1、首先将项目拷贝到ubantu虚拟机中

如何在ubantu虚拟机拷贝文件：

安装vmware-tools-distrib

<https://jingyan.baidu.com/article/93f9803f0d9d9be0e46f55ce.html>

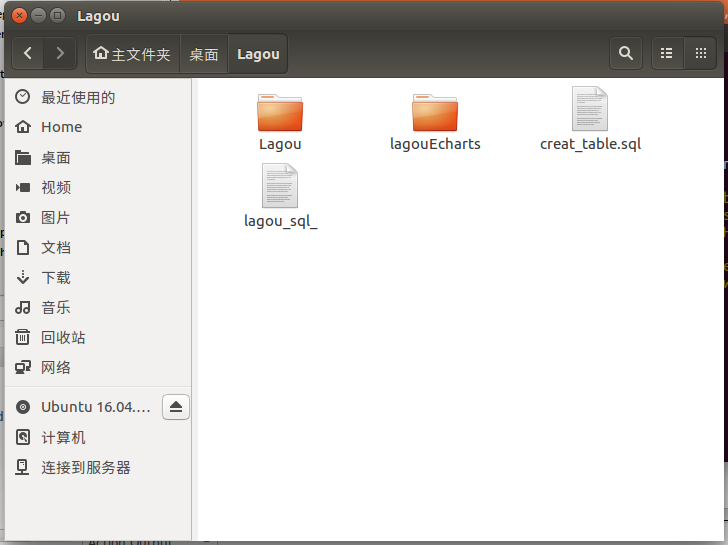
安装完成后重启拷贝文件

<https://blog.csdn.net/u010899985/article/details/54999487>

Lagou为爬虫模块

LagouEcharts为可视化模块

Sql为建表语句与统计表的建表语句与查询语句



2、安装mysql

https://www.linuxidc.com/Linux/2017-05/143864.htm

在 Ubuntu 16.04 中，默认情况下，只有最新版本的 MySQL 包含在 APT 软件包存储库中。在撰写本文时，那是 MySQL 5.7

要安装它，只需更新服务器上的包索引并安装默认包 apt-get。

sudo apt-get update

sudo apt-get install mysql-server

系统将提示您在安装过程中创建 root 密码。选择一个安全的密码，并确保你记住它，因为你以后需要它。

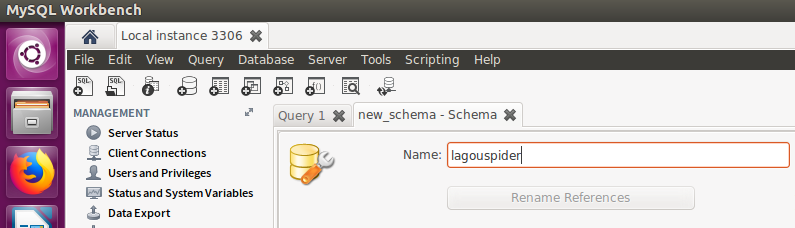
为了方便操作数据库，安装mysql可视化界面的WorkBench

sudo apt-get install mysql-workbench 就可以了 它会自动安装，中间会问你Y/N 选择Y就行

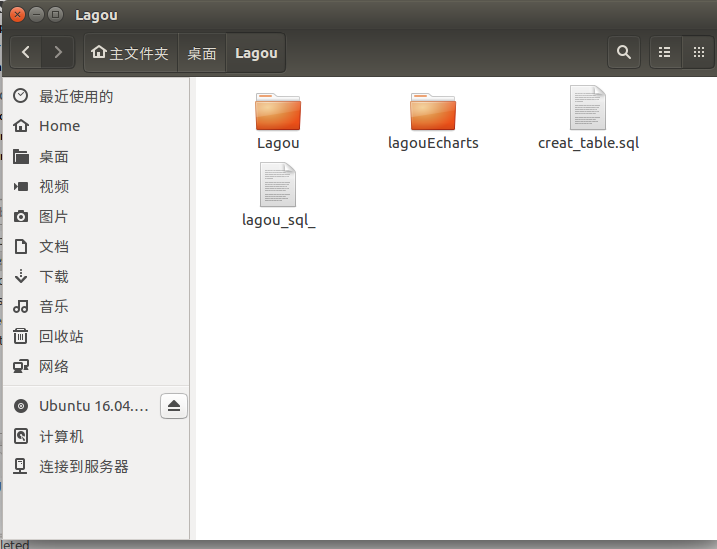
执行完以上命令，mysql-workbench安装完成，在命令行执行 mysql-workbench 就会出现软件界面

