**Environment Requirements:**

1. Linux Server – In this guide, we are using Amazon EC2 Server (Ubuntu 16.04)
2. Python 3.5 – Build-in with Ubuntu 16.04
3. Required Software: Please refer to the step 3.

**Important notes:**

* + This guide will not guide you on how to configure Amazon EC2 Instance.
  + We are using Amazon EC2 Instance to host this project. If you can follow this guide with Amazon EC2 Instance, then it should be working. However, if you changed the settings (like: location of the folders) then please modified / update the paths accordingly.

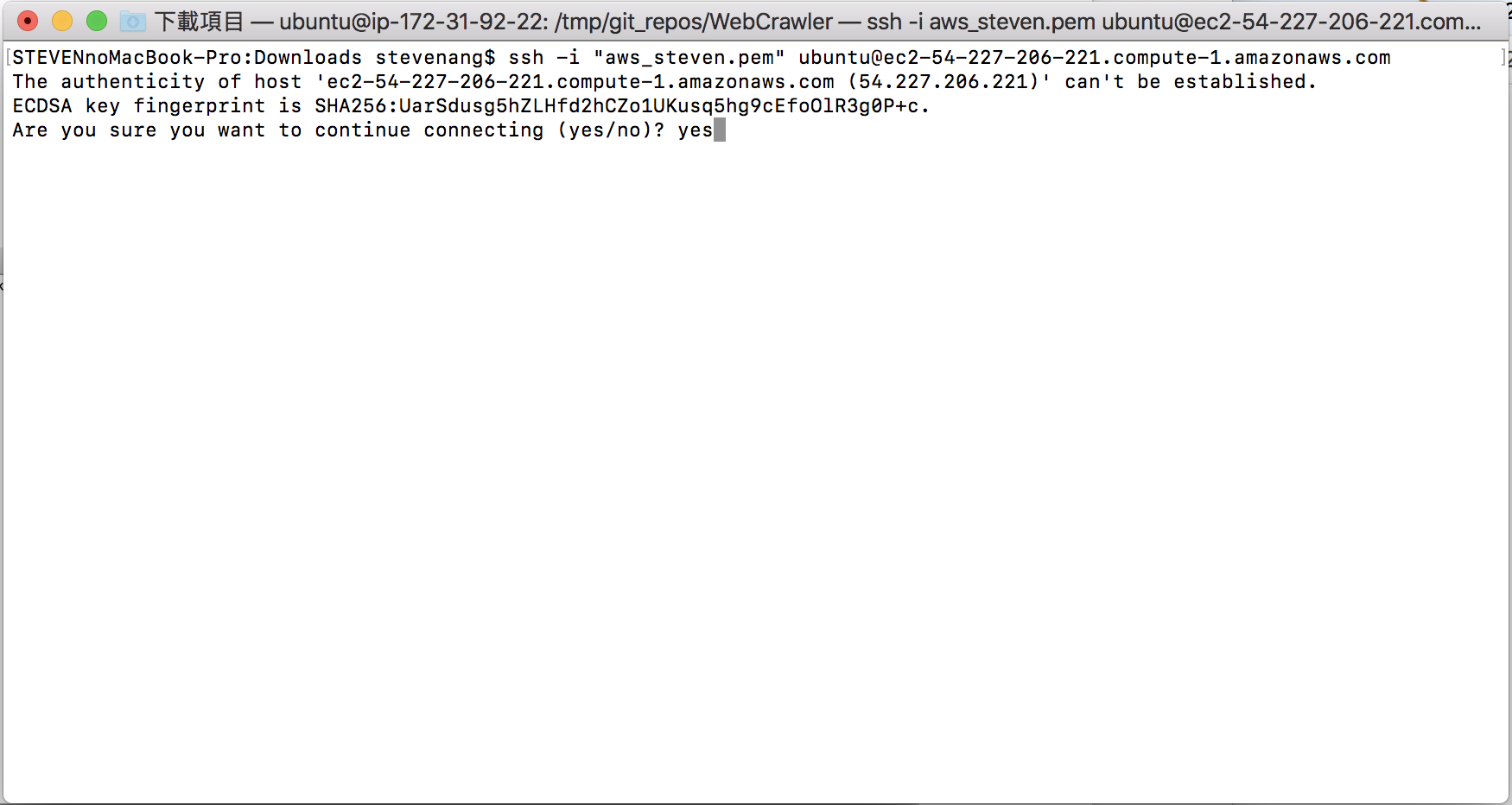
**The outcome:**

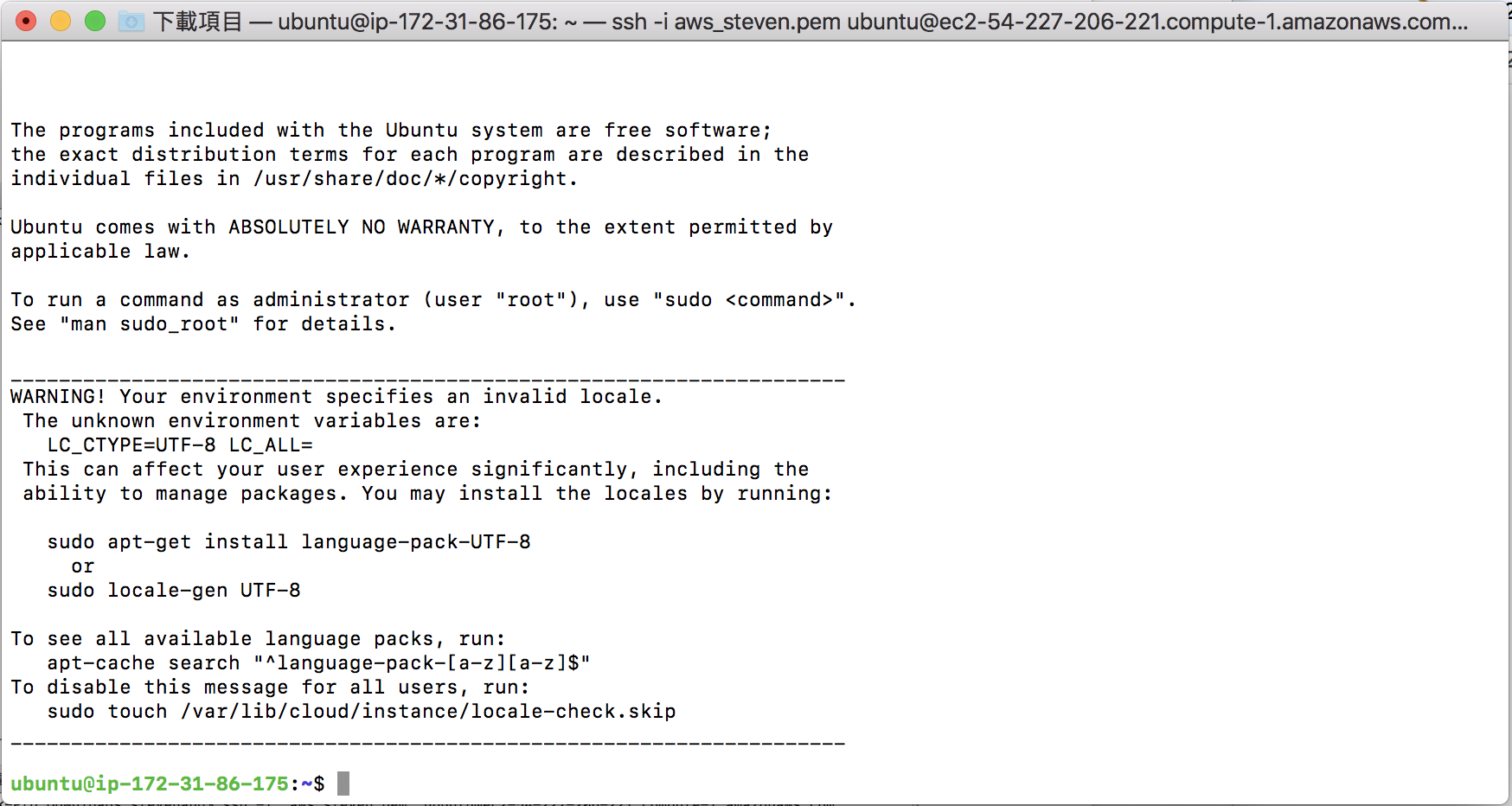
1. You will have a simple Web Crawler (Spider) program with can crawled the link (top page or home page of a web site) you specified and return all the links on that website.
2. You will have a web application running using django as its development language
3. You will have a program which will be running automatically every Monday to Friday at 11 AM to collect foreign currency’s data from the bank specified in this program.

Setup Guide:

