

数据爬取、存储、可视化分析

一、数据的爬取

1. 环境准备

◦ 配置虚拟环境

- 安装virtualenv工具，在cmd控制台执行以下命令：

```
pip install virtualenv -i https://pypi.tuna.tsinghua.edu.cn/simple
```

- 使用virtualenv在D:\2025人工智能辅修目录下创建虚拟环境，比如my_env

```
# 切换到 D:\2025人工智能辅修 目录
cd D:\2025人工智能辅修

# 在D:\2025人工智能辅修目录下，执行：
virtualenv my_env
```

- 创建成功后，可以看到D:\2025人工智能辅修目录下创建了一个文件夹，名称为my_env
- 激活my_env环境

```
cd 到D:\2025人工智能辅修\my_env\Scripts目录，执行activate命令

# 此时自动进入my_env虚拟环境
(my_env) D:\2025人工智能辅修\my_env\Scripts>
```

- 安装组件twisted和scrapy

```
pip install twisted -i https://pypi.tuna.tsinghua.edu.cn/simple

pip install scrapy -i https://pypi.tuna.tsinghua.edu.cn/simple

# 以上下载的组件都会存放到D:\2025人工智能辅修\my_env\Lib\site-packages目录
中
```

- 切换到创建D:\2025人工智能辅修目录，执行以下命令创建scrapy项目，项目名为scrapy_demo

```
scrapy startproject scrapy_demo
```

- 使用pycharm打开scrapy_demo项目，并设置该项目的虚拟环境为my_env
- 使用scrapy命令创建一个“蜘蛛”文件

```
scrapy genspider spider01 "aiit.edu.cn"
```

二、数据的存储

三、数据的可视化分析