PYTHON数据采集

18-19第二学期-软件1705（创新）

0-准备知识-HTML

1.什么是HTML?

计算机科学与技术学院
College of Computer Science and Technology

- HTML 是用来描述网页的一种语言，指的是**超文本标记语言** (Hyper Text Markup Language)，不是一种编程语言（不具备计算机编程语言的选择和循环结构），而是一种**标记语言** (markup language)。
- HTML 标记标签通常被称为 **HTML 标签** (HTML tag)。
- HTML 标签是由**尖括号包围的关键词**，比如 <html>。HTML 标签**通常是成对出现**的，比如 和 。标签对中的第一个标签是**开始标**签，第二个标签是**结束标签**。开始和结束标签也被称为**开放标签**和**闭合标签**。

0-准备知识-HTML

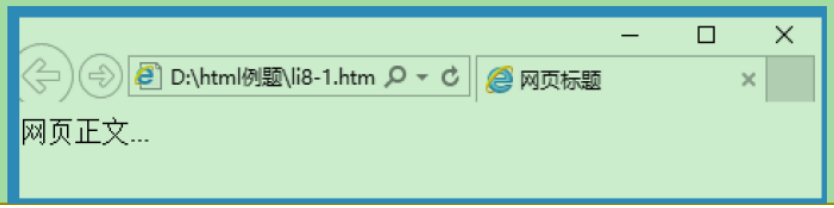
2.HTML文件的基本结构

【例8-1】我的第一个网页.html。

源代码

```
1  <html>
2    <head>
3      <title>网页标题</title>
4    </head>
5    <body>
6      网页正文...
7    </body>
8  </html>
```

运行结果



1.HTML 元素

HTML 文档是由 HTML 元素定义的。HTML 元素指的是从开始标签（start tag）到结束标签（end tag）之间的所有代码，如下表格中所示。

注释：开始标签常被称为开放标签（opening tag），结束标签常称为闭合标签（closing tag）。

开始标签	元素内容	结束标签
	这是一个段落！	

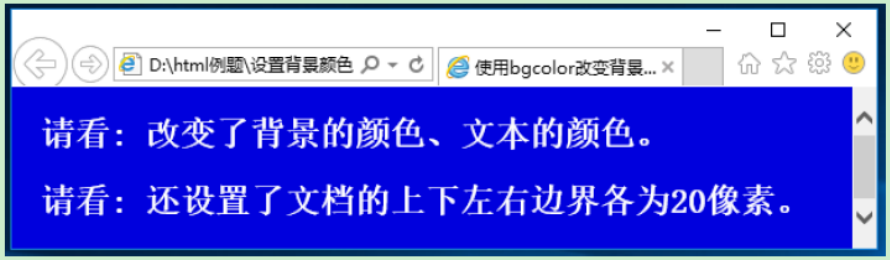
2.HTML 属性

- HTML 标签可以拥有属性。属性提供了有关 HTML 元素的更多的信息，属性总是以名称/值对的形式出现，比如：name="value"，属性总是在 HTML 元素的开始标签中规定，属性值应该始终被包括在引号内。双引号是最常用的，不过使用单引号也没有问题。
- 例如：HTML 链接由 <a> 标签定义，链接的地址在 href 属性中指定：
- 链接到山东理工大学

4.【例8-2】设置网页文档的背景、前景颜色及上下左右四个页面边距。

计算机科学与技术学院
College of Computer Science and Technology

```
<html>
  <head>
    <title>使用bgcolor改变背景的颜色</title>
  </head>
  <!--设置网页的背景色（兰色）、前景色（白色）及上下左右四个边界值-->
  <body bgcolor="#0000ff" text="ffffff"  leftmargin="20" rightmargin="20"  topmargin="20"  bottommargin="20">
    <h2>请看：改变了背景的颜色、文本的颜色。</h2>
    <h2>请看：还设置了文档的上下左右边界各为20像素。</h2>
  </body>
</html>
```



8.3.4 HTML主要标记

1.文档开始与结束标记<html>…</html>
2.头部标记<head>…</head>
3.标题标记 <title>…</title>
4.正文标记<body>…</body>
5.水平线标记<hr >
6.注释标记<! --注释内容-->
7.字体设置…
8.标题样式<hn>…<hn>（n取值可以是1至6）

8.3.4 HTML主要标记

9.文本格式化

10.段落标记 …

11.换行标记

12.列表标记

13.图像标记<img>

14.超链接标记<a>

15.表格标记<table>…</table>

16.字符实体标记

转到附录B

【实例】例8-7插入图像示例。

```
<html>
  <body>
    <p>
      可爱的小兔：
      <img src="小兔子.jpg" width="500" height="323" alt="可爱的小兔子" align="middle" border=
    </p>
  </body>
</html>
```

可爱的小兔：

0-准备知识

- 传统数据采集与基于互联网的数据采集-译者序
- https://github.com/REMitchell/python-scraping

3 创建和运行 “.ipynb” 文件

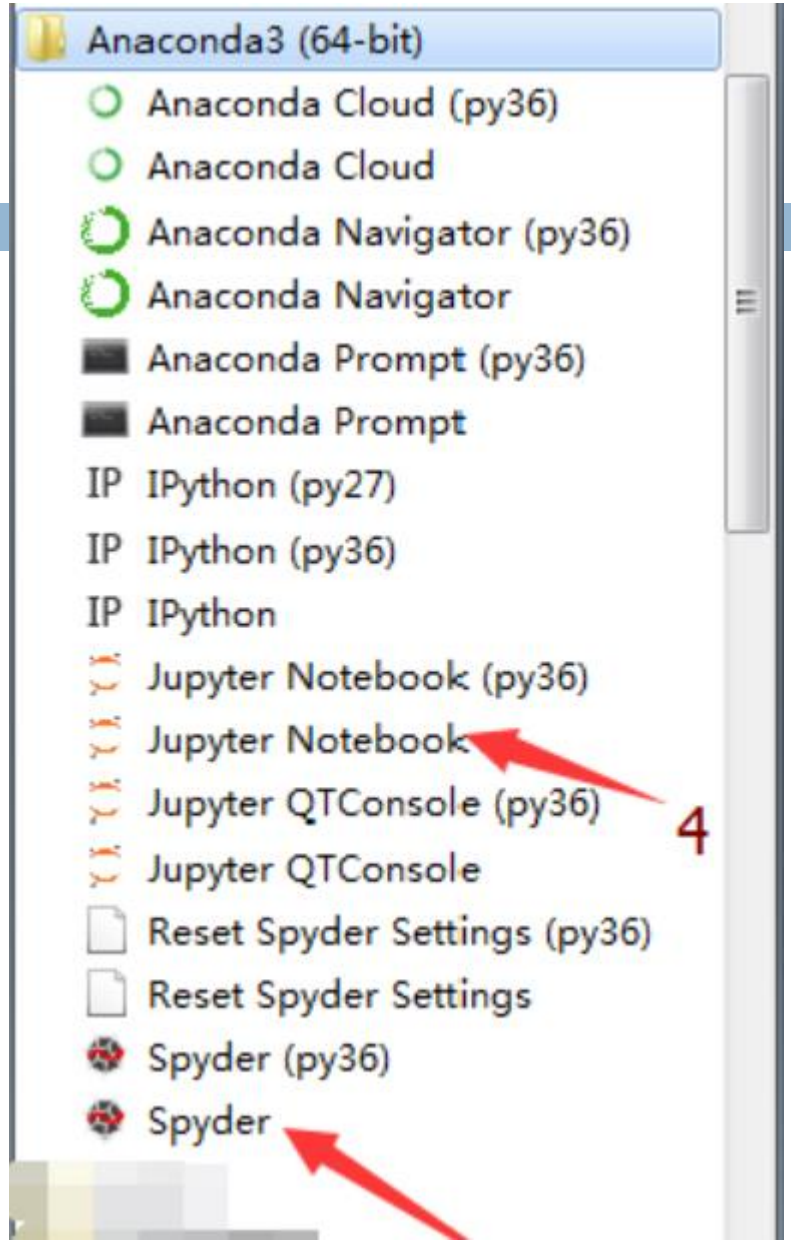
“.ipynb” 文件是使用 Jupyter Notebook 来编写 Python 程序时的文件。