1. Connect to your Amazon EC2 Instance

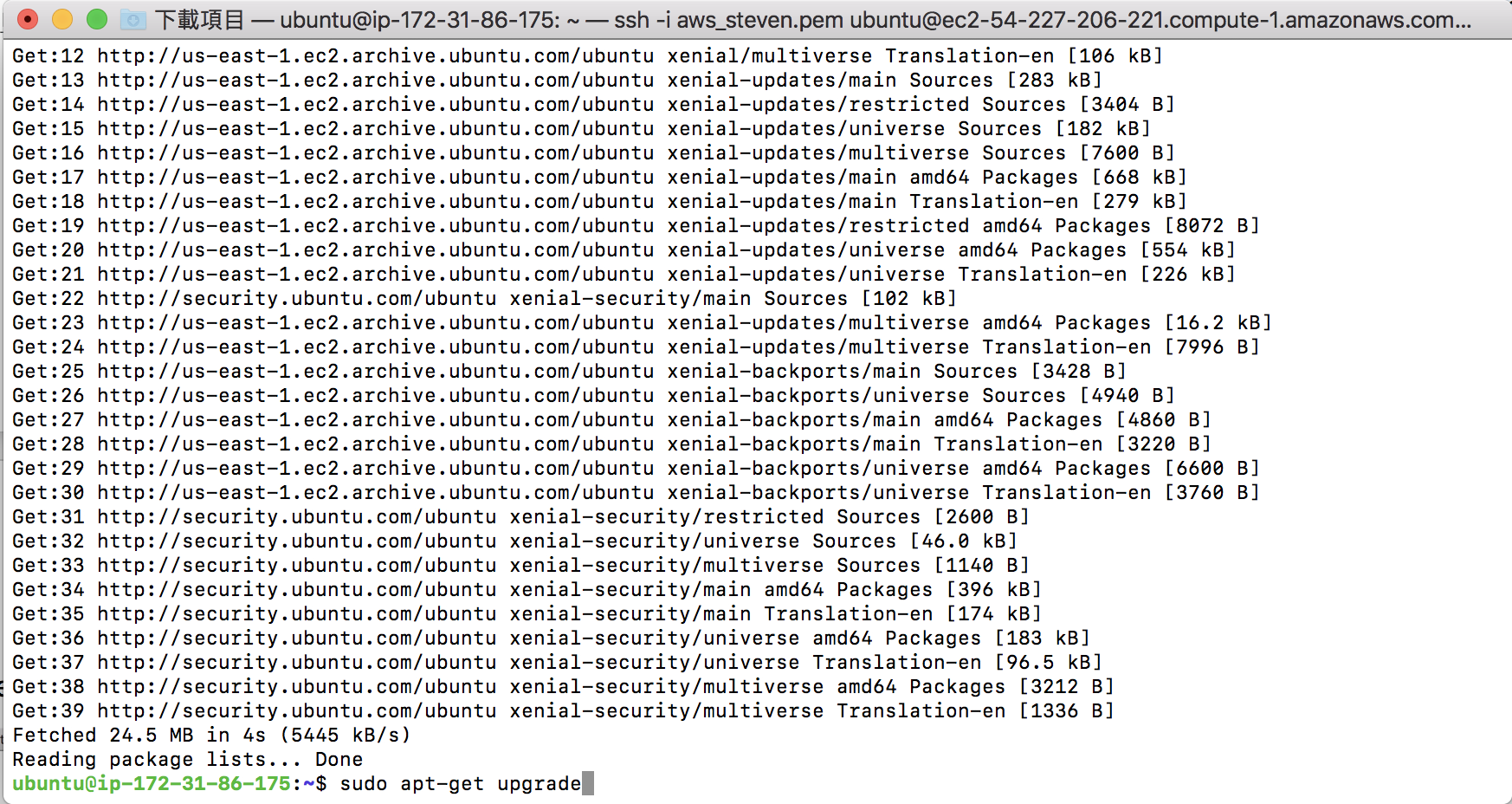


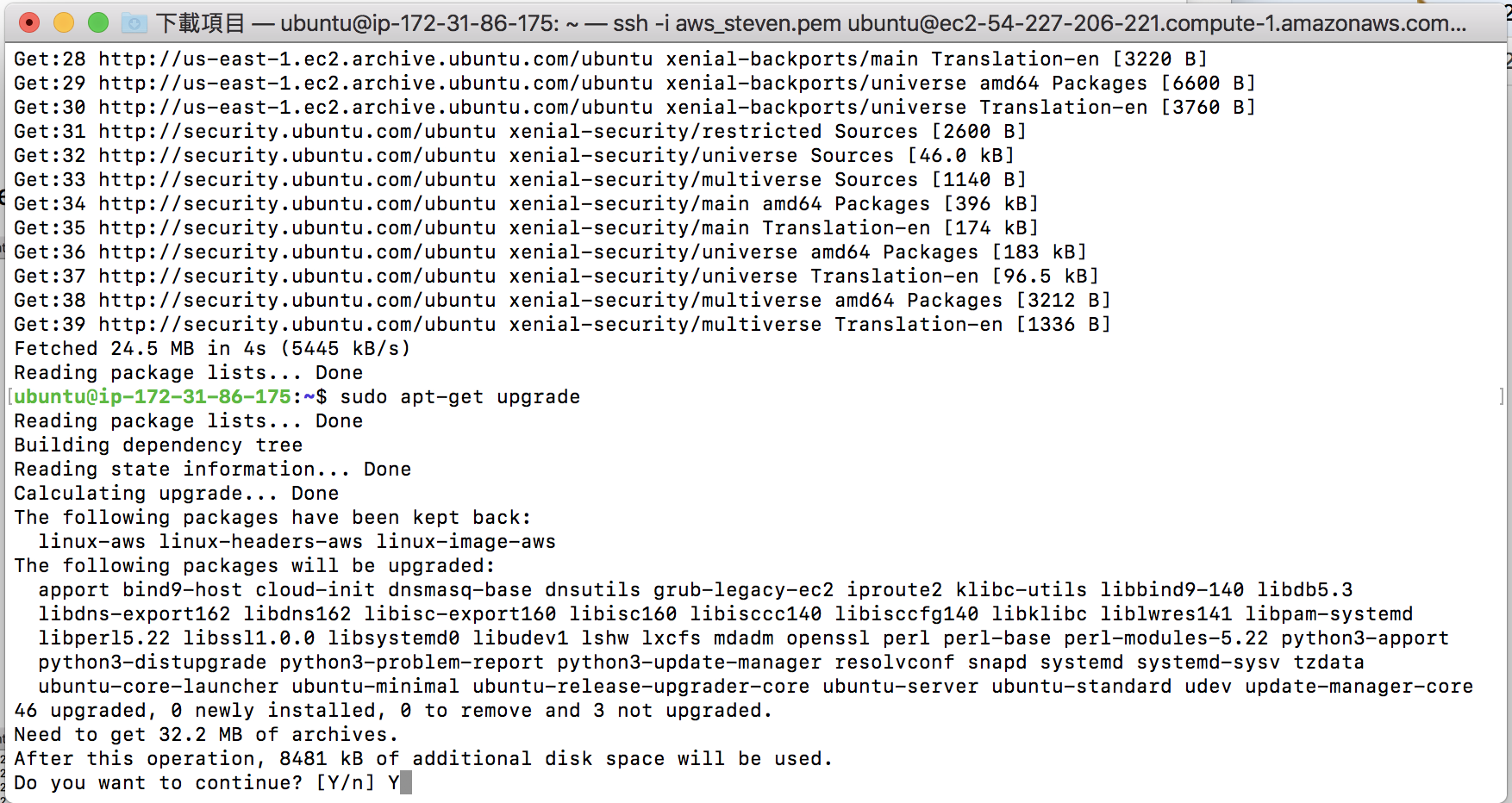


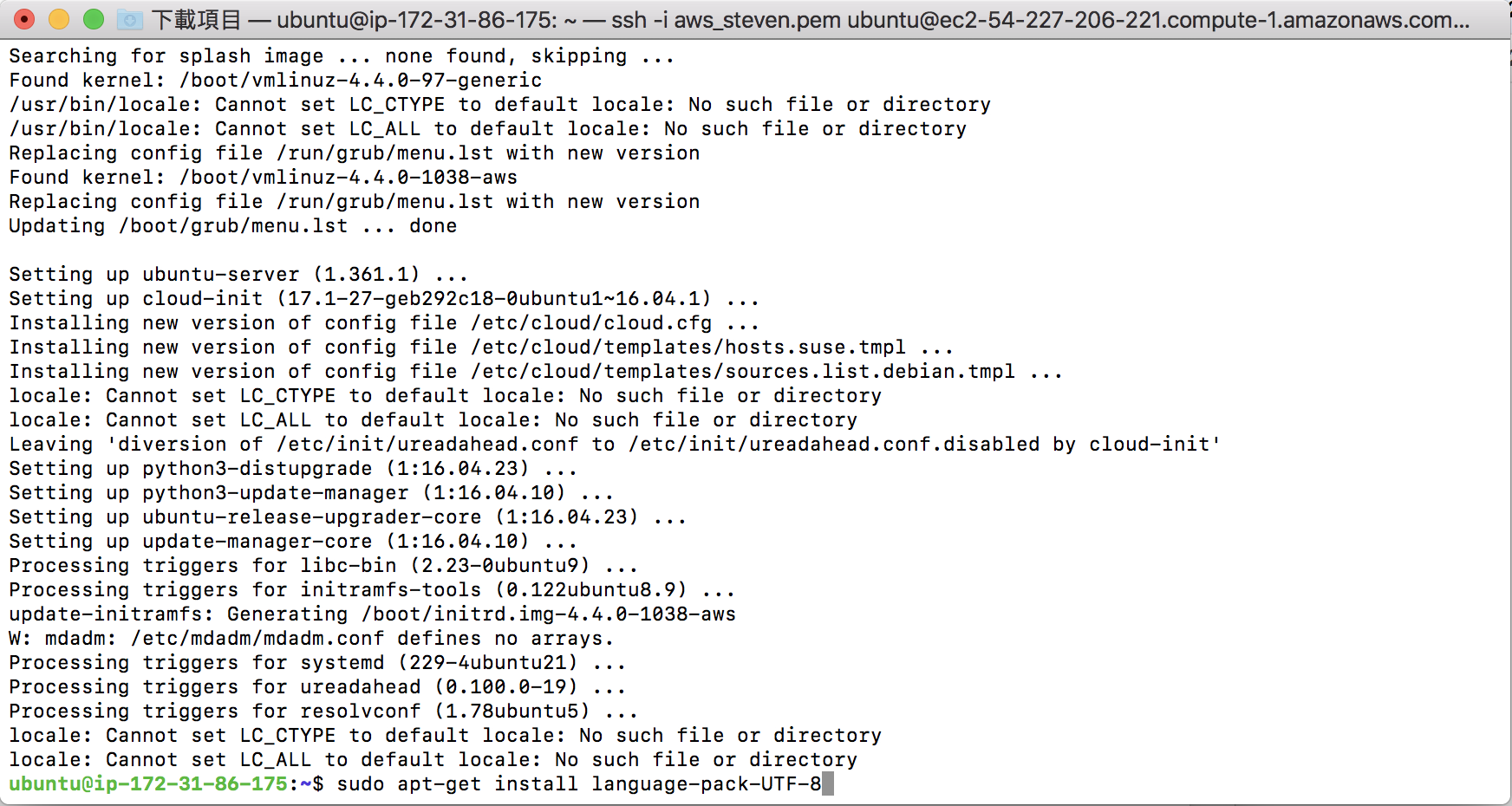


1. Execute “sudo apt-get update” and “sudo apt-get upgrade” to update your server

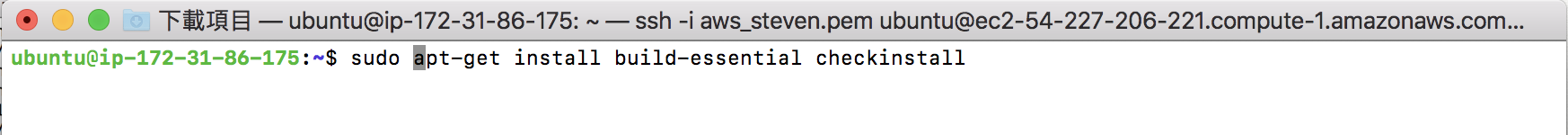








