1. Python 爬虫框架知识点
   1. **Scrapy**是用python实现的一个为了爬取网站数据，提取结构性数据而编写的应用框架。使用Twisted高效异步网络框架来处理网络通信。
   2. **ScrapyEngine**：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。 此组件相当于爬虫的“大脑”，是 整个爬虫的调度中心。
   3. **Schedule**：调度器。接收从引擎发过来的requests，并将他们入队。初始爬取url和后续在页面里爬到的待爬取url放入调度器中，等待被爬取。调度器会自动去掉重复的url。
   4. **Downloader**：下载器。负责获取页面数据，并提供给引擎，而后提供给spider。
   5. Spider：爬虫。用户编些用于分析response并提取item和额外跟进的url。将额外**跟进**的url提交给ScrapyEngine，加入到Schedule中。将每个spider负责处理一个特定(或 一些)网站。
   6. **ItemPipeline**：负责处理被spider提取出来的item。当页面被爬虫解析所需的数据存入Item后，将被发送到Pipeline，并经过设置好次序
   7. **DownloaderMiddlewares**：下载中间件。是在引擎和下载器之间的特定钩子(specific hook)，处理它们之间的请求(request)和响应(response)。提供了一个简单的机制，通过插入自定义代码来扩展Scrapy功能。通过设置DownloaderMiddlewares来实现爬虫自动更换user-agent,IP等。
   8. **SpiderMiddlewares**：Spider中间件。是在引擎和Spider之间的特定钩子(specific hook)，处理spider的输入(response)和输出(items或requests)。提供了同样简单机制，通过插入自定义代码来扩展Scrapy功能。
2. 按照使用场景，可以将爬虫划分为哪几种？通用网络爬虫、聚焦网络爬虫
3. Scrapy 框架提供了基类 **Scrapy.Item** 类表示实体数据
4. Python 支持一些解析网页的技术，主要有 XPath、正则表达式、Beautiful Soup
5. 若要提取第二个 head 标签里的 a 标签的 href 属性值，正确的 XPath 语句为 selector.xpath(‘//head/li[2]/a/@href’)
6. Selenium 提供了两种等待方式，分别是隐式等待和显式等待
7. robots.txt 文件必须放置在一个站点的根目录下，且文件名必须全部小写
8. 在使用 xpath 解析网页内容时，通常可以通过 text() 来访问元素对象的文本
9. 使用正则表达式进行解析时，匹配不区分大小写，应使用 re.I
10. selenium的WebDriver中提供了多种定位页面元素的方法,若要通过链接文本定位页面元素,则选用 find\_element(s)\_by\_link\_text
11. 下列命令中，scrapy startproject mySpider 可以新建一个 Scrapy 项目
12. 从字符串的任意位置与pattern进行匹配，匹配成功返回第一个匹配的对象，否则返回None，则使用search
13. 使用 BeautifulSoup 解析网页时，需要导入 bs4 库
14. ActionChains 类能够模拟鼠标完成浏览器页面的相关操作
15. requests库的( requests.get() )函数是获取HTML网页的主要方法,对应于GET请求的方式。
16. 下列关于 Scrapy 框架描述正确的是
    1. 应用爬虫开发、数据挖掘、数据检测等领域
    2. 自带 shell 工具，方便调试爬虫项目
    3. 支持自定义 Item 和管道
    4. Scrapy 是一个用 Python 语言编写的爬虫框架
17. 用于创建一个 Image 类对象的是 Image.open() 、Image.new()
18. 结合爬取动态网页的技术，模拟完成豆瓣网的登录
    1. 模拟创建谷歌浏览器
       1. driver = webdriver.Chrome(chromedirver)
    2. 登录豆瓣网
       1. driver.get(<http://www.douban.com>)
    3. 跳转到第一个 iframe 模块
       1. driver.switch\_to.frame(0)
    4. 点击“密码登录”选项
       1. “密码登录”
    5. 输入账号
       1. usesname
    6. 输入密码
       1. password
    7. 点击“登录豆瓣”
       1. click()
    8. 识别“滑块”
       1. tcaptcha\_drag\_thumb
    9. 水平拖动滑块
       1. thumb,i,0
    10. 退出浏览器
        1. driver.quit()
19. 在PyMongo库中,用来负责与MongoDB建立连接的类是 MongoClient
20. 使用 Scrapy 框架爬取数据的工作流程，写出核心的代码指令

