基于scrapy-redis的第二种形式的分布式爬虫：

1. 基于RedisSpider实现的分布式爬虫（网易新闻）
   1. 代码修改（爬虫类）：
      1. 导包：from scrapy\_redis.spiders import RedisSpider
      2. 将爬虫类的父类修改成RedisSpider
      3. 将起始url列表注释，添加一个redis\_key（调度器队列的名称）的属性
   2. redis数据库配置文件的配置redisxxx.conf：
      1. #bind 127.0.0.1
      2. protected-mode no
   3. 对项目中settings进行配置：

#配置redis服务的ip和端口  
REDIS\_HOST = 'redis服务的ip地址'  
REDIS\_PORT = 6379  
#REDIS\_PARAMS = {‘password’:’123456’}

# 使用scrapy-redis组件的去重队列  
DUPEFILTER\_CLASS = "scrapy\_redis.dupefilter.RFPDupeFilter"  
# 使用scrapy-redis组件自己的调度器  
SCHEDULER = "scrapy\_redis.scheduler.Scheduler"  
# 是否允许暂停  
SCHEDULER\_PERSIST = True

* + 1. 使用可以被共享的管道

ITEM\_PIPELINES = {  
 #'wangyiPro.pipelines.WangyiproPipeline': 300,  
 'scrapy\_redis.pipelines.RedisPipeline': 400,  
}

* 1. 开启redis数据库的服务：redis-server 配置文件
  2. 执行爬虫文件：scrapy runspider wangyi.py
  3. 向调度器的队列中扔一个起始url：
     1. 开启redis客户端
     2. 向调度器队列中扔一个起始url

lpush wangyi https://news.163.com

1. UA池:
   1. 在中间价类中进行导包：

from scrapy.contrib.downloadermiddleware.useragent import UserAgentMiddleware

b) 封装一个基于UserAgentMiddleware的类，且重写该类的process\_requests方法

1. 代理池：注意请求url的协议后到底是http·还是https
2. selenium如何被应用到scrapy
   1. 在爬虫文件中导入webdriver类
   2. 在爬虫文件的爬虫类的构造方法中进行了浏览器实例化的操作
   3. 在爬虫类的closed方法中进行浏览器关闭的操作
   4. 在下载中间件的process\_response方法中编写执行浏览器自动化的操作

需求：爬取的是基于文字的新闻数据（国内，国际，军事，航空）

PROXY = [

   '173.82.219.113:3128',

   '92.243.6.37:80',

   '117.102.96.59:8080',

   '213.234.28.94:8080',

   '101.51.123.88:8080',

   '158.58.131.214:41258' ]

user\_agent\_list = [  
 "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.1 "  
 "(KHTML, like Gecko) Chrome/22.0.1207.1 Safari/537.1",  
 "Mozilla/5.0 (X11; CrOS i686 2268.111.0) AppleWebKit/536.11 "  
 "(KHTML, like Gecko) Chrome/20.0.1132.57 Safari/536.11",  
 "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.6 "  
 "(KHTML, like Gecko) Chrome/20.0.1092.0 Safari/536.6",  
 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/536.6 "  
 "(KHTML, like Gecko) Chrome/20.0.1090.0 Safari/536.6",  
 "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/537.1 "  
 "(KHTML, like Gecko) Chrome/19.77.34.5 Safari/537.1",  
 "Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/536.5 "  
 "(KHTML, like Gecko) Chrome/19.0.1084.9 Safari/536.5",  
 "Mozilla/5.0 (Windows NT 6.0) AppleWebKit/536.5 "  
 "(KHTML, like Gecko) Chrome/19.0.1084.36 Safari/536.5",  
 "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1063.0 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1063.0 Safari/536.3",  
 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_8\_0) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1063.0 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1062.0 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1062.0 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1061.1 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1061.1 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.1) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1061.1 Safari/536.3",  
 "Mozilla/5.0 (Windows NT 6.2) AppleWebKit/536.3 "  
 "(KHTML, like Gecko) Chrome/19.0.1061.0 Safari/536.3",  
 "Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/535.24 "  
 "(KHTML, like Gecko) Chrome/19.0.1055.1 Safari/535.24",  
 "Mozilla/5.0 (Windows NT 6.2; WOW64) AppleWebKit/535.24 "  
 "(KHTML, like Gecko) Chrome/19.0.1055.1 Safari/535.24"  
 ]