Jupyter Notebook（此前被称为 IPython notebook）是一个交互式笔记本，支持运行 40 多种编程语言。它会在浏览器中打开并运行相关程序，在这里，我们主要介绍其在编写和运行 Python 程序方面的应用。

在安装好 Anaconda 后，已经自动安装好了 Jupyter Notebook，如下所示（红色数字4所指）：

优势在于

数据分析和绘图



什么是网络数据采集-前言

网络爬虫 [编辑]

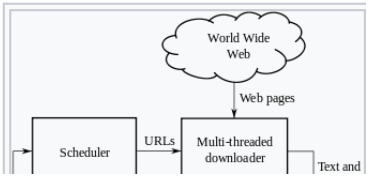
维基百科，自由的百科全书

本条目存在以下问题，请协助**改善本条目**或在**讨论页**针对议题发表看法。 [折叠]

- 本条目**需要扩充。** *(2015年3月15日)*
  请协助改善这篇条目，更进一步的信息可能会在讨论页或扩充请求中找到。请在扩充条目后将此模板移除。
- 本条目需要**精通或熟悉相关主题的编者**参与及协助编辑。 *(2015年3月15日)*
  请邀请适合的人士改善本条目。更多的细节与详情请参见讨论页。
- 本条目**需要补充更多来源。** *(2015年3月15日)*
  请协助添加多方面可靠来源以改善这篇条目，无法查证的内容可能会因为异议提出而移除。
- 本条目**可参照英语维基百科相应条目来扩充。**
  若您熟悉来源语言和主题，请协助参考外语维基扩充条目。请勿直接提交机械翻译，也不要翻译不可靠、低品质内容。
  依版权协议，译文需在编辑摘要注明来源，或于讨论页顶部标记 `{{Translated page}}` 标签。

网络爬虫（英语：web crawler），也叫网络蜘蛛（spider），是一种用来自动浏览万维网的网络机器人。其目的一般为编纂网络索引。

网络搜索引擎等站点通过爬虫软件更新自身的网站内容或其对其他网站的索引。网络爬虫可以将自己所访问的页面保存下来，以便搜索引擎事后生成索引供用户搜索。



网络爬虫用英文怎么说?

我来答　分享　举报　浏览 5440 次

3个回答

#活动# 参与《复联4》问答讨论，赢免费影票!

今天我好高兴3

来自科学教育类芝麻团　2017-07-12

英文是： Internet worm。

[例句]分布式多主题网络爬虫系统的研究与实现

Research and Implementation of Distributed and Multi-topic Web Crawler System

词汇解释:

internet

n. 互联网;

abbr. interconnection network, internetwork 互联网; internetwork 互联网网;

展开全部

2　43　评论　分享　举报

2000年1月1日，李彦宏在北京市中关村创建了百度在线网络技术（北京）有限公司（Baidu Online Network Technology (Beijing) Co.,Ltd.，简称“百度在线”）。[6]

2000年1月18日，李彦宏在英属开曼群岛注册了上市主体百度公司（Baidu.com,Inc），又在英属维尔京群岛注册了一个百度公司的全资子公司——百度控股有限公司（Baidu Holdings Limited）。

网站	ww
Alexa排名	4

2001年6月5日，李彦宏和徐勇在中国大陆注册了一个公司——北京百度网讯科技有限公司（Baidu Netcom Science and Technology (Beijing) Co.,Ltd.），注册这样的公司是为了规避中国政府的关于外资不能进入新闻广告等领域的法规，而之前的百度在线也成为了百度控股有限公司的全资子公司。