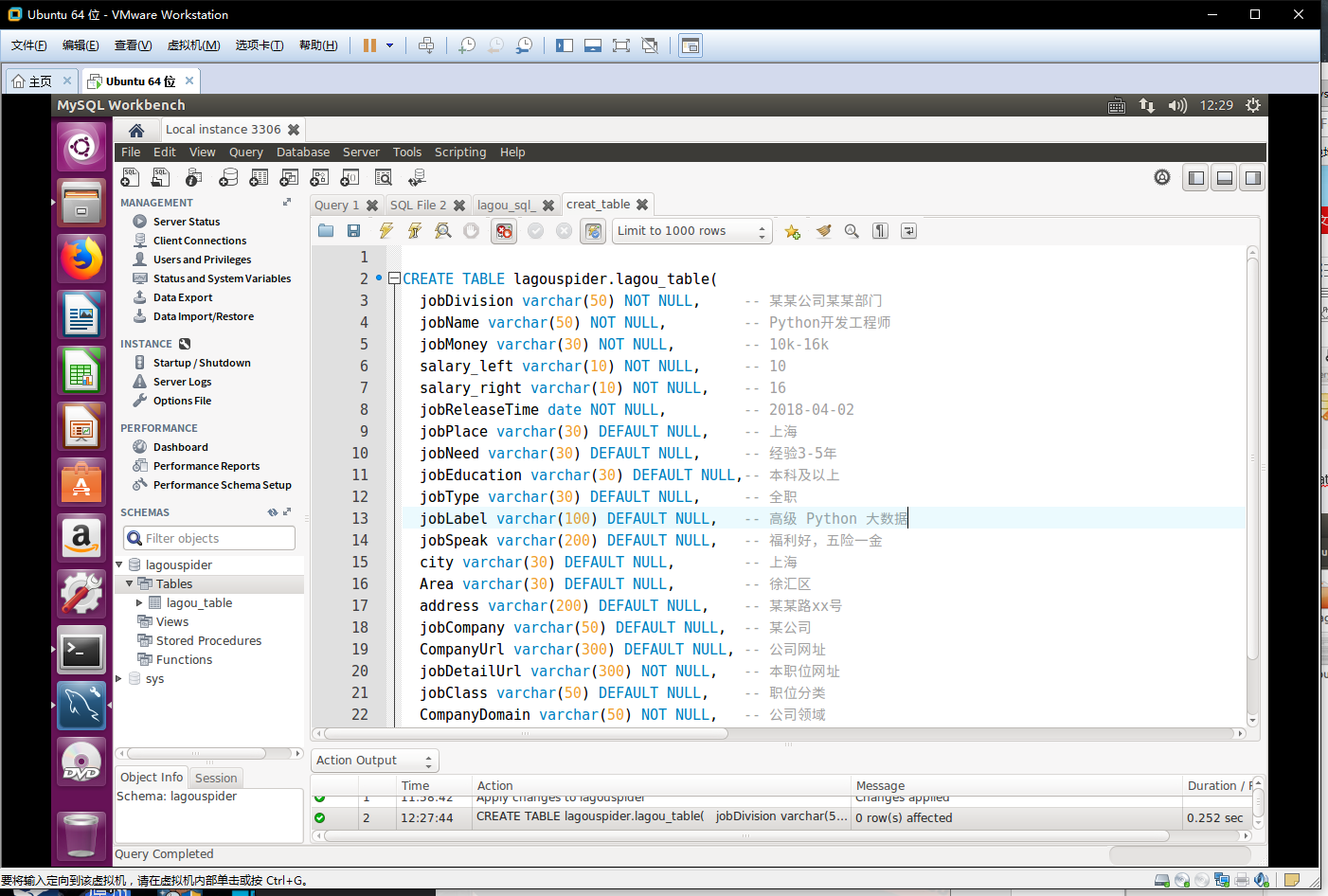
输入密码进入

创建一个lagouspider数据库



建立新的查询，打开项目中的create\_table.sql在mysql中执行建表





表已经建立好，接下来配置爬虫环境

3、python3环境

第一步：

检查一下有没有python3，在ubuntu下很好检查，直接敲命令即可。命令如下：

python3

ubuntu一般都自带python和python3。如果没有的话，需要安装一下。如下网站。

顺便找一个网站：http://www.cnblogs.com/yasmi/p/4832461.html （自行百度）

第二步：

我这里先安装了一下pip3，以防万一。

先检查一下：

pip3 –version

没有的话，会提示怎么安装。

sudo apt-get install python3-pip

安装完需要更新一下

pip3 install –upgrade pip

第三步：

安装一些依赖

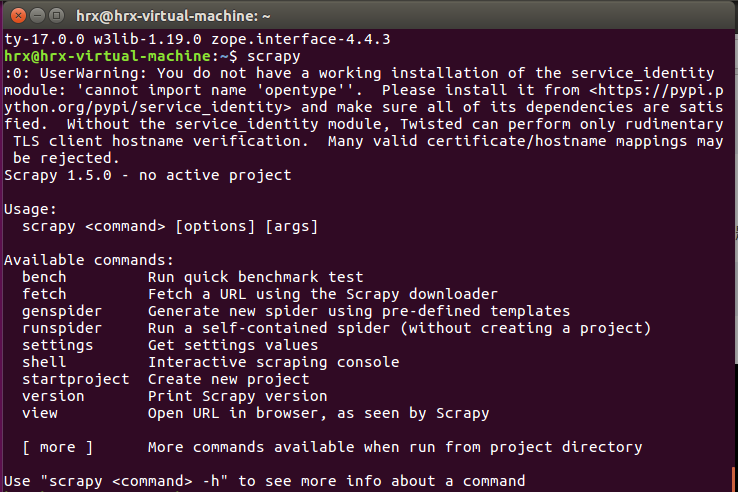
sudo apt-get install python3 python3-dev python-pip libxml2-dev libxslt1-dev zlib1g-dev libffi-dev libssl-dev

第四步：

直接用pip3工具安装scrapy即可。

