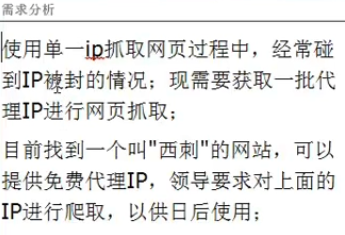
对“西刺”网站的爬取

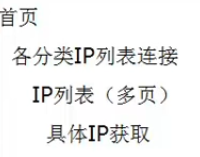
# 步骤：

## 需求分析

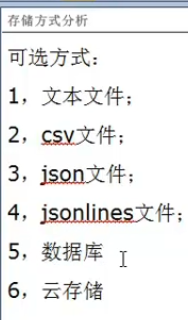


## 分析网站结构

<http://www.xicidaili.com/>



## 存储方式分析



## 编写代码：Scrapy代码

### 创建工程及spider文件

### 编写item

### 编写spider

### 编写及配置pipeline

### 调试运行代码

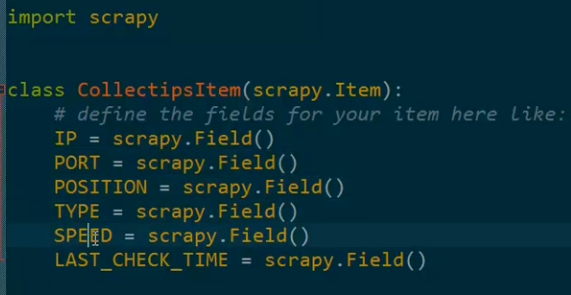
# 编写代码

## 创建工程及spider文件

**scrapy startproject collectIps**

**scrapy genspider xicidaili xicidaili.com**

## 编写item



**import** scrapy  
  
**class** CollectipsItem(scrapy.Item):  
 *# define the fields for your item here like:* IP = scrapy.Field()  
 PORT = scrapy.Field()  
 POSITION = scrapy.Field()  
 TYPE = scrapy.Field()  
 SPEED = scrapy.Field()  
 LAST\_CHECK\_TIME = scrapy.Field()

## 编写spider

代码如下：

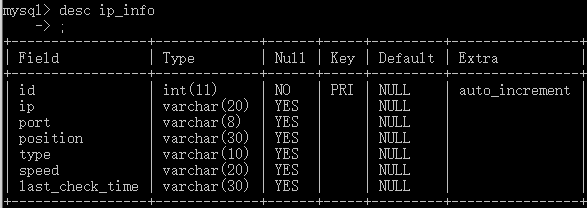
*# -\*- coding: utf-8 -\*-***import** scrapy  
**from** collectIps.items **import** CollectipsItem  
**import** logging  
  
**class** XicidailiSpider(scrapy.Spider):  
 name = **'xicidaili'** *# allowed\_domains = ['xicidaili.com']  
 # start\_urls = ['http://www.xicidaili.com/',  
 # ]* **def** start\_requests(self):  
 reqs = []  
 **for** i **in** range(1,2):  
 req = scrapy.Request(url=**"http://www.xicidaili.com/nn/%s"**%i,callback=self.parse,headers={**'user-agent'**:**'Mozilla/5.0'**})  
 reqs.append(req)  
 **return** reqs  
  
 **def** parse(self, response):  
 ip\_list = response.xpath(**"//table[@id='ip\_list']"**)  
 trs = ip\_list[0].xpath(**'tr'**)  
 *# logging.info(trs)* items = []  
 **for** ip **in** trs[1:]:  
 **try**:  
 pre\_item = CollectipsItem()  
 pre\_item[**'IP'**] = ip.xpath(**'td[2]/text()'**)[0].extract()  
 pre\_item[**'PORT'**] = ip.xpath(**'td[3]/text()'**)[0].extract()  
 pre\_item[**'POSITION'**] = ip.xpath(**'td[4]'**)[0].xpath(**'a/text()'**)[0].extract()  
 pre\_item[**'TYPE'**] = ip.xpath(**'td[5]/text()'**)[0].extract()  
 pre\_item[**'SPEED'**] = ip.xpath(**'td[7]/div[@class="bar"]/@title'**).re(**'\d{0,2}\.\d{0,}'**)[0]  
 pre\_item[**'LAST\_CHECK\_TIME'**] = ip.xpath(**'td[10]/text()'**)[0].extract()  
 items.append(pre\_item)  
 **except**:  
 logging.info(**'!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!'**)  
 **continue**;  
 **return** items

