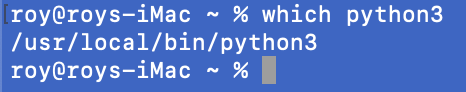
# **Yelp Spider Program**

**1. 检查Python环境是否安装**

打开Terminal，输入which python3，查看是否能得到python3的path，如果是，则环境没有问题。（python3为系统默认安装，若python3环境存在问题，麻烦给我说哦）



**2.1 安装HomeBrew（如已安装可跳过）**

打开Terminal，输入

/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"

（如以上命令有误，访问https://brew.sh/复制他官网上的）

待命令运行完成（提示有如complete文字），HomeBrew则安装完成。

**2.2 安装Chrome WebDriver**

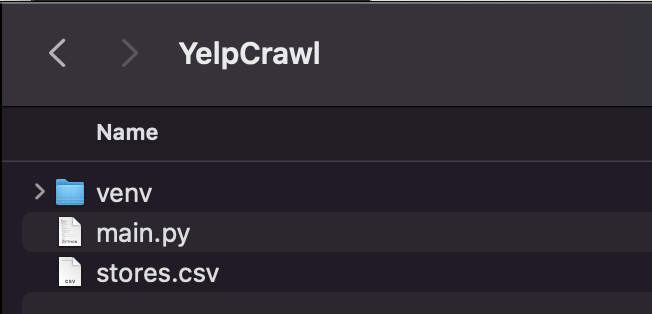
打开Terminal，输入brew install chromedriver，等待命令运行完成，如下图：



则说明driver安装完成。

**3. 文件层级**

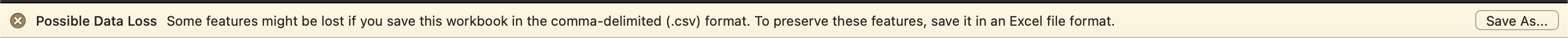
压缩包解压后，得到如下文件：



其中：

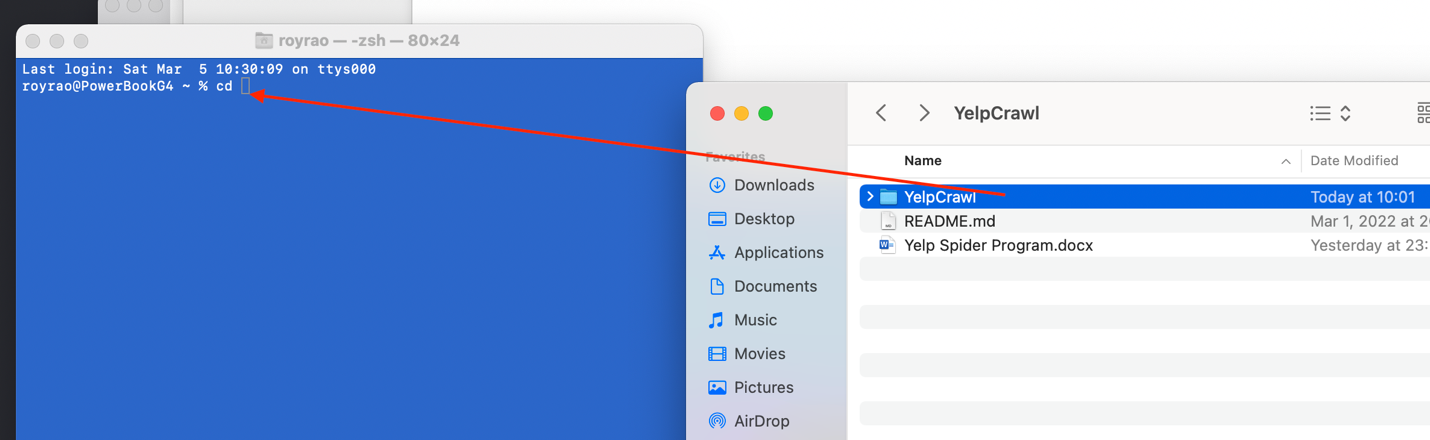
* venv: 里面包含的是这个项目的dependencies，里面的文件请妥善保留。
* main.py: 项目的脚本文件，稍候会运行它来执行命令。
* stores.csv: 你想抓取的business\_id依次保存在这里面（一排一个id，纵向输入），如果不想抓取某个商店了，直接删除掉对应的id就行（删除后产生的空行可以不用管，csv文件会自动缩进，删除id后保存就行。

