**用基于Scrapy的简单爬虫程序爬取中国科大主页通知信息**

**张劲暾 PB16111485**

1. **爬虫概述：**

**Scrapy是一种轻量级的Python 爬虫开发框架，一个简单的Scrapy爬虫项目结构如下所示：**

homework1/  
 scrapy.cfg  
 homework1/  
 \_\_init\_\_.py  
 items.py  
 pipelines.py  
 settings.py  
 spiders/  
 \_\_init\_\_.py  
 …

**我利用Scrapy框架编写了一个非常简单的爬虫程序homework1，用来爬取中国科大学校主页上的通知。**

1. **爬虫设计：**
2. **items.py：内容抓取容器定义**

import scrapy

class Homework1Item(scrapy.Item):

title = scrapy.Field()

link = scrapy.Field()

desc = scrapy.Field()

1. **ustc\_spider.py：爬虫程序，爬取中国科大学校主页上的通知成为item列表返回**

import scrapy

from scrapy.selector import Selector

from homework1.items import Homework1Item

class USTCSpider(scrapy.Spider):

name = "ustc"

allowed\_domains = ["ustc.edu.cn"]

start\_urls = [ "https://www.ustc.edu.cn/"]

def parse(self, response):

sel = Selector(response)

sites = sel.xpath('//ul/div/li')

#sites = sel.xpath('//ul/li')

items = []

for site in sites:

item = Homework1Item()

item['title'] = site.xpath('a/text()').extract()

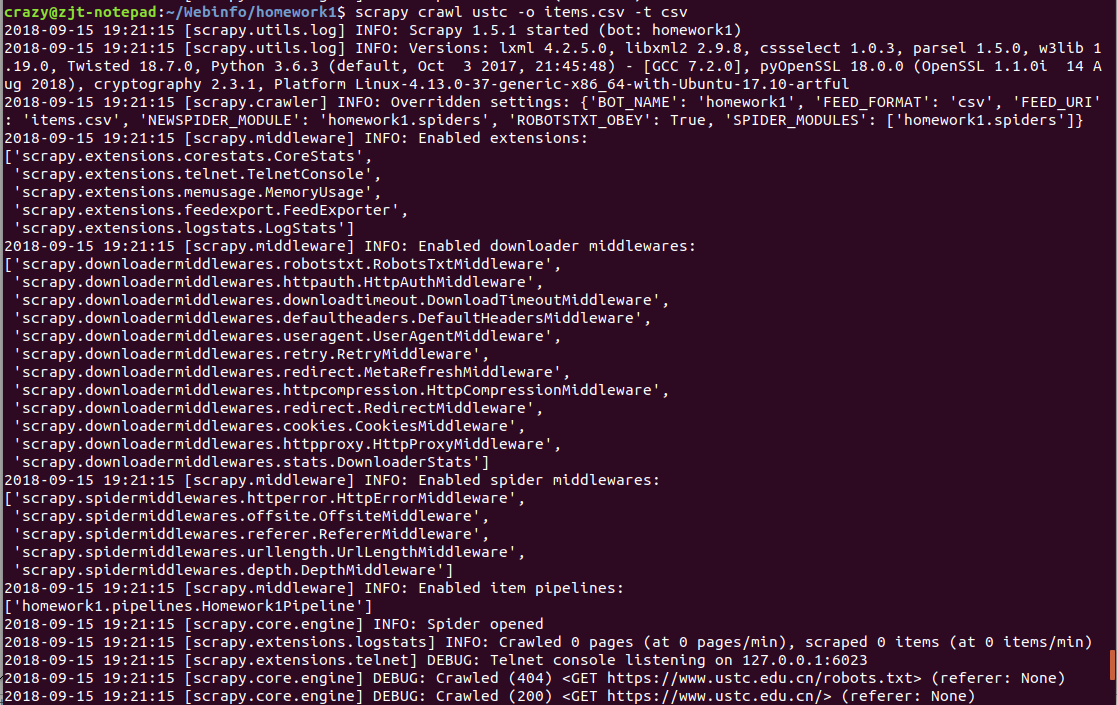
item['link'] = site.xpath('a/@href').extract()