## 在MySQL数据库创建表格

**创建数据库**：create database xicidaili default character set utf8;

**创建设计表格**：

**create table ip\_info(id int primary key auto\_increment,ip varchar(20),  
port varchar(8),position varchar(30),  
type varchar(10),speed varchar(20),last\_check\_time varchar(30));**



## 编写pipeline及settings.py配置

**import** pymysql  
**import** pymysql.cursors  
**import** logging  
  
**class** CollectipsPipeline(object):  
 **def** process\_item(self, item, spider):  
 *# config = {  
 # 'host': '127.0.0.1',  
 # 'port': 3306,  
 # 'user': 'root',  
 # 'password': 'zhao12',  
 # 'db': 'xicidaili',  
 # 'charset': 'utf8mb4',  
 # 'cursorclass': pymysql.cursors.DictCursor  
 # }  
 # myConnector = pymysql.connect(\*\*config)* CONFIG\_MYSQL = spider.settings.get(**'CONFIG\_MYSQL'**)  
 myConnector = pymysql.connect(\*\*CONFIG\_MYSQL)  
  
 logging.info(**"%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%"**)  
 logging.info(type(CONFIG\_MYSQL)) *#dict* logging.info(**"%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%"**)  
 *#  
 # myConnector = pymysql.connect(host= '127.0.0.1',port= 3306,user='root',password='zhao12',  
 # db='xicidaili',charset='utf8mb4',cursorclass=pymysql.cursors.DictCursor)* cur = myConnector.cursor() *# 获得游标* logging.info(**'\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*'**)  
 logging.info(type(myConnector))  
 logging.info(type(cur))  
 sql = (**"insert into ip\_info(ip,port,type,position,speed,last\_check\_time)"  
 "values(%s,%s,%s,%s,%s,%s)"**)  
 lis = (item[**"IP"**],item[**"PORT"**],item[**"TYPE"**],item[**"POSITION"**],item[**"SPEED"**],item[**"LAST\_CHECK\_TIME"**])  
 **try**:  
 **with** myConnector.cursor() **as** cunsor:  
 cur.execute(sql, lis)  
 **except** Exception **as** e:  
 logging.info(**"Insert error:"**)  
 logging.error(e)  
 myConnector.rollback()  
 **else**:  
 myConnector.commit()  
 cur.close()  
 myConnector.close()  
 **return** item

**settings.py 文件**：

**import** pymysql.cursors  
BOT\_NAME = **'collectIps'**SPIDER\_MODULES = [**'collectIps.spiders'**]  
NEWSPIDER\_MODULE = **'collectIps.spiders'**CONFIG\_MYSQL={**'db'**:**'xicidaili'**,**'port'**:3306,**'user'**:**'root'**,**'passwd'**:**'zhao12'**,  
 **'host'**:**'127.0.0.1'**,**'use\_unicode'**:**'True'**,**'charset'**:**'utf8mb4'**,**'cursorclass'**:pymysql.cursors.DictCursor}

# 默认是注释的，所以需要打开  
**ITEM\_PIPELINES = {  
 'collectIps.pipelines.CollectipsPipeline': 300,  
}**

## 调试运行代码及结构

首先，查看工程下的爬虫，然后运行爬虫。

**scrapy list**

**scrapy crawl xicidaili**

数据库的结构：