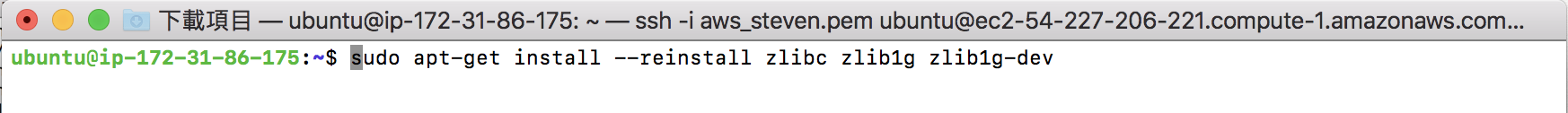
1. Install necessary packages
   1. sudo apt-get install build-essential checkinstall



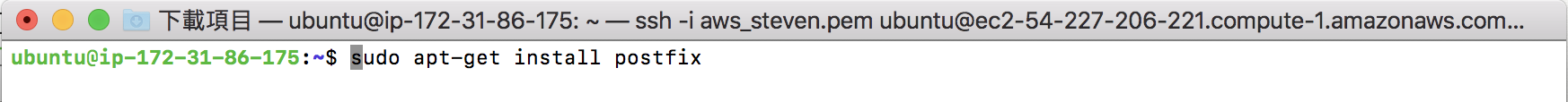
* 1. sudo apt-get install libreadline-gplv2-dev libncursesw5-dev libssl-dev libsqlite3-dev tk-dev libgdbm-dev libc6-dev libbz2-dev



