Scrapy框架课程介绍：

1. 框架的简介和基础使用
2. 持久化存储
3. 代理和cookie
4. 日志等级和请求传参
5. CrawlSpider
6. 基于redis的分布式爬虫

分布式爬虫：

* + - 1. 概念：多台机器上可以执行同一个爬虫程序，实现网站数据的分布爬取。
      2. 原生的scrapy是不可以实现分布式爬虫？
         1. 调度器无法共享
         2. 管道无法共享
      3. scrapy-redis组件:专门为scrapy开发的一套组件。该组件可以让scrapy实现分布式。
         1. 下载：pip install scrapy-redis
      4. 分布式爬取的流程：
         1. redis配置文件的配置：

bind 127.0.0.1 进行注释

protected-mode no 关闭保护模式

* + - * 1. redis服务器的开启：基于配置配置文件
        2. 创建scrapy工程后，创建基于crawlSpider的爬虫文件
        3. 导入RedisCrawlSpider类，然后将爬虫文件修改成基于该类的源文件
        4. 将start\_url修改成redis\_key = ‘XXX’
        5. 在配置文件中进行相应配置：将管道配置成scrapy-redis集成的管道
        6. 在配置文件中将调度器切换成scrapy-redis集成好的调度器
        7. 执行爬虫程序：scrapy runspider xxx.py
        8. redis客户端：lpush 调度器队列的名称 “起始url”

【补充】

#如果redis服务器不在自己本机，则需要在setting中进行如下配置

REDIS\_HOST = 'redis服务的ip地址'

REDIS\_PORT = 6379

【注意】近期糗事百科更新了糗图板块的反爬机制，更新后该板块的页码链接/pic/page/2/s=5135066，末尾的数字每次页面刷新都会变化，因此爬虫文件中链接提取器的正则不可写为/pic/page/\d+/s=5135066而应该修改成/pic/page/\d+