sudo pip3 install scrapy

如下，即安装完成



在终端执行 sudo pip3 install pymsql安装python连接mysql的库

4、项目设置

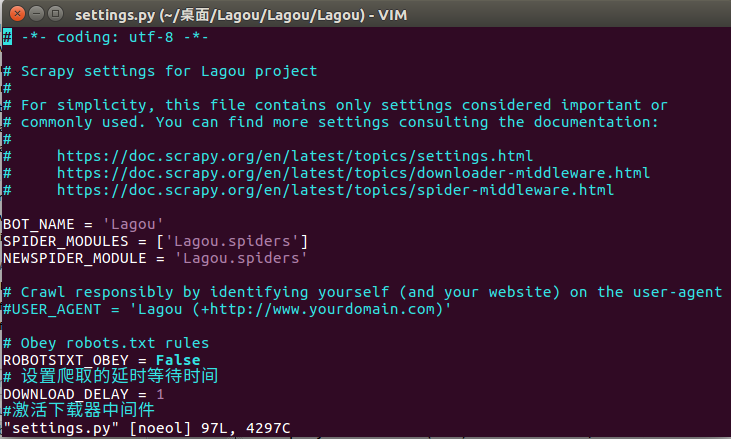
到项目中修改数据库配置。

可以安装sudo apt-get install vim-gtk编辑代码

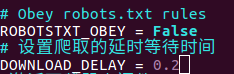
也可使用pycharm编辑

跳转到Lagou有settings的目录

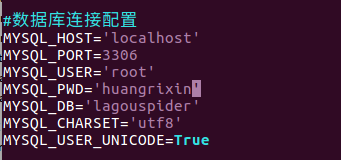
执行vi settings.py进入



修改爬取等待时间，在初次爬取设置速度较快

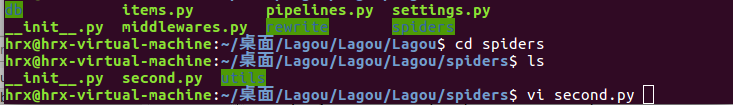


修改数据库配置为本机配置



Esc :wq 保存并退出

再跳转到spiders文件夹下



编辑second.py

由于是首次爬虫的启动，需要将page设置为30

将cookie\_city设置为 QUANGUO

在第二个函数中将Url去重关闭，注释部分代码，然后ESC ：wq保存退出

在进行了每个cookie\_city的爬取后，职位更新速度由于不是那么的频繁，我们将url去重启动，再将下图中上面4行代码注释，启用下面被注释掉的代码，此时settings中的爬取延时可以设置为1s-3s