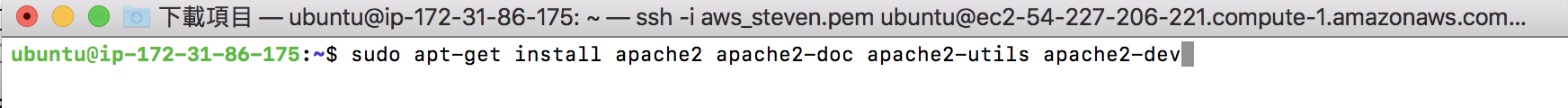
* 1. sudo apt-get install --reinstall zlibc zlib1g zlib1g-dev



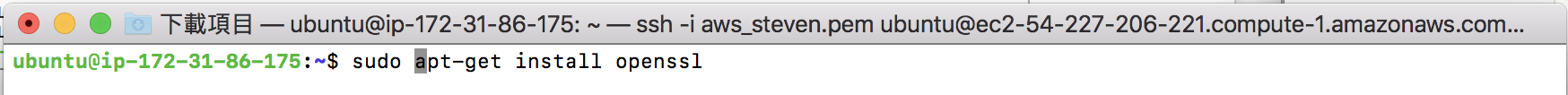
* 1. sudo apt-get install postfix



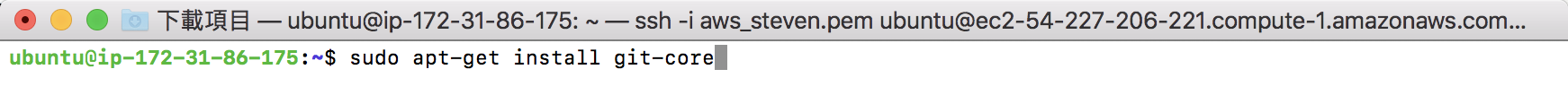
* 1. sudo apt-get install apache2 apache2-doc apache2-utils apache2-dev