百度是美国公司，联想不是中国公司？

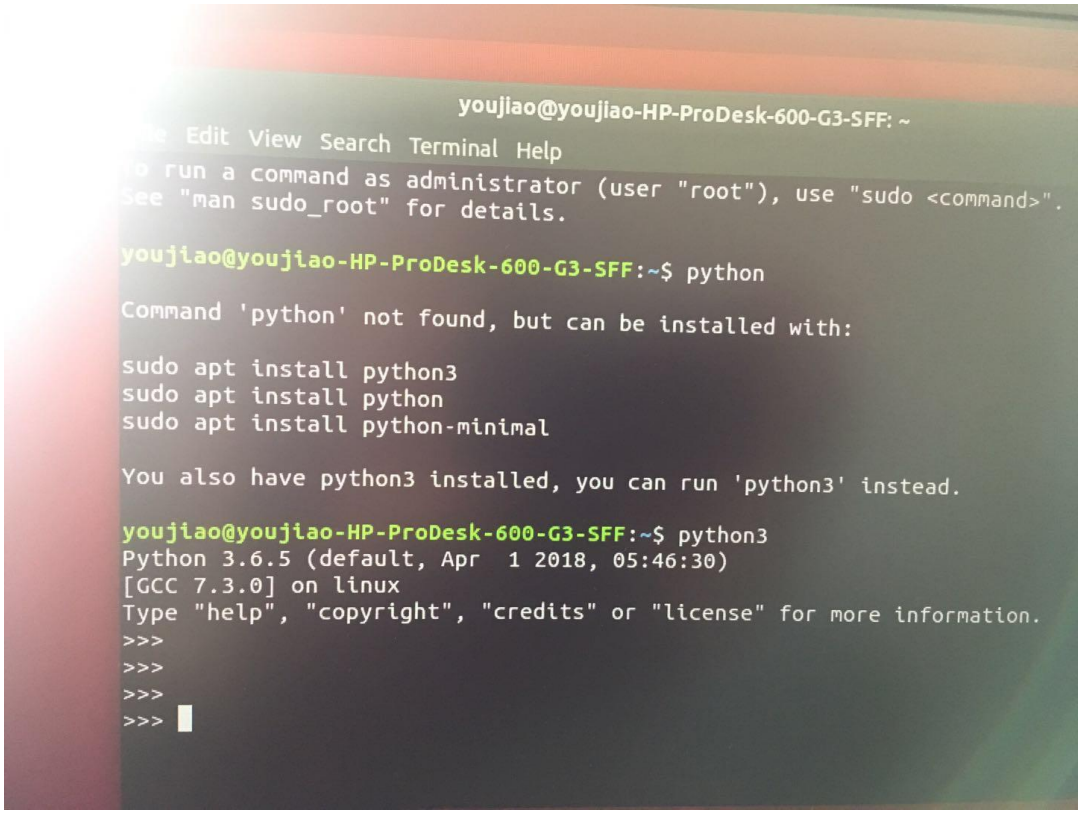
2018-09-18 23:47

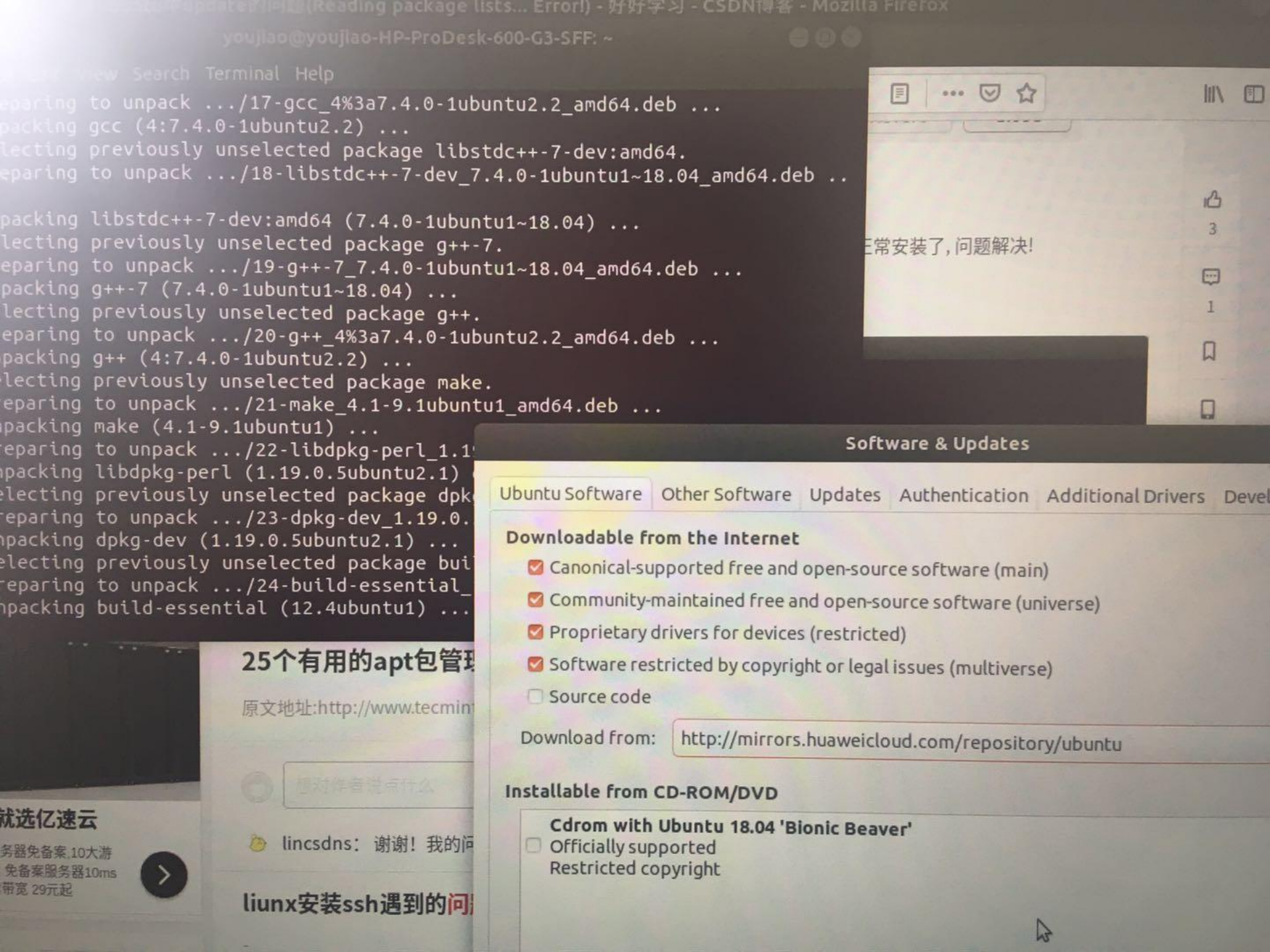
在9月16-17日的“2018中国发展高层论坛专题研讨会”上，IDG资本全球董事

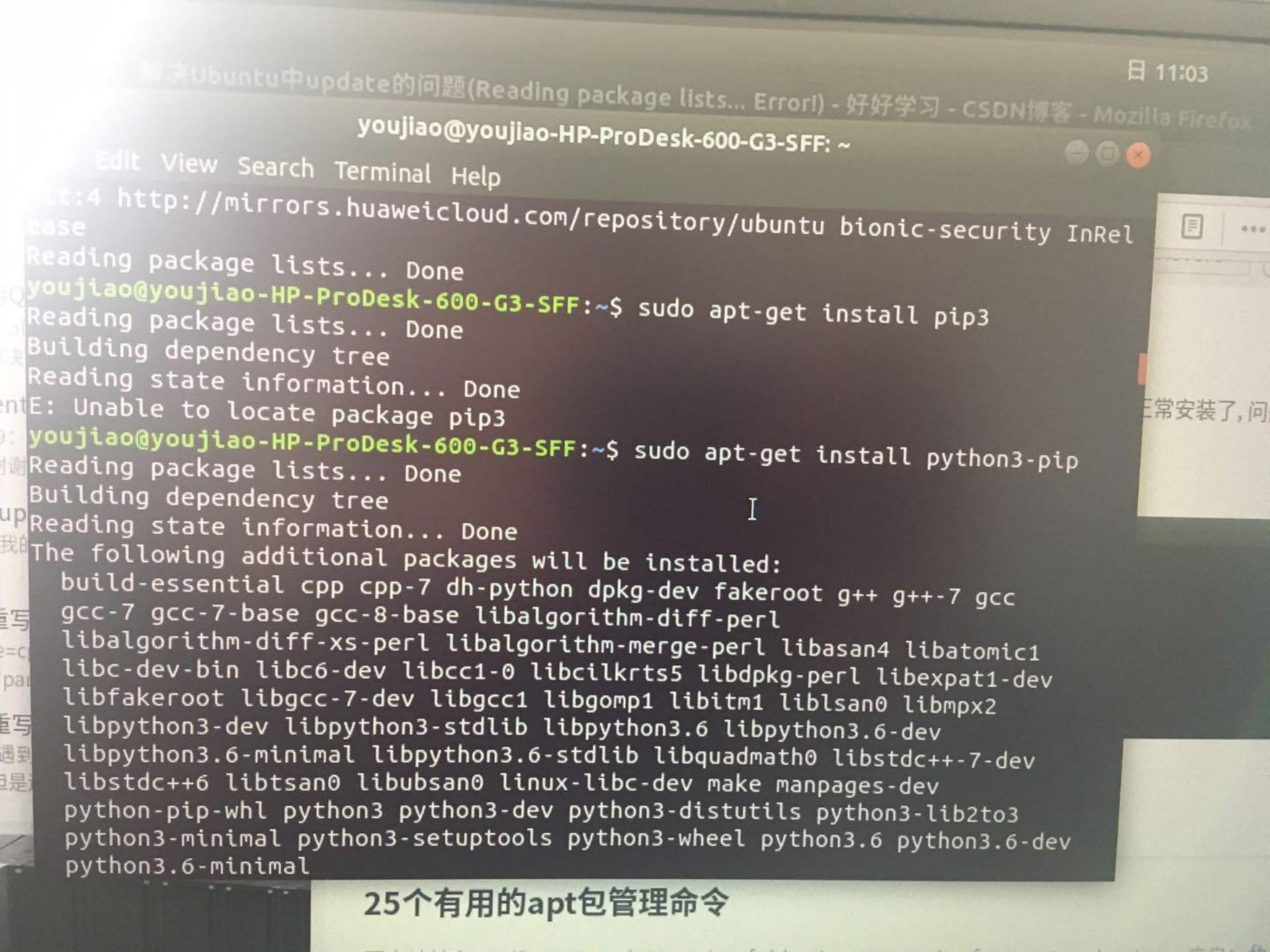
人：“像百度在美国上市，就是美国的公司。”