代码设置如下：

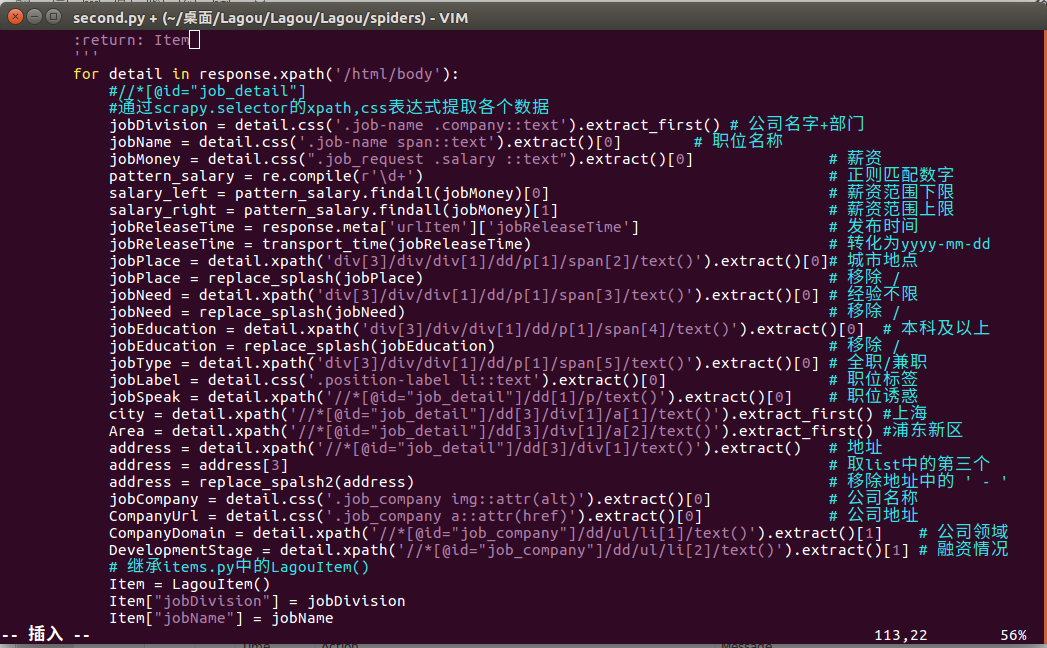


由于网站的不定时更新问题，parse\_Details方法中的xpath表达式可能发生一些变化

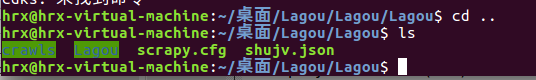
其中xpth开头为div[2] 或者div[3]的 如果其中一个不可用即替换成另外一个