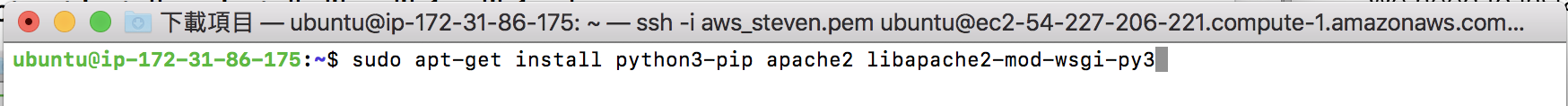
* 1. sudo apt-get install openssl



* 1. sudo apt-get install git-core

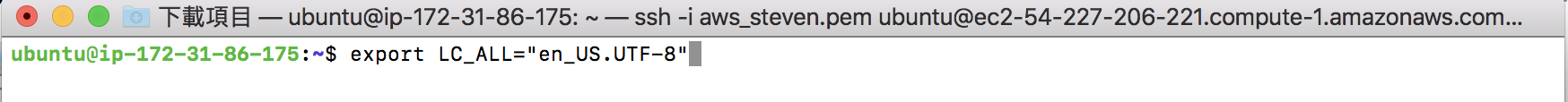


* 1. sudo apt-get install python3-pip apache2 libapache2-mod-wsgi-py3

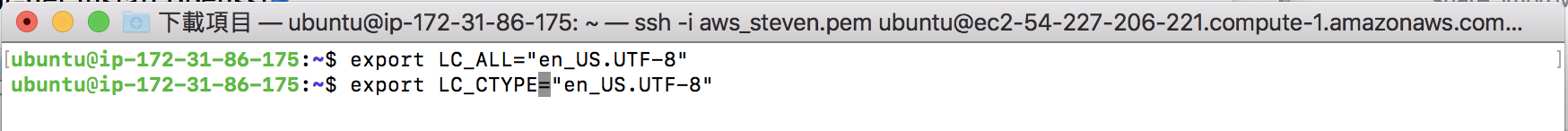


* 1. Execute the following commands

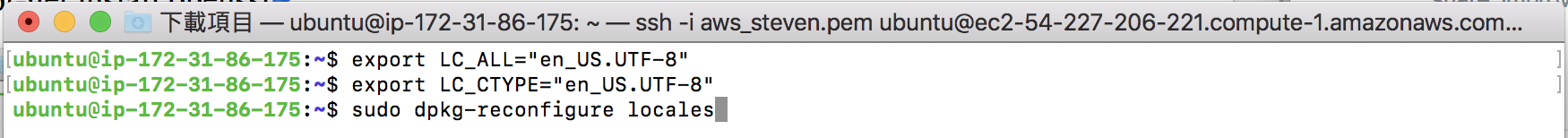
export LC\_ALL=”en\_US.UTF-8”



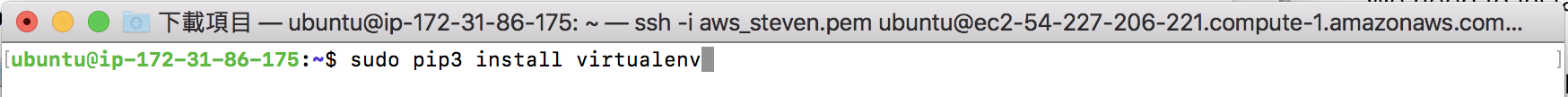
export LC\_CTYPE=”en\_US.UTF-8”



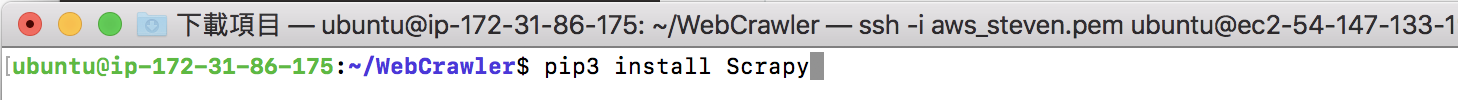
sudo dpkg-reconfigure locales



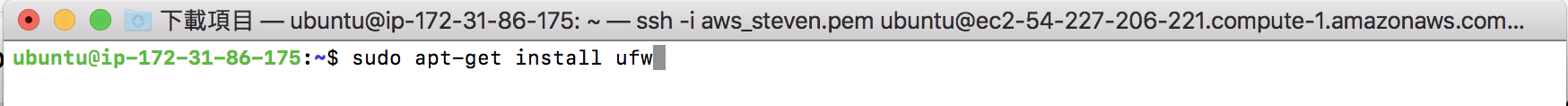
* 1. sudo pip3 install virtualenv



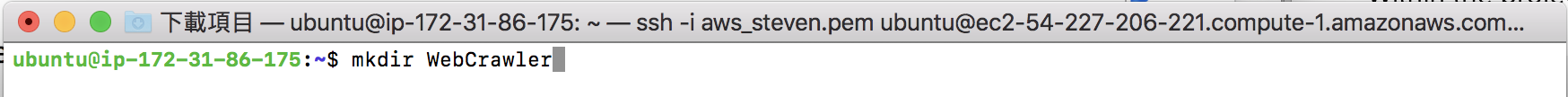
* 1. sudo pip3 install Scrapy html5lib pandas beautifulsoup4 lxml plotly



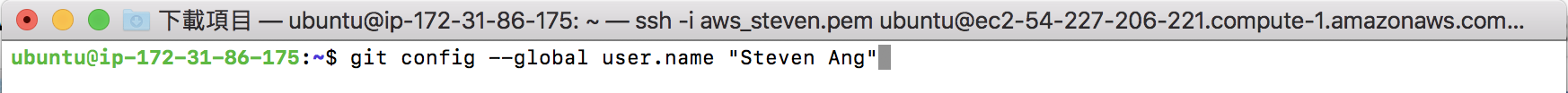
* 1. sudo apt-get install ufw



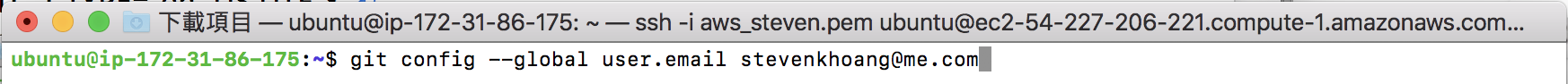
1. Download Project Source Code
   1. mkdir WebCrawler



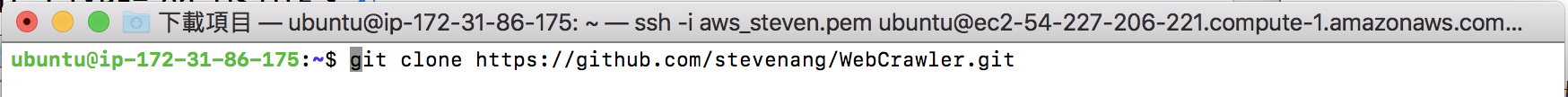
* 1. Setup Git
     1. git config –global user.name “Your Name”