修改stores.csv的时候，直接复制到每排第一个格子里就好，请勿修改格式（例如行、列高），否则无法保存为.csv格式；同时，.csv文件打开后会有如下warning，请忽略：

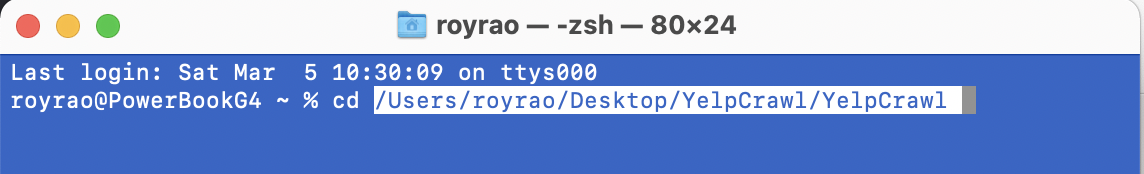


**3. 运行方法：**

首先打开Terminal。然后导航到main.py的父目录（YelpCrawl）。保持Terminal和Finder二者的窗口开启，然后拖拽python3.8到Terminal的窗口中，如下图：



如此一来，Terminal里将会自动填写文件夹的path，如下图：

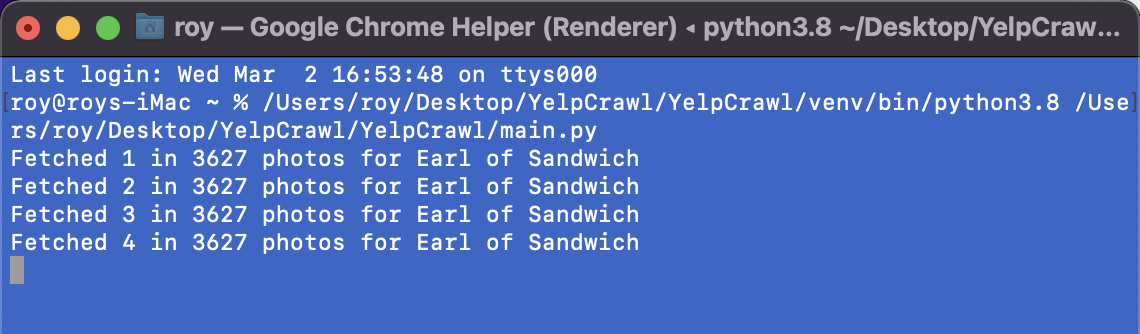


接下来，输入python3 ./main.py，并Enter：



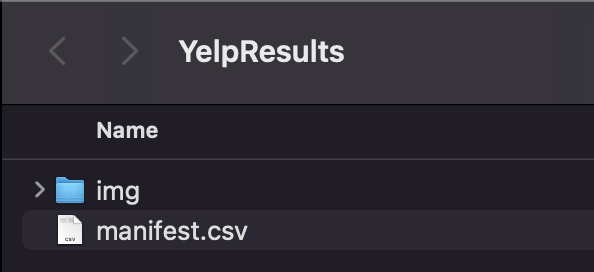
注意：python3命令后有一个空格。

两个文件拖拽完毕后，在Terminal里敲回车运行，稍作等待，出现如下output：



说明爬虫在正常运作，并开始爬取数据了。

当爬虫开始运作后，会在桌面生成一个文件夹（/Desktop/YelpResults），里面存有每个business\_id对应的文件夹，打开后有两个文件：



其中：

* img: 下载的图片会保存在这里，名称在manifest.csv中对应
* manifest.csv: 爬取到的数据，图片名称在img文件夹中能找到对应

注：项目默认依次爬取所有保存在stores.csv里的id，如果想停止爬取，直接退出Terminal即可。默认爬取间隔为2s，在将一个商店爬取完毕后，会自动爬取下一个商店，直到stores.csv中所有的商店爬取完毕后会自动停止。