将div[2]修改为div[3]ls

如下

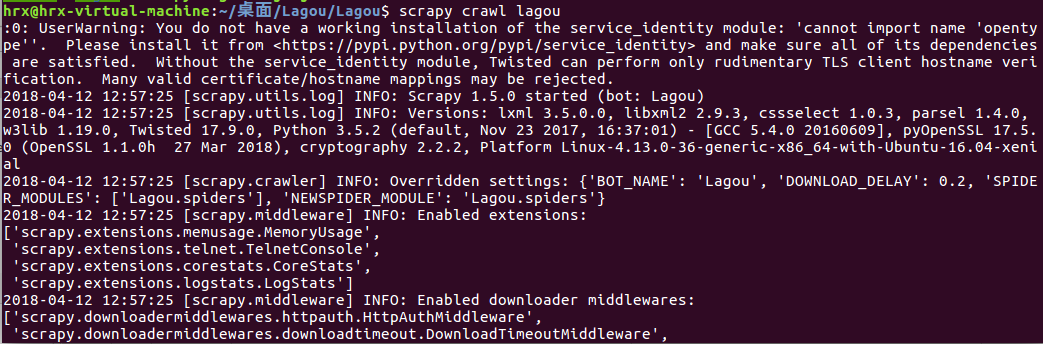


Cd .. 跳转至上一层目录

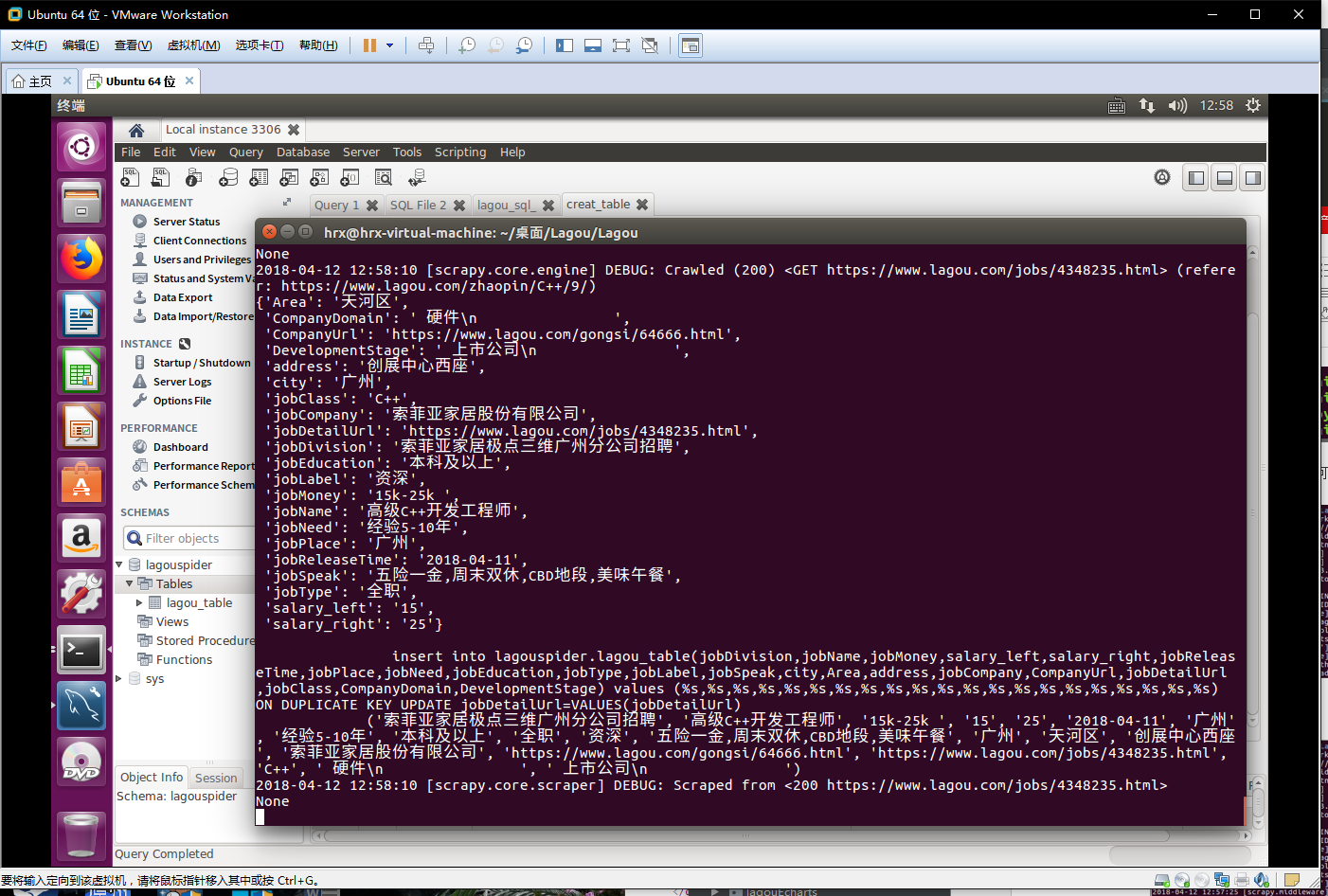


输入 Scrapy Crawl lagou 即可启动爬虫

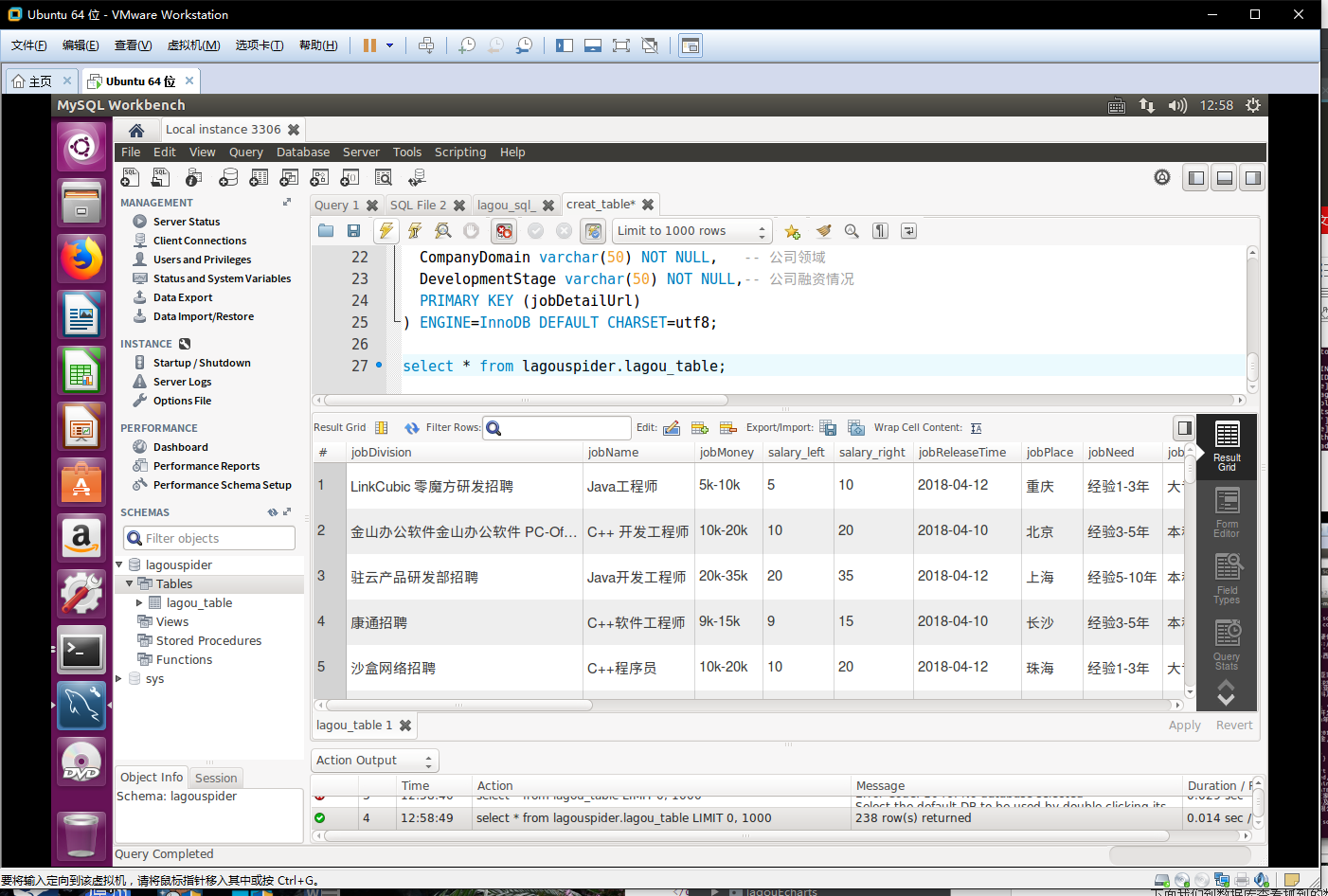
如图



下图为爬虫运行过程



下面我们到数据库查看抓到的数据

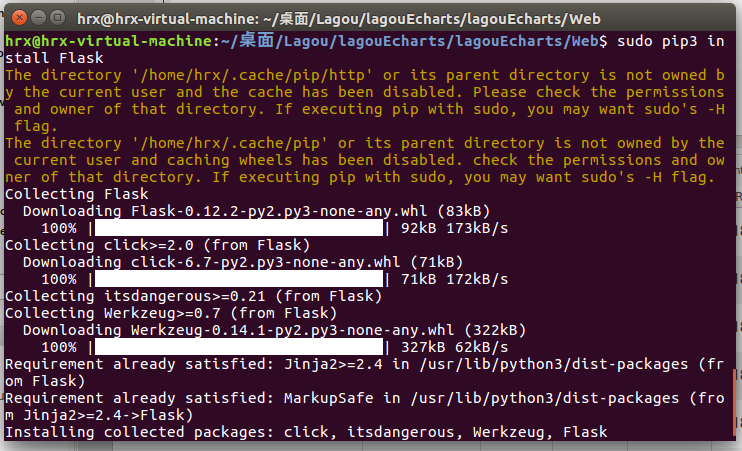


大功告成！

等待second.py中的城市设置为全国爬取完毕后，要轮流切换至设定中的不同城市站点，例如北京对应BEIJING ，上海对应SHANGHAI ，具体要到settings.py中查看。依次爬取各个站点的数据后，我们就可以实现可视化了！

4、Flask配置

执行sudo pip3 install Flask



安装pyecharts： sudo pip3 install pyecharts

安装pyecharts所需要的资源库

1. Sudo pip3 install echarts-countries-pypkg
2. Sudo pip3 install echarts-china-provinces-pypkg
3. Sudo pip3 install echarts-china-cities-pypkg