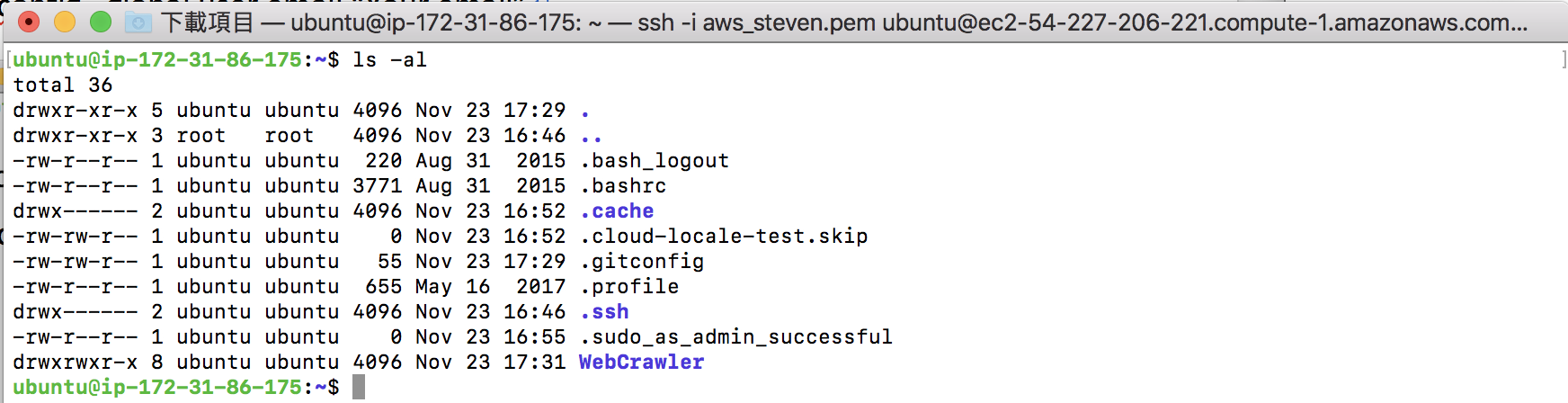
* + 1. git config –global user.email “Your email”

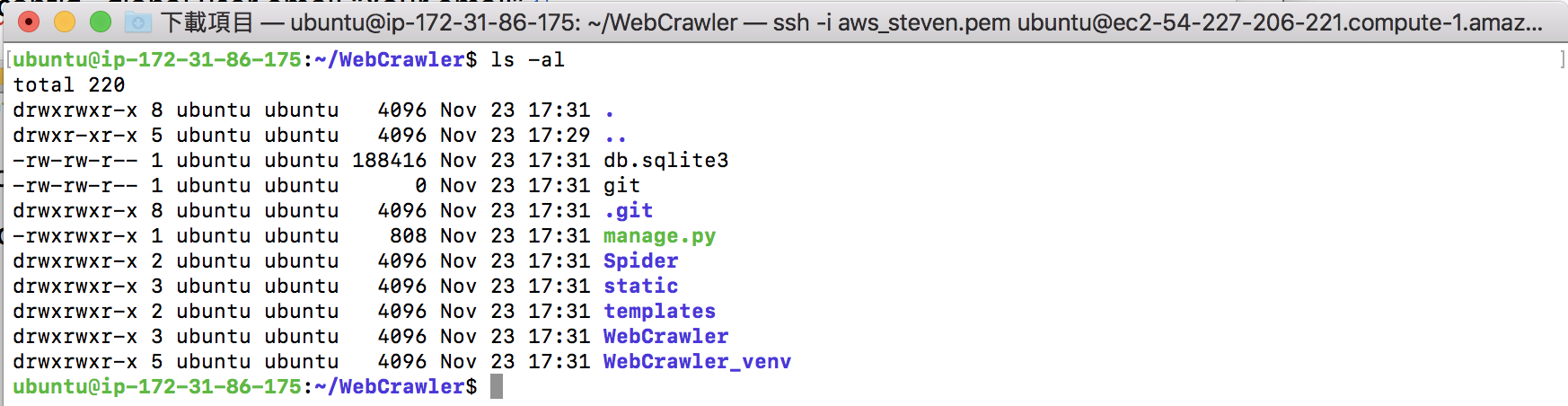


* 1. Clone source code from GitHub
     1. git clone <https://github.com/stevenang/WebCrawler.git>



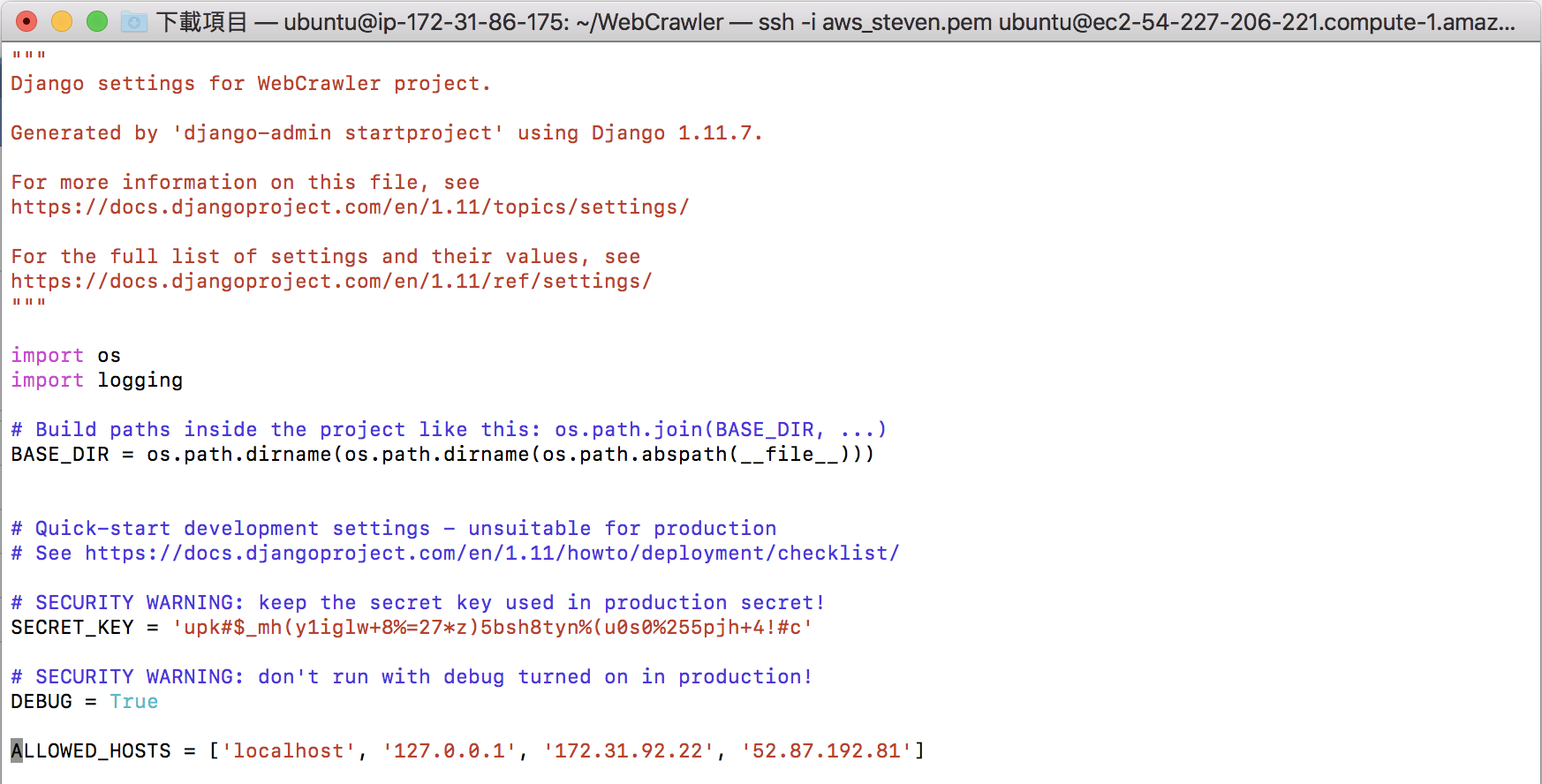
* + 1. Check whether the source code was downloaded or not





