# 数据采集子系统——使用说明

## 硬件要求

Internet访问功能：可以访问博物馆网站、连接数据库
支持爬虫相关软件，本组采用Pycharm、Navicat、谷歌浏览器（其他亦可）

## 软件要求

安装所需要使用的软件：
查看要爬取博物馆的网站浏览器：谷歌浏览器（其他亦可）
爬虫编写、运行软件：Pycharm（虚拟环境Python3.8.8）
数据库连接、查看、编辑软件：Navicat

## 系统环境要求

操作系统：Windows（Win10及以上为佳）
环境：Pycharm Python 3.8.8
安装配件：pip、Scrapy、Scrapyd（根据爬虫代码不同需要可自行补充）

## 后端服务器地址

远程连接数据库管理系统：MySQL
Ip地址：123.56.13.242
数据库端口：3306
username：root
password：Aliyun2021
数据库：museum
表：Collection、Exhibition、Museum

代码实现：

```
host="123.56.13.242",
port=3306,
user="root",
passwd="Aliyun2021",
db="museum",
charset="utf8"
```

## 文件的功能

`spiders`： 目录下的 `.py` 文件为爬虫脚本文件
`_init_.py`： 项目建立的配置文件，可空白
`items.py`： 向数据库提交数据的类结构
`middlewares.py`： 中间件，其中包含了下载器类，用于高速下载网页源码

`pipelines.py` ： 实体管道，用于处理爬虫提取的实体。即连接数据库并向其传入数据。
`settings.py` ： 配置文件，用于激活下载器和管道，配置爬虫头部信息等
`setup.py` ： 将爬虫部署到服务器上时，下载的配件的安装文件

## 爬虫的运行

### 在Pycharm上运行

终端：进入到该爬虫文件夹，打开命令窗口，输入： `scrapy crawl [爬虫名称]`
启动按钮：打开爬虫文件，点击右上角运行按钮

### 在ScrapydWeb上运行

进入到指定爬虫：Projects-指定上传人-指定爬虫，运行