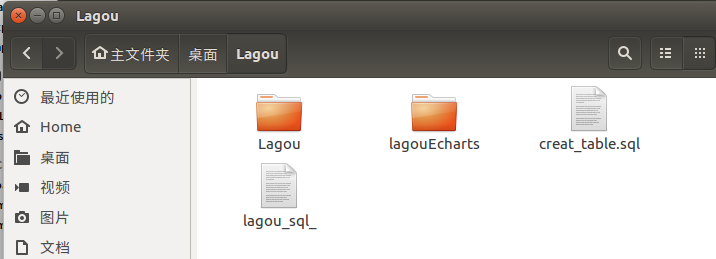
安装jieba分词库: sudo pip3 install jieba

安装wordcloud词云库: sudo pip3 install wordcloud

安装matplotlib数学绘图库:sudo pip3 install matplotlib

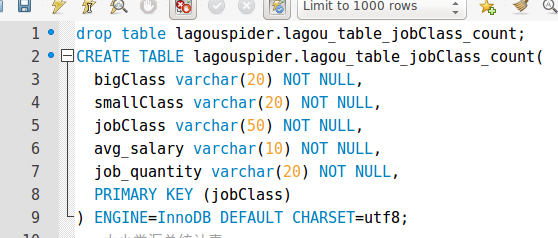
有关大小类分类汇总统计表：

在我们爬取到一定量的数据后，由于本次查询用到的统计数据比较复杂，需要使用现有的表中数据生成统计表，打开主文件下的lagou\_sql\_

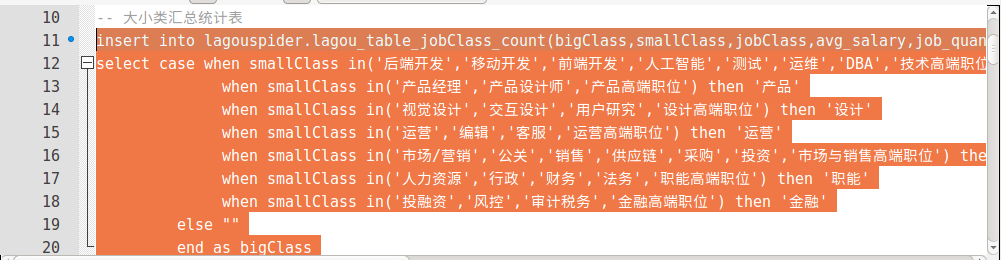


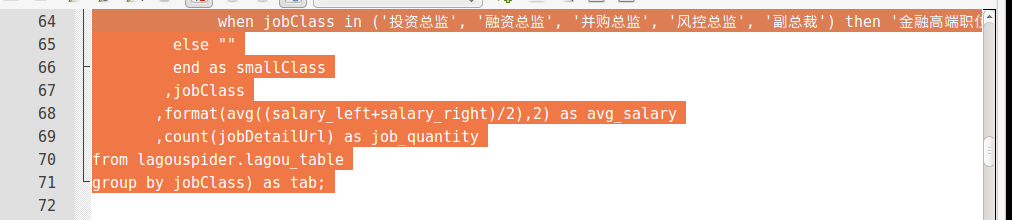
执行下面的代码，创建大小分类汇总统计表

在我们数据新增很多的时候，就需要重新运行下面的代码，先删除表，再建立表，再将当前的数据统计后插入到目标表中。



执行如下代码：

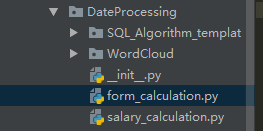


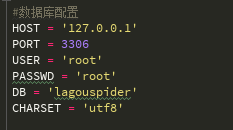


创建了大小分类汇总统计表

接下来需要将Web项目中的一些配置修改为当前计算机环境的配置：

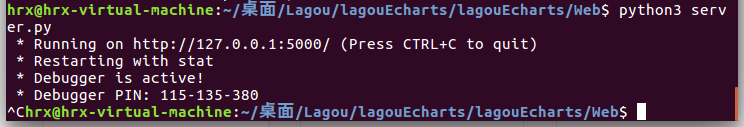