If you can see those folders, that means you already have our source code in your machine.

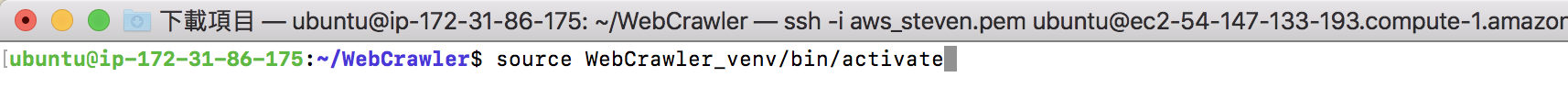
1. Edit setting file
   1. Execute command “vim WebCrawler/settings.py”
   2. Look for “ALLOWED HOST”



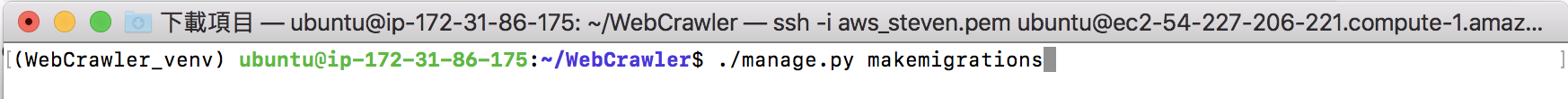
* 1. Add your IP address and save the changes

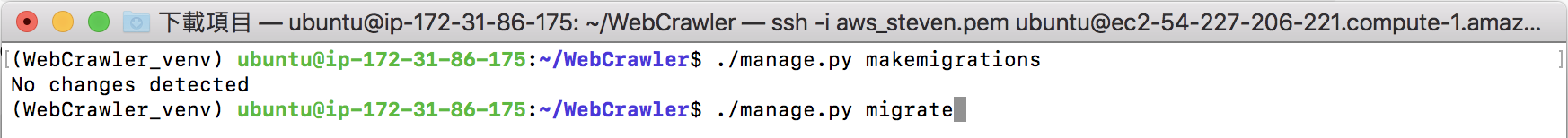


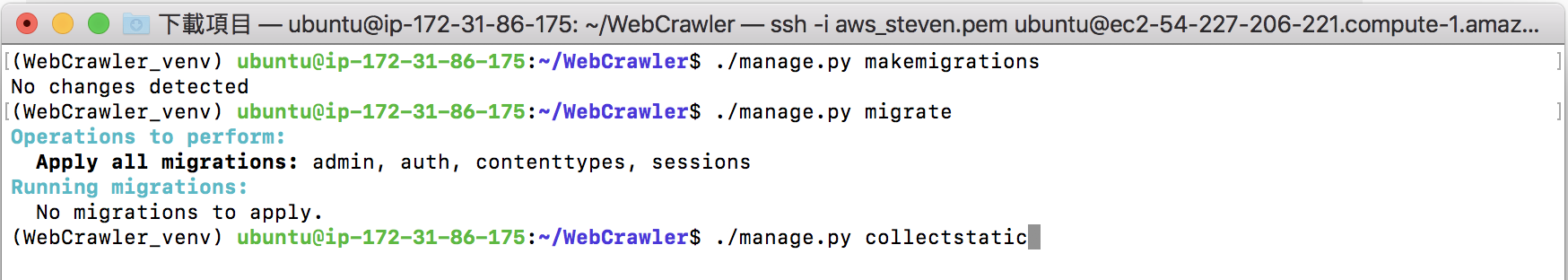
* 1. Execute command “source WebCrawler\_venv/bin/activate”



