1. 安装虚拟环境:virtualenv

(1)安装包Pip install virtualenv

(2)安装虚拟环境工具pip install virtualenwrapper

(3)先去环境变量里配置：新建WORKON\_HOME:...

(4)重启cmd,创建虚拟文件夹mkirtualenv ...

(5)自动进入，以后进入只需workon ...，退出deactivate

1. 豆瓣源:-i <https://pypi.douban.com/simple/>
2. 安装scrapy:

如果失败去www.lfd.uci.edu/~gohlke/pythonlibs/下找

下载后用pip install安装

1. 选择虚拟环境的安装的Python版本用：mkvirtualenv --python=C:\...\python.exe name
2. 进入虚拟环境后新建scrapy工程:scrapy startproject ArticleSpider

创建项目：先进入项目cd ArticleSpider

再scrapy genspider 项目名 要爬取的域名

其中genspider是一个模板，默认选择basic，输入scrapy genspider --list可以查看模板的种类，输入 scrapy genspider -t 模板名 项目名 要爬取的域名 可以创建指定模板

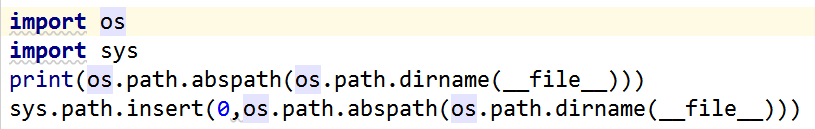
1. parse.join用于连接两个网址形成新的网址
2. 下载PIL包：pip install -i https://pypi.douban.com/simple pillow
3. 获取当前文件路径:os.path.dirname(\_\_file\_\_)

访问某个文件下级目录：os.path.join(url,子路径)

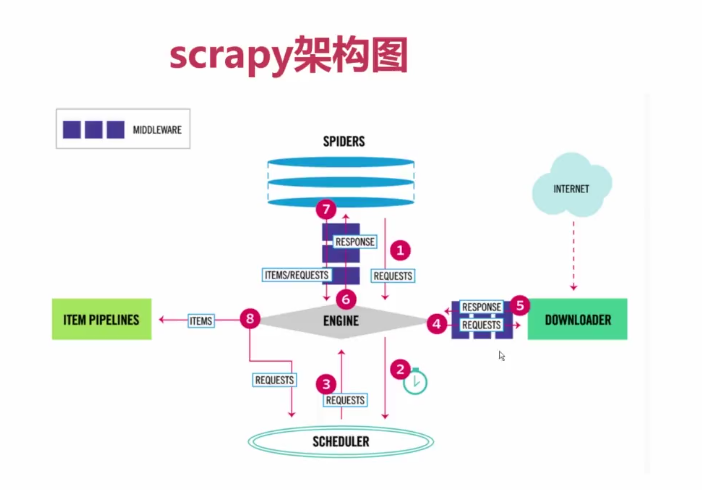
1. md5不支持unicode，需要编码成utf-8：(...).encode(‘utf-8’)
2. 安装mysql的客户端 pip install -i https://pypi.douban.com/simple/ mysqlclient
3. 安装pip install requests
4. 找到浏览器访问请求头中的user-agent来设置代码中请求头
5. 注意pycharm导入自己项目中的py文件时可以自己设置sources root就可以省略被设置为sources root的包，但是这只是在pycharm中有用，要在命令行中有用需要在settings.py中添加下列代码，这样命令行会依次在原来的路径和新设的路径下分别查找import的组件，

其中insert里面的0代表优先级最高，即命令行先去此路径下查找import的组件。

需要注意的是，一般需要把路径写活，如下所示，结合第8点



14.scrapy架构图



1. fake-useragent的useragent库地址：(用来生成user-agent)

fake-useragent.herokuapp.com/browers/0.1.4

最后一项是版本号，可以通过在workon+虚拟环境中输入pip list即可查看版本