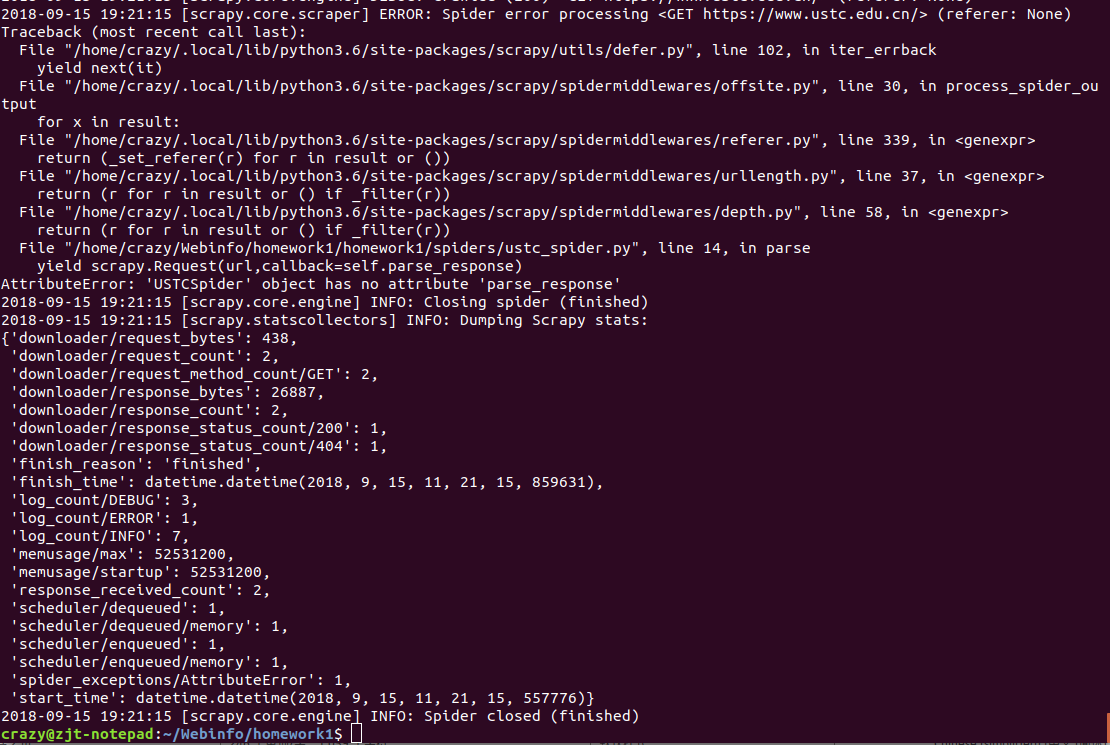
item['desc'] = site.xpath('text()').extract()

items.append(item)

return items

1. **爬取过程：**





1. **爬取结果：**

**用命令：**

**scrapy crawl ustc -o items.csv -t csv**

**将爬取到的结果输出为csv表格文件如下：**

desc,link,title

,http://news.ustc.edu.cn/2018/0914/c8091a312259/page.htm,习近平出席全国教育大会并发表重要讲话

,http://news.ustc.edu.cn/2018/0913/c15884a312180/page.htm,我校举办“我是科大人”迎新暨教师节专场联欢晚会

,http://news.ustc.edu.cn/2018/0912/c15884a312153/page.htm,学校举办第十场科大精神报告会 汤洪高教授畅谈科大精...

,http://news.ustc.edu.cn/2018/0914/c15884a312257/page.htm,我校举行纪念建校60周年《种子天堂》专场话剧演出

,http://news.ustc.edu.cn/2018/0914/c15884a312308/page.htm,中国科大成功研制具有纳升级精度的手持式数字移液器

,http://news.ustc.edu.cn/2018/0914/c15884a312260/page.htm,生命科学与医学部召开研究生教育研讨交流会

,http://news.ustc.edu.cn/2018/0914/c15884a312254/page.htm,2018年学生社团游园会举办

,http://news.ustc.edu.cn/2018/0914/c15884a312307/page.htm,第四届中日几何会议召开

,/2018/0909/c15545a311807/page.htm,一周会议安排（2018年9月10日—17日）

,/2018/0914/c2014a312234/page.htm,关于校庆期间加强校园安全和交通管理工作的通知

,/2018/0915/c2014a312319/page.htm,关于纪念中国科学技术大学建校六十周年“永恒的东风”校...

,/2018/0914/c2014a312306/page.htm,关于举办庆祝建校60周年教职工健步走活动的通知­

,/2018/0914/c2014a312248/page.htm,关于中秋节、国庆节放假的通知

,/2018/0914/c2014a312302/page.htm,关于发放2018年度住房货币补贴的通知

,/2018/0913/c2014a312195/page.htm,关于纪念中国科学技术大学建校六十周年“珞珈遇见瀚海”...

,/2018/0913/c2014a312191/page.htm,关于校博物馆开馆的通知

,/\_redirect?siteId=44&columnId=1995&articleId=196280,学习宣传贯彻十九大精神

,http://news.ustc.edu.cn/8050/list.htm,习近平总书记考察中国科大专题

,http://www.news.cn/nzzt/16/,新华社：解码中国科大

,http://djyszw.ustc.edu.cn/xxzt/gxszhy/,学习贯彻高校思政工作会议精神

,http://djyszw.ustc.edu.cn/,中国科大理论学习网

,http://xxgk.ustc.edu.cn/,中国科大信息公开平台

,/\_redirect?siteId=44&columnId=1995&articleId=196279,中国科大60周年校庆

,http://finance.ustc.edu.cn/list.php?MenuID=012,中央科研资金管理政策专题

,http://idea.cas.cn/,科学智慧火花

,http://ustcnet.ustc.edu.cn,网络信息服务

,http://www.job.ustc.edu.cn/,中国科大就业信息网

,http://qybx.ustc.edu.cn/main.htm,“全院办校、所系结合”专题网

,http://yz.ustc.edu.cn/,

,https://zsb.ustc.edu.cn/,

,http://djyszw.ustc.edu.cn/xxzt/xysbd/,

,http://fc.ustc.edu.cn/,

,http://iat.ustc.edu.cn/,

,http://cpst.ustc.edu.cn,

,http://www.quantum2011.org/main.htm,