据传输过程如下:

1. 引擎打开一个网站(open a domain)，找到处理该网站的Spider并向该spider请求第一个要爬取的URL(s)。
2. 引擎从Spider中获取到第一个要爬取的URL并在调度器(Scheduler)以Request调度。
3. 引擎向调度器请求下一个要爬取的URL。
4. 调度器返回下一个要爬取的URL给引擎，引擎将URL通过下载中间件(请求(request)方向)转发给下载器(Downloader)。
5. 一旦页面下载完毕，下载器生成一个该页面的Response，并将其通过下载中间件(返回(response)方向)发送给引擎。
6. 引擎从下载器中接收到Response并通过Spider中间件(输入方向)发送给Spider处理。
7. Spider处理Response并返回爬取到的Item及(跟进的)新的Request给引擎。
8. 引擎将(Spider返回的)爬取到的Item给Item Pipeline，将(Spider返回的)Request给调度器。
9. (从第二步)重复直到调度器中没有更多地request，引擎关闭该网站。

3.4.2 Spiders

Spider类定义了如何爬取某个(或某些)网站。包括了爬取的动作(例如:是否跟进链接)以及如何从网页的内容中提取结构化数据(爬取item)。 换句话说，Spider就是定义爬取的动作及分析某个网页(或者是有些网页)的模块。

对spider来说，爬取的循环过程如下:

1. 以初始的URL初始化Request，并设置回调函数。 当该request下载完毕并返回时，将生成response，并作为参数传给该回调函数。spider中初始的request是通过调用 start\_requests() 来获取。 start\_requests() 读取 start\_urls 中的URL， 并以 parse 为回调函数生成 Request 。
2. 在回调函数内分析返回的(网页)内容，返回 Item 对象、dict、 Request 或者一个包括三者的可迭代容器。 返回的Request对象之后会经过Scrapy处理，下载相应的内容，并调用设置的callback函数(函数可相同)。
3. 在回调函数内，可以使用 选择器(Selectors) (也可以使用BeautifulSoup, lxml 或者想用的任何解析器) 来分析网页内容，并根据分析的数据生成item。
4. 最后，由spider返回的item将被存到数据库(由某些 Item Pipeline 处理)或使用 Feed exports 存入到文件中。

3.4.3设置（Settings）

Scrapy设定(settings)提供了定制Scrapy组件的方法。可以控制包括核心(core)，插件(extension)，pipeline及spider组件。设定为代码提供了提取以key-value映射的配置值的的全局命名空间(namespace)。 设定可以通过多种方式设置，每个方式具有不同的优先级。

3.4.4 Item Pipeline

当Item在Spider中被收集之后，它将会被传递到Item Pipeline，按照先后顺序执行对Item的处理。

3.4.5 下载器中间件

下载器中间件是介于Scrapy的request/response处理的钩子框架。 是用于全局修改Scrapy request和response的一个轻量、底层的系统。

3.4.6 Spider中间件

Spider中间件是介入到Scrapy的spider处理机制的钩子框架，可以添加代码来处理发送给 Spiders 的response及spider产生的item和request。

3.5 扩展功能模块

由于实际使用中我们可能还会需要用到日志、页面服务、Item Exporters、自动限速设置和自定义扩展这几个扩展模块以便编写实际的爬虫或者调试程序，因此我们对这些模块功能也给出了如下的描述。

3.5.1 日志（Logging）

Scrapy使用Python的内置日志记录系统进行事件日志记录。日志记录可以立即使用，并且可以在记录设置中列出的Scrapy设置在某种程度上进行配置。Scrapy调用scrapy.utils.log.configure\_logging（）设置一些合理的默认值，并在运行命令时在日志设置中处理这些设置。

3.5.2 页面服务（Web Service）

Scrapy提供的用于监控及控制运行中的爬虫的web服务(service)， 并且服务是可扩展的并且默认是启用的。

3.5.3 Item Exporters

当抓取了需要的数据(Items)，就会想要将他们持久化或导出它们，并应用在其他的程序。这是整个抓取过程的目的。为此，Scrapy提供了Item Exporters 来创建不同的输出格式，如XML，CSV或JSON。

3.5.4 自动限速(AutoThrottle)扩展

该扩展能根据Scrapy服务器及爬取的网站的负载自动限制爬取速度。

设计目标：

1. 更友好的对待网站，而不使用默认的下载延迟0。
2. 自动调整scrapy来优化下载速度，使得用户不用调节下载延迟及并发请求数来找到优化的值。 用户只需指定允许的最大并发请求数，剩下的都交给扩展来完成。