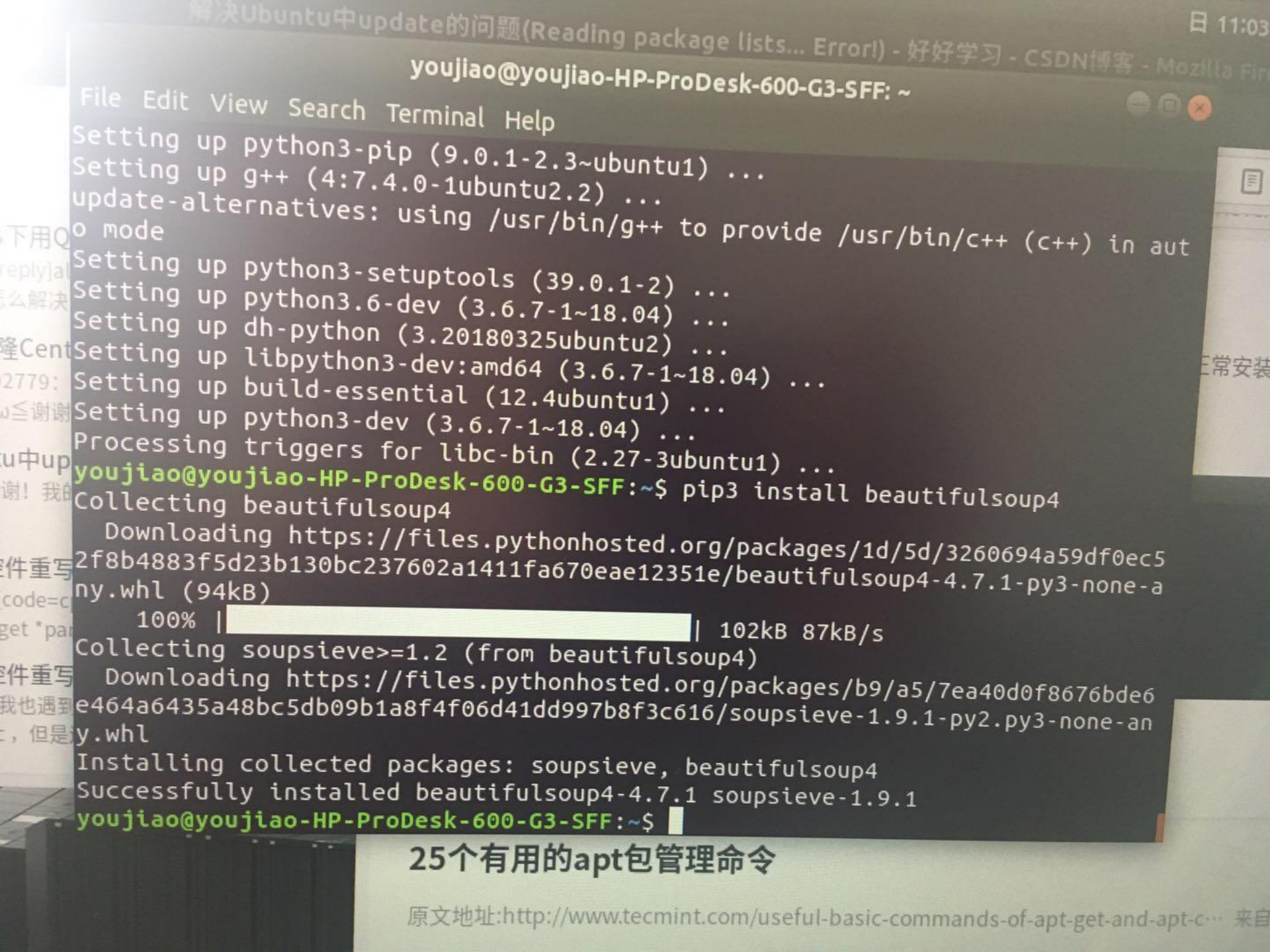
版本的选择

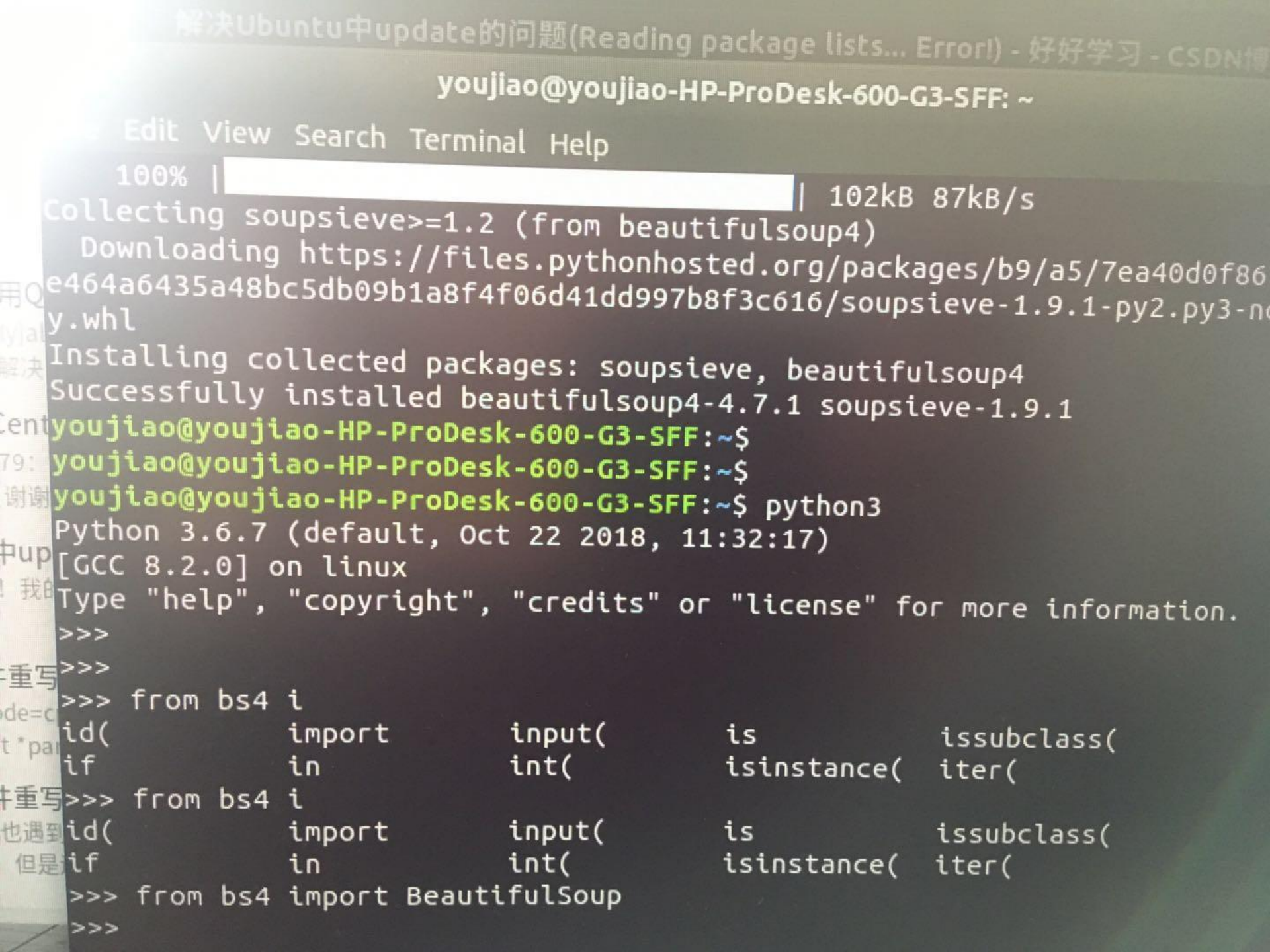
- 关于本书-附录A



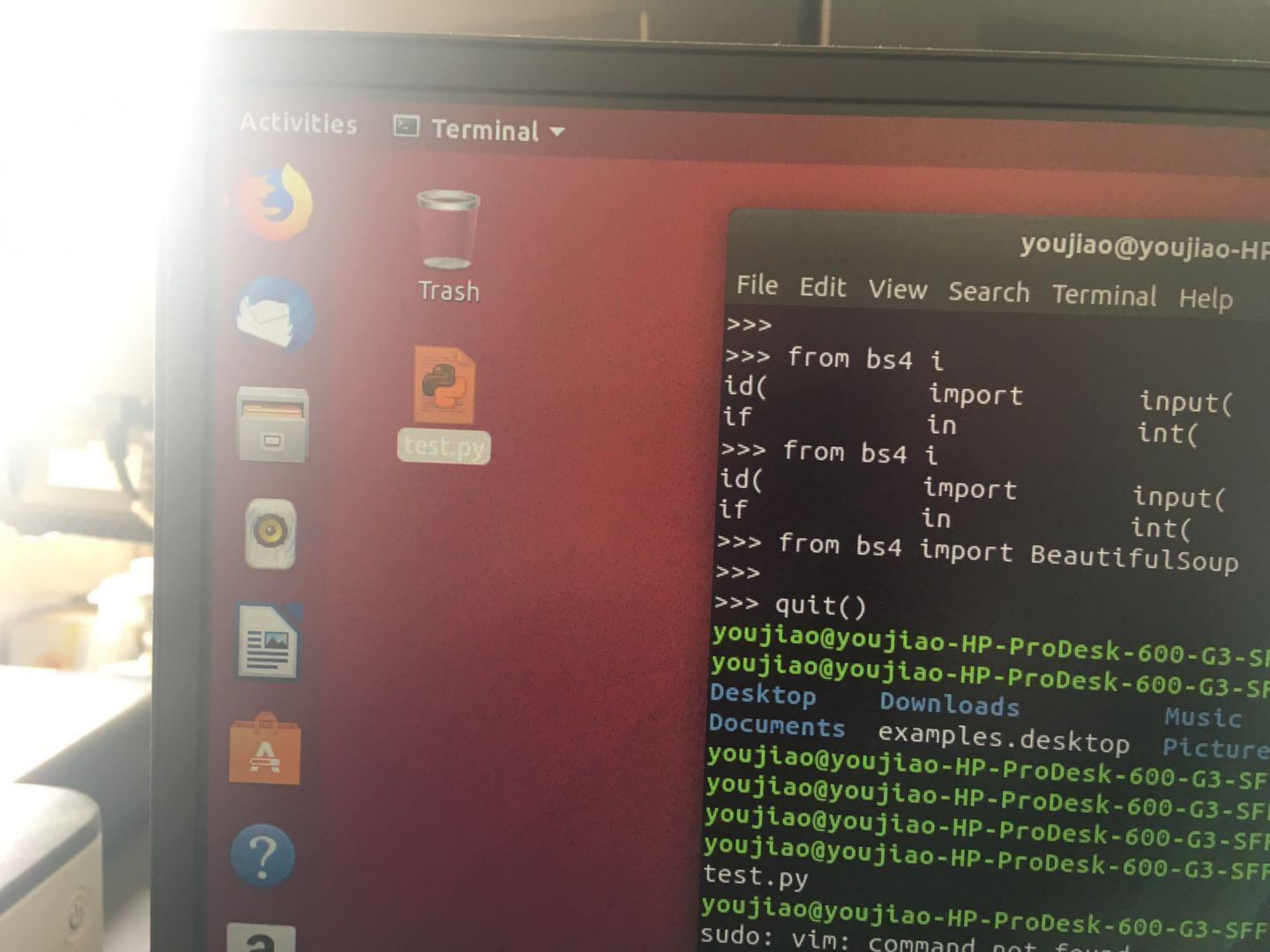

youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~
Search Terminal Help
eparing to unpack .../17-gcc_4%3a7.4.0-1ubuntu2.2_amd64.deb ...
packing gcc (4:7.4.0-1ubuntu2.2) ...
lecting previously unselected package libstdc++-7-dev:amd64.
eparing to unpack .../18-libstdc++-7-dev_7.4.0-1ubuntu1~18.04_amd64.deb ..
packing libstdc++-7-dev:amd64 (7.4.0-1ubuntu1~18.04) ...
lecting previously unselected package g++-7.
eparing to unpack .../19-g++-7_7.4.0-1ubuntu1~18.04_amd64.deb ...
packing g++-7 (7.4.0-1ubuntu1~18.04) ...
lecting previously unselected package g++.
eparing to unpack .../20-g++_4%3a7.4.0-1ubuntu2.2_amd64.deb ...
packing g++ (4:7.4.0-1ubuntu2.2) ...
lecting previously unselected package make.
eparing to unpack .../21-make_4.1-9.1ubuntu1_amd64.deb ...
packing make (4.1-9.1ubuntu1) ...
eparing to unpack .../22-libdpkg-perl_1.1
packing libdpkg-perl (1.19.0.5ubuntu2.1)
lecting previously unselected package dpk
eparing to unpack .../23-dpkg-dev_1.19.0.
packing dpkg-dev (1.19.0.5ubuntu2.1) ...
lecting previously unselected package bui
eparing to unpack .../24-build-essential_
packing build-essential (12.4ubuntu1) ...
正常安装了，问题解决!
3
1
Software & Updates
Ubuntu Software
Other Software
Updates
Authentication
Additional Drivers
Downloadable from the Internet
Canonical-supported free and open-source software (main)
Community-maintained free and open-source software (universe)
Proprietary drivers for devices (restricted)
Software restricted by copyright or legal issues (multiverse)
Source code
Download from:
http://mirrors.huaweicloud.com/repository/ubuntu
Installable from CD-ROM/DVD
Cdrom with Ubuntu 18.04 'Bionic Beaver'
Officially supported
Restricted copyright
25个有用的apt包管理
原文地址:http://www.tecmint
lincsdns: 谢谢！我的问
liunx安装ssh遇到的问
就选亿速云
务器免备案,10大游
免备案服务器10ms
带宽 29元起