From 与 salary 文件





本次已经将爬好的数据制作了词云图放置在资源文件中

在Web目录下执行python3 server.py即可启动Flask服务端



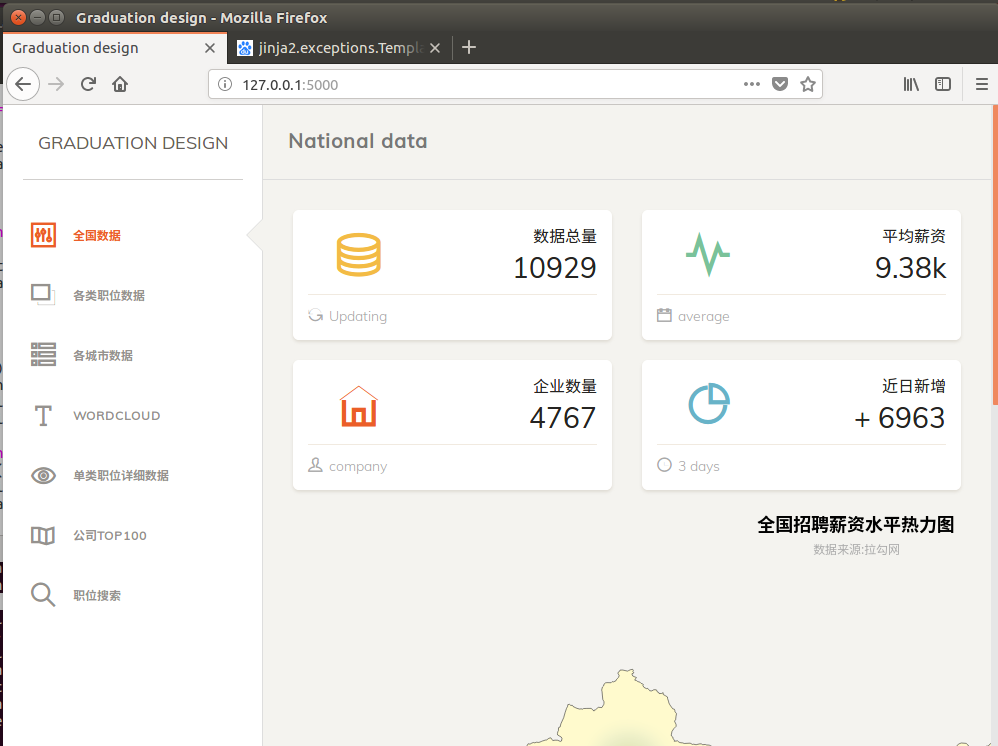
输入 127.0.0.1：5000访问首页

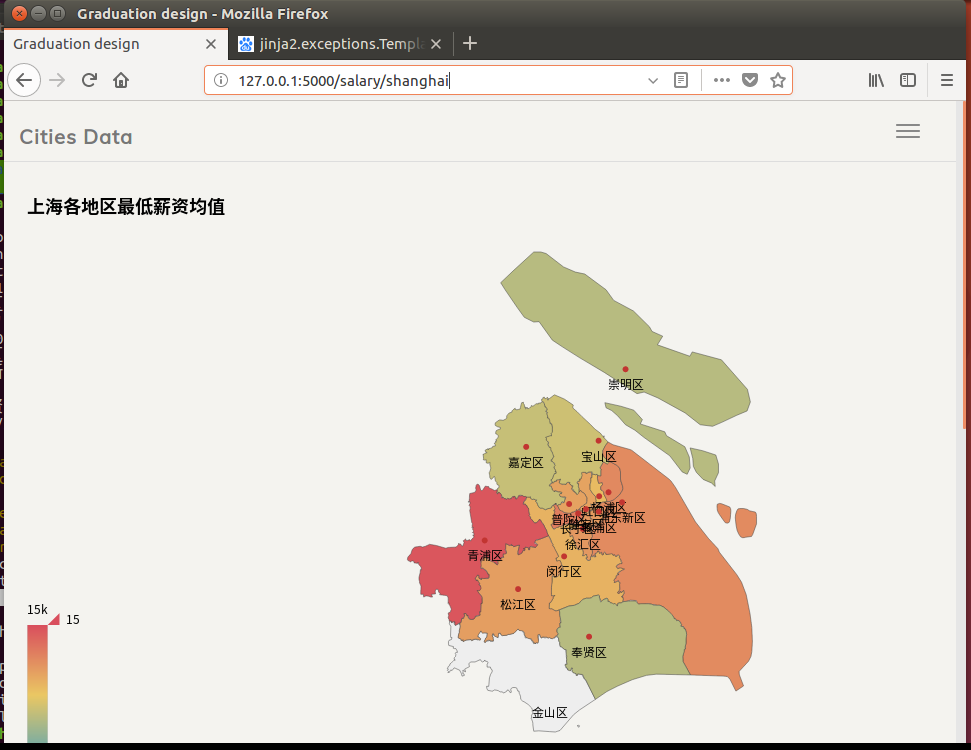
踩过的坑：由于ubantu中python代码区分大小写，如遇到

jinja2.exceptions.TemplateNotFound:

的错误，需要到templates与server.py代码查看对应的模板文件名与server.py中的是否一致

下图为首页

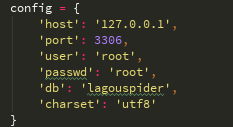


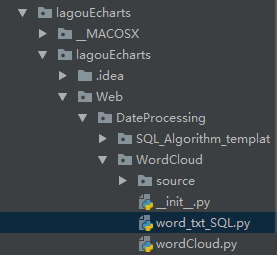


有关词云图的绘制：

想要生成新的词云图，需要到运行word\_txt\_SQL.py文件读取当前数据库的词语，生成txt文件会存入source文件中.

记得先修改数据库配置，用户名与密码





再进入wordCloud.py文件中设置当前环境的目录，



需要将文件写入static/images，修改上面的目录为当前系统此文件的绝对目录，运行wordCloud.py，生成了新的词语图