下载延迟是通过计算建立TCP连接到接收到HTTP包头(header)之间的时间来测量的。

限速算法根据以下规则调整下载延迟及并发数:

1. spider永远以1并发请求数及 AUTOTHROTTLE\_START\_DELAY 中指定的下载延迟启动。
2. 当接收到回复时，下载延迟会调整到该回复的延迟与之前下载延迟之间的平均值。

3.5.5 自定义扩展

自定义扩展框架提供一个机制，使得能将自定义功能绑定到Scrapy。扩展只是正常的类，它们在Scrapy启动时被实例化、初始化。

扩展使用 Scrapy settings 管理它们的设置，通常扩展需要给它们的设置加上前缀，以避免跟已有(或将来)的扩展冲突。

扩展在扩展类被实例化时加载和激活。 因此，所有扩展的实例化代码必须在类的构造函数(\_\_init\_\_)中执行。要使得扩展可用，需要把它添加到Scrapy的 EXTENSIONS 配置中。

为了禁用一个默认开启的扩展(比如，包含在 EXTENSIONS\_BASE 中的扩展)， 需要将其顺序(order)设置为 None 。

每个扩展是一个单一的Python class。 Scrapy扩展(包括middlewares和pipelines)的主要入口是 from\_crawler 类方法， 它接收一个 Crawler 类的实例。通过这个对象访问settings，signals，stats，控制爬虫的行为。通常来说，扩展关联到 signals 并执行它们触发的任务。如果 from\_crawler 方法抛出 NotConfigured 异常， 扩展会被禁用。否则，扩展会被开启。

3.6 故障处理要求

下面是Scrapy提供的故障异常抛出及其产生的场景。

DropItem该异常由item pipeline抛出，用于停止处理item。

CloseSpider该异常由spider的回调函数(callback)抛出，来暂停/停止spider。

IgnoreRequest该异常由调度器(Scheduler)或其他下载中间件抛出，声明忽略该request。

NotConfigured该异常由Extensions、Item pipelines、Downloader middlwares、Spider middlewares组件抛出，声明其仍然保持关闭。该异常必须由组件的构造器(constructor)抛出。

NotSupported该异常声明一个不支持的特性。

日志log记录完整详细的操作动作与操作数据，用于调试和查错。

3.7 其他专门要求

无

4 运行环境规定

4.1 设备

系统：Linux、Mac、Windows

内存：256MB以上

硬盘：20G以上

显示器分辨率：800×600以上

用于开发python爬虫的PC机或网络服务器

4.2 支持软件

Python3.0+ 或 2.7

基于Windows、Linux、MacOs等平台

4.3 接口

4.3.1 硬件接口

无

4.3.2 软件接口

Python类库

4.3.3 通信接口

HTTP协议、TCP/IP协议、HTTPS协议

4.3.4 用户接口

命令行工具: 通过 scrapy 命令行工具进行控制。

Scrapy终端:一个交互终端，提供在未启动spider的情况下尝试及调试爬取代码。

telnet终端：以供检查和控制Scrapy运行的进程。

5拓展和改进

5.1 方案思路

1. 首先在框架的基础上实现一个具体的爬虫，完整爬取一个网站。通过演示将整个爬取的过程展现出来，给出最终的爬取结果。并且以实际爬取过程测试框架的并发量和负载能力，尝试给出优化方案。

2. 在实际爬取中需要对使用框架后的爬取行为过程做限定和约束，避免重复爬取和循环爬取，同时避免爬取中数据丢失或者遗漏。爬取时也需要尝试多种反爬取措施，模拟人为动作行为，避免爬取失败。

3. 根据爬取的实际效果，进一步对爬虫的下载器做优化，加速爬取下载。

4. 爬取方式增加分布式爬取,在多个主机环境下进行分布式爬取,加速爬取的速度，爬取的数据写入后端的MangoDB数据库存储。

5. 在实现的Scrapy分布式爬虫的基础上，结合布隆过滤器实现两级过滤器。单机爬虫首先在本机的布隆过滤器中对新生成的url进行过滤，过滤失败则在继续分布式的过滤器中进行过滤。