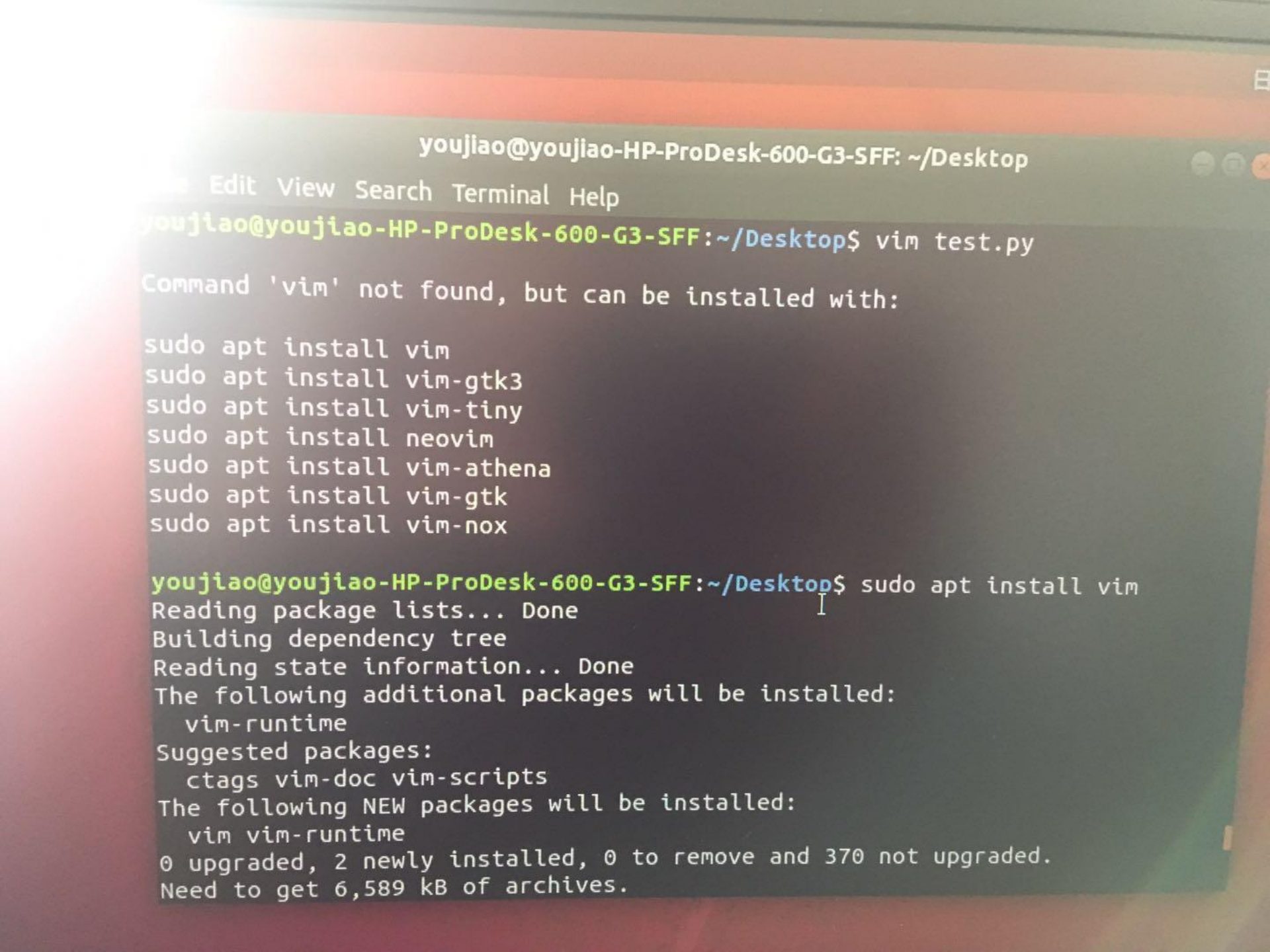

youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~
Edit View Search Terminal Help
:4 http://mirrors.huaweicloud.com/repository/ubuntu bionic-security InRel
ase
Reading package lists... Done
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ sudo apt-get install pip3
Reading package lists... Done
Building dependency tree
Reading state information... Done
E: Unable to locate package pip3
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ sudo apt-get install python3-pip
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
build-essential cpp cpp-7 dh-python dpkg-dev fakeroot g++ g++-7 gcc
gcc-7 gcc-7-base gcc-8-base libalgorithm-diff-perl
libalgorithm-diff-xs-perl libalgorithm-merge-perl libasan4 libatomic1
libc-dev-bin libc6-dev libcc1-0 libcilkrts5 libdpkg-perl libexpat1-dev
libfakeroot libgcc-7-dev libgcc1 libgomp1 libitm1 liblsan0 libmpx2
libpython3-dev libpython3-stdlib libpython3.6 libpython3.6-dev
libpython3.6-minimal libpython3.6-stdlib libquadmath0 libstdc++-7-dev
libstdc++6 libtsan0 libubsan0 linux-libc-dev make manpages-dev
python-pip-whl python3 python3-dev python3-distutils python3-lib2to3
python3-minimal python3-setuptools python3-wheel python3.6 python3.6-dev
python3.6-minimal
25个有用的apt包管理命令