6. 增加一个前端展示的界面，方便进行爬取的操作和显示相关爬取结果，前端显示的数据从数据库中提取，需要实现相关数据接口。

7. 提供给用户一个网站链接关系设置的接口，用于设置爬取网站链接。

5.2 详细设计

5.2.1搭建测试环境

为了显示的相关方案的改进效果，需要搭建测试环境。在改进前和改进后，分别统计爬取同一网站所有目标的url所花费的起止时间，并将爬取到的内容输出到一个文档记录中，最后根据统计结果，证明改进方案是否成功。爬取对象为北航计算机学院官方网站主站（http://scse.buaa.edu.cn/buaa-css-web/initAction.action）。

5.2.2反爬改进

Scrapy本身也提供了一个ProxyMiddleware，但是它只能使用固定的IP地址。由于固定的IP地址可能遭到IP被禁用的情况，这么一来，你就不能使用本地IP进行相应的爬虫工作了。此时就需要进行反爬虫的实现，使用到的两种基本策略：伪装user agent和使用免费代理。伪装user agent方式通过准备若干个浏览器的user agent，每次发送请求的时候就从这几个user agents中随机选一个填上去，从而达到伪装的目的。但是如果对方用某段时间内某IP的访问次数来判定爬虫，然后将这些爬虫的IP都封掉的话，以上伪装就失效了。通过使用免费的代理减小单个IP的访问量，从而规避反爬。因而需要使用python代码进行代理IP地址的抓取和可用性测试，之后将可用的代理保存到相应的文件中提供给scrapy使用。

5.2.3 下载器改进

Scrapy采用Python内置的集合（set）实现内存过滤器，同时提供文件过滤器让用户选择。两者并不排斥，文件过滤器的设置是为了可以停止爬虫（结束爬虫进程），然后在之后的某一个时间内重新接着上次的爬取继续时，可以直接从文件中读取已经爬取的url。实际运行期间的过滤都是在内存过滤器（set集合实现的过滤器）进行的。默认的Python集合需要存储网页本身的url，同时为了实现管理集合以及其他一些针对集合的操作，Python需要保存的内容除了URL之外还有许多，因此，当爬取大量网页的时候占用内存会比较大。而分析发现，我们实际上只需要判断一个url是否已经出现过，而至于url是什么或者何时出现并不在乎。其次，我们也不会对已经爬取的url进行索引等操作，所以Python内置的set结合本身的许多操作相对来说是冗余的。根据以上分析，可以使用布隆过滤器（布隆过滤器可以用于检索一个元素是否在一个集合中，它的优点是空间效率和查询时间都远远超过一般的算法）来实现对已下载网页的过滤。

5.2.4 分布式爬取

Scrapy是基于Twisted异步框架的单机爬虫，目前只能在一台机器上进行爬取，而且受限于Python的GIL，在多核CPU上也只能利用到一个CPU核心。所以，我们决定将scrapy改造成一个多机分布式爬虫，实现多台机器对同一个爬取任务的共同完成。

5.2.5 分布式爬取两级过滤器

Scrapy采用Python内置的集合（set）实现内存过滤器，同时提供文件过滤器让用户选择。两者并不排斥，文件过滤器的设置是为了可以停止爬虫（结束爬虫进程），然后在之后的某一个时间内重新接着上次的爬取继续时，可以直接从文件中读取已经爬取的url。实际运行期间的过滤都是在内存过滤器（set集合实现的过滤器）进行的。

默认的Python集合需要存储网页本身的url，同时为了实现管理集合以及其他一些针对集合的操作，Python需要保存的内容除了URL之外还有许多，因此，当爬取大量网页的时候占用内存会比较大。而分析发现，我们实际上只需要判断一个url是否已经出现过，而至于url是什么或者何时出现并不在乎。其次，我们也不会对已经爬取的url进行索引等操作，所以Python内置的set集合本身的许多操作相对来说是冗余的。

结合我们决定要实现的Scrapy分布式爬虫，我们决定结合布隆过滤器实现两级过滤器。单机爬虫首先在本机的布隆过滤器中对新生成的url进行过滤，过滤失败则在继续分布式的过滤器中进行过滤。

5.2.6 爬取前端界面显示

爬取的结果在命令行下显示并不直观，同时在命令行下进行操作需要输入相关命令，操作起来相对较为复杂。我们在命令行命令的基础上，进行前端UI用以包装，设计一个前端界面，用以显示爬取的结果和操作爬取动作，进一步简化操作和方便查看爬取结果。相关显示的数据均来源于后端的数据库存储。

5.2.7 爬取链接设置

提供一个爬取网站链接设置的接口，在命令行下进行设置后，即可改变爬取的目标。