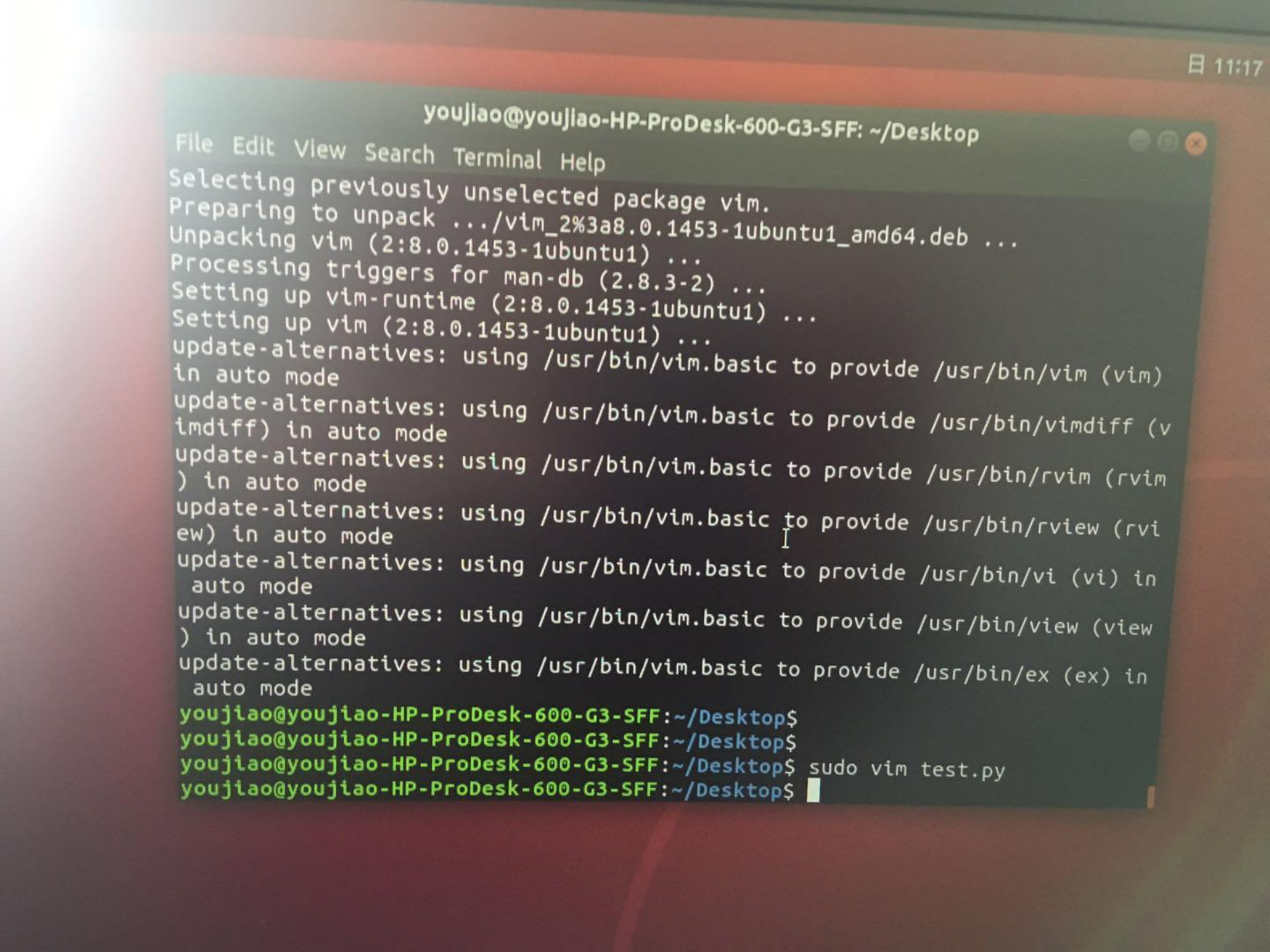

youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~
File Edit View Search Terminal Help
Setting up python3-pip (9.0.1-2.3~ubuntu1) ...
Setting up g++ (4:7.4.0-1ubuntu2.2) ...
update-alternatives: using /usr/bin/g++ to provide /usr/bin/c++ (c++) in aut
o mode
Setting up python3-setuptools (39.0.1-2) ...
Setting up python3.6-dev (3.6.7-1~18.04) ...
Setting up dh-python (3.20180325ubuntu2) ...
Setting up libpython3-dev:amd64 (3.6.7-1~18.04) ...
Setting up build-essential (12.4ubuntu1) ...
Setting up python3-dev (3.6.7-1~18.04) ...
Processing triggers for libc-bin (2.27-3ubuntu1) ...
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ pip3 install beautifulsoup4
Collecting beautifulsoup4
Downloading https://files.pythonhosted.org/packages/1d/5d/3260694a59df0ec5
2f8b4883f5d23b130bc237602a1411fa670eae12351e/beautifulsoup4-4.7.1-py3-none-a
ny.whl (94kB)
100% | | 102kB 87kB/s
Collecting soupsieve>=1.2 (from beautifulsoup4)
Downloading https://files.pythonhosted.org/packages/b9/a5/7ea40d0f8676bde6
e464a6435a48bc5db09b1a8f4f06d41dd997b8f3c616/soupsieve-1.9.1-py2.py3-none-an
y.whl
Installing collected packages: soupsieve, beautifulsoup4
Successfully installed beautifulsoup4-4.7.1 soupsieve-1.9.1
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$
25个有用的apt包管理命令
原文地址:http://www.tecmint.com/useful-basic-commands-of-apt-get-and-apt-c…

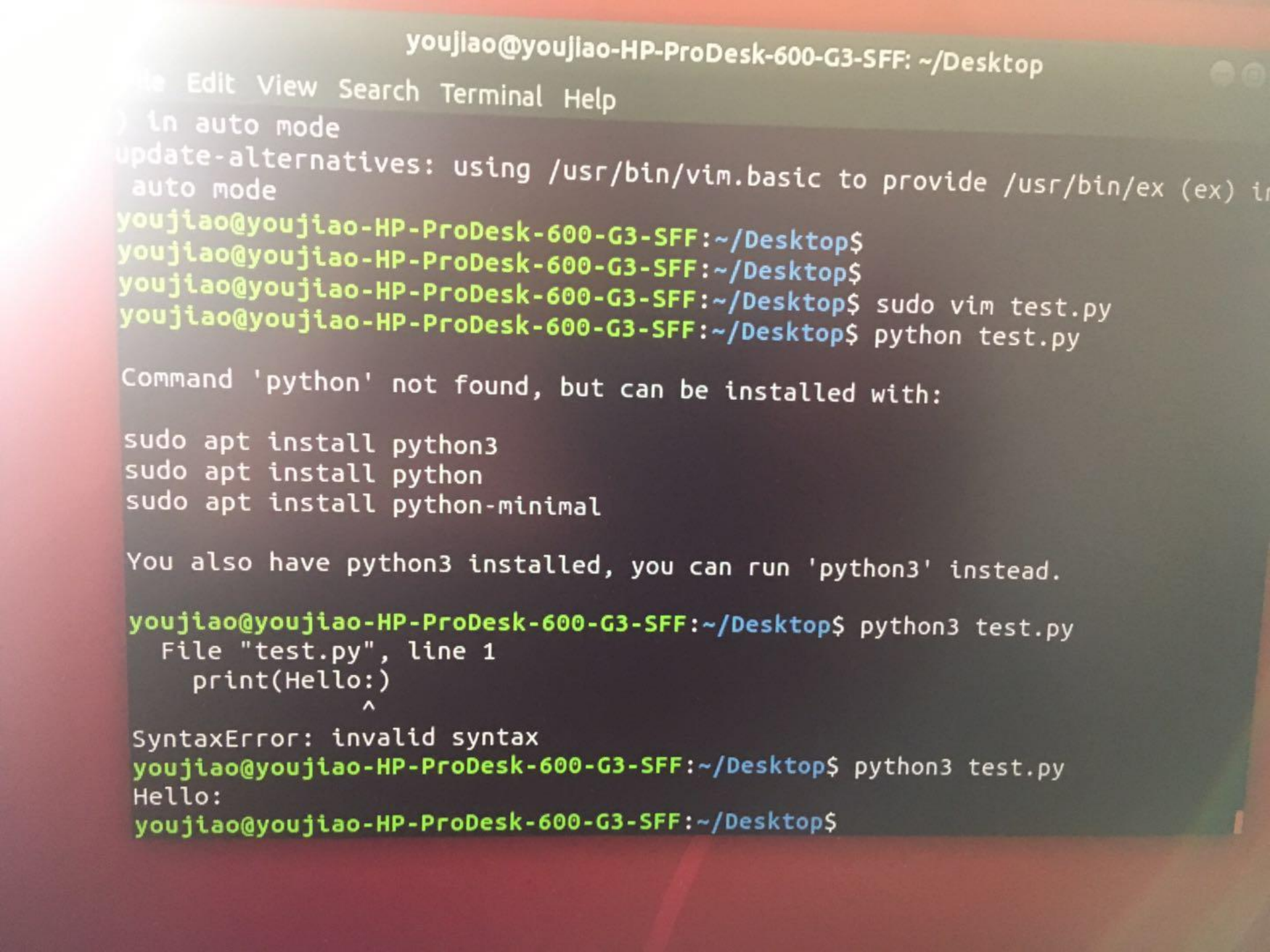

解决Ubuntu中update的问题(Reading package lists... Error!) - 好好学习 - CSDN博
youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~
Edit View Search Terminal Help
100% | | 102kB 87kB/s
Collecting soupsieve>=1.2 (from beautifulsoup4)
Downloading https://files.pythonhosted.org/packages/b9/a5/7ea40d0f86
e464a6435a48bc5db09b1a8f4f06d41dd997b8f3c616/soupsieve-1.9.1-py2.py3-n
y.whl
Installing collected packages: soupsieve, beautifulsoup4
Successfully installed beautifulsoup4-4.7.1 soupsieve-1.9.1
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ python3
Python 3.6.7 (default, Oct 22 2018, 11:32:17)
[GCC 8.2.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
>>>
>>> from bs4 i
id(import input(is issubclass(
if in int(isinstance(iter(
>>> from bs4 i
id(import input(is issubclass(
if in int(isinstance(iter(
>>> from bs4 import BeautifulSoup
>>>

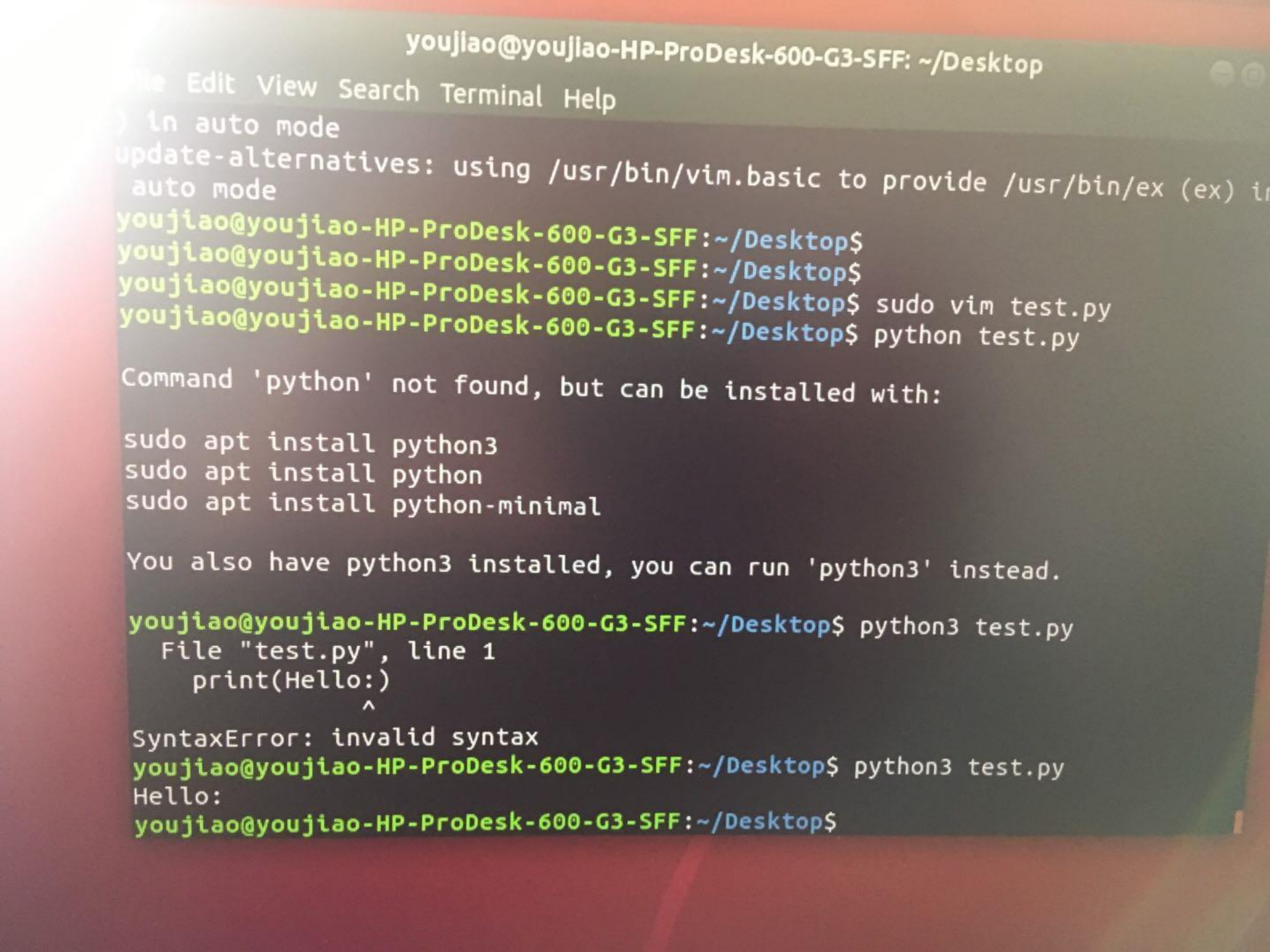
```
>>> from bs4 i
id(             import          input(          is              issubclass(
if              in              int(            isinstance(     iter(
>>> from bs4 i
id(             import          input(          is              issubclass(
if              in              int(            isinstance(     iter(
>>> from bs4 import BeautifulSoup
>>>
>>> quit()
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ ls
Desktop    Downloads          Music     Public     Videos
Documents  examples.desktop   Pictures  Templates
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~$ cd Desktop/
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ touch test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ ls
test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ sudo vim test.py
sudo: vim: command not found
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ vim test.py

Command 'vim' not found, but can be installed with:
```


Activities
Terminal
Trash
test.py
youjiao@youjiao-HP
File Edit View Search Terminal Help
>>>
>>> from bs4 i
id(import input(
if in int(
>>> from bs4 i
id(import input(
if in int(
>>> from bs4 import BeautifulSoup
>>>
>>> quit()
youjiao@youjiao-HP-ProDesk-600-G3-SF
youjiao@youjiao-HP-ProDesk-600-G3-SF
Desktop Downloads Music
Documents examples.desktop Picture
youjiao@youjiao-HP-ProDesk-600-G3-SF
youjiao@youjiao-HP-ProDesk-600-G3-SF
youjiao@youjiao-HP-ProDesk-600-G3-SF
youjiao@youjiao-HP-ProDesk-600-G3-SF
test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF
sudo: vim: command not found


youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~/Desktop
Edit View Search Terminal Help
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ vim test.py
Command 'vim' not found, but can be installed with:
sudo apt install vim
sudo apt install vim-gtk3
sudo apt install vim-tiny
sudo apt install neovim
sudo apt install vim-athena
sudo apt install vim-gtk
sudo apt install vim-nox
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ sudo apt install vim
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
vim-runtime
Suggested packages:
ctags vim-doc vim-scripts
The following NEW packages will be installed:
vim vim-runtime
0 upgraded, 2 newly installed, 0 to remove and 370 not upgraded.
Need to get 6,589 kB of archives.


youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~/Desktop
File Edit View Search Terminal Help
Selecting previously unselected package vim.
Preparing to unpack .../vim_2%3a8.0.1453-1ubuntu1_amd64.deb ...
Unpacking vim (2:8.0.1453-1ubuntu1) ...
Processing triggers for man-db (2.8.3-2) ...
Setting up vim-runtime (2:8.0.1453-1ubuntu1) ...
Setting up vim (2:8.0.1453-1ubuntu1) ...
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/vim (vim) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/vimdiff (vimdiff) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/rvim (rvim) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/rview (rview) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/vi (vi) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/view (view) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/ex (ex) in auto mode
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ sudo vim test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$


youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~/Desktop
File Edit View Search Terminal Help
) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/ex (ex) i
auto mode
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ sudo vim test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python test.py
Command 'python' not found, but can be installed with:
sudo apt install python3
sudo apt install python
sudo apt install python-minimal
You also have python3 installed, you can run 'python3' instead.
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python3 test.py
File "test.py", line 1
print(Hello:)
^
SyntaxError: invalid syntax
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python3 test.py
Hello:
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$

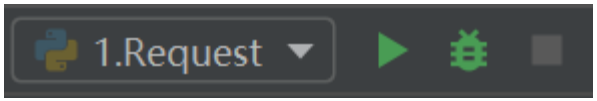

youjiao@youjiao-HP-ProDesk-600-G3-SFF: ~/Desktop
File Edit View Search Terminal Help
) in auto mode
update-alternatives: using /usr/bin/vim.basic to provide /usr/bin/ex (ex) i
auto mode
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ sudo vim test.py
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python test.py
Command 'python' not found, but can be installed with:
sudo apt install python3
sudo apt install python
sudo apt install python-minimal
You also have python3 installed, you can run 'python3' instead.
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python3 test.py
File "test.py", line 1
print(Hello:)
^
SyntaxError: invalid syntax
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$ python3 test.py
Hello:
youjiao@youjiao-HP-ProDesk-600-G3-SFF:~/Desktop$

第一章 初见网络爬虫

- Windows下安装BeautifulSoup
- Pycharm设置
  - https://zhuanlan.zhihu.com/p/26066151
- Prettify() 方法将Beautiful Soup的文档树格式化后以Unicode编码输出,每个XMLHTML标签都独占一行

Python爬虫

- 进行必要设置（添加库）
  - File →Settting →Project: XXX →Project Interpreter
  - Import requests
  - From bs4 import BeautifulSoup
- 新建Python文档
  - 右击工程名→ New → Python File
- 运行
  - 注意多个源文件之间的切换、运行、调试



- Chrome基础操作与设置

Request和BeautifulSoup库

- 参考1.Requests.py
- 200说明联网成功
- 有时爬虫需要加入请求头来伪装成浏览器，增强稳定性，以便更好地抓取数据
- Chrome://version
- 参考2.BeautifulSoup4.py，观察输出结果
- 防止出现乱码：
  - 利用urllib处理HTTP协议
  - BeautifulSoup4解析HTML文档

定位需要的信息

- 准确定位：
  - copy selector，参考3.Select.py
- 模糊定位：
  - 参考4.SelectTitle.py & 4.1.SelectTitle.py
- 综合案例1：爬取相声列表（5.xiangsheng.py）
  - https://www.ximalaya.com/xiangsheng/2667276/
- 综合案例2：爬取跨网页图像（6.Photo.py）