* 1. Complete Initial Project Setup







1. Modified Apache 2 Setting
   1. sudo vim /etc/apache2/sites-available/000-default.conf

Add the following lines

    Alias /static /home/ubuntu/WebCrawler/static

        <Directory /home/ubuntu/WebCrawler/static>

                Require all granted

        </Directory>

        <Directory /home/ubuntu/WebCrawler/WebCrawler>

                <Files wsgi.py>

                        Require all granted

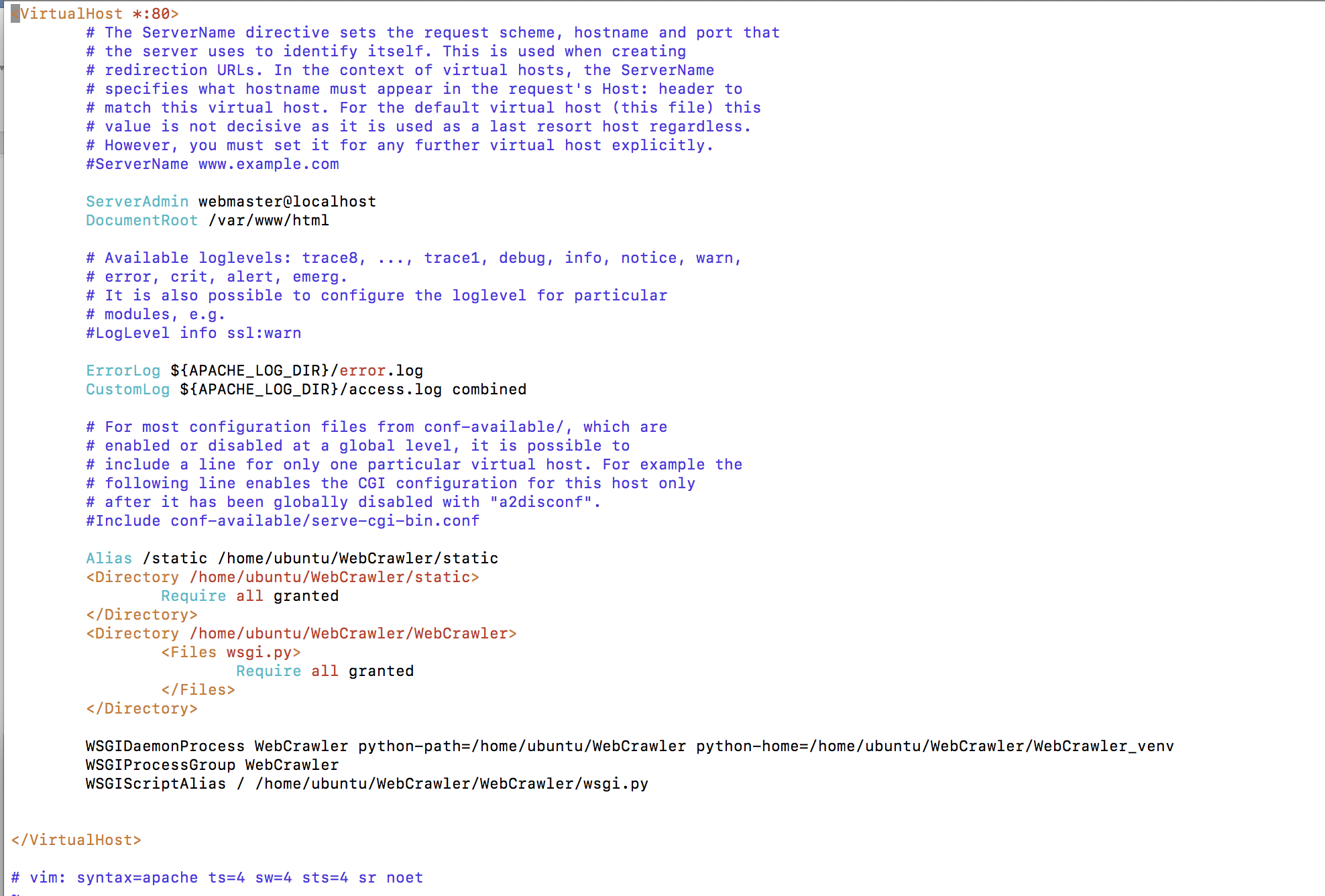
                </Files>

        </Directory>

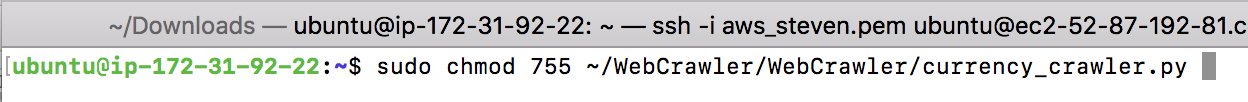
        WSGIDaemonProcess WebCrawler python-path=/home/ubuntu/WebCrawler python-home=/home/ubuntu/WebCrawler/WebCrawler\_venv

        WSGIProcessGroup WebCrawler

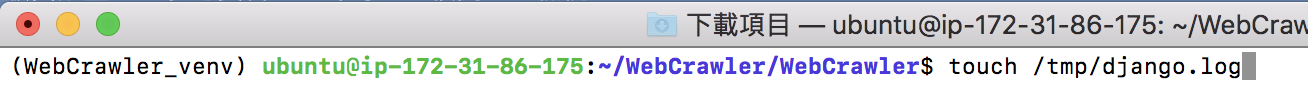
        WSGIScriptAlias / /home/ubuntu/WebCrawler/WebCrawler/wsgi.py



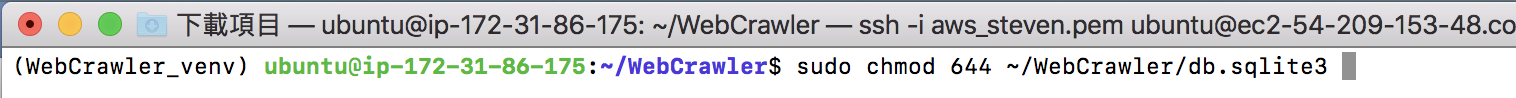
* 1. Execute the following commands:
     1. sudo chmod 755 ~/WebCrawler/WebCrawler/currency\_crawler.py



* + 1. touch /tmp/django.log



* + 1. sudo chmod 777 /tmp/django.log 
    2. sudo chmod 664 ~/WebCrawler/db.sqlite3



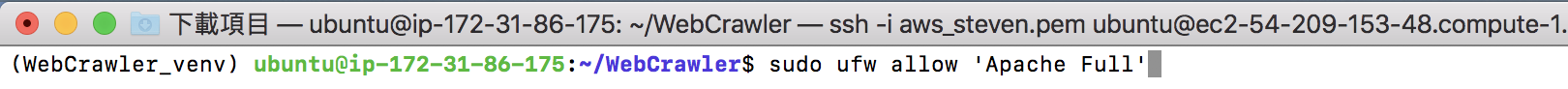
* + 1. sudo chown :www-data ~/WebCrawler/db.sqlite3



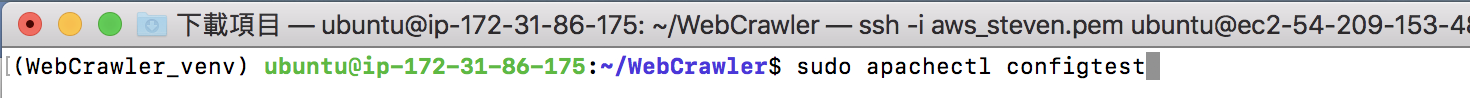
* + 1. sudo chown :www-data ~/WebCrawler



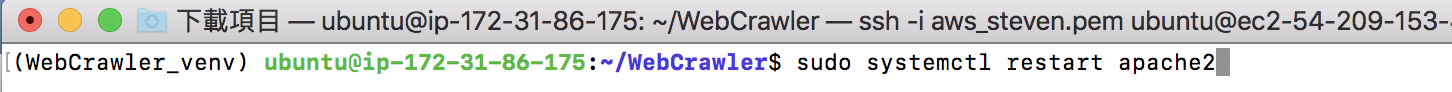
* + 1. sudo ufw allow ‘Apache Full’



* + 1. sudo apachectl configtest



* + 1. sudo systemctl restart apache2



1. Setup Crontab
   1. Execute command “sudo crontab –e” and enter the following information at the bottom of the file. Save the file after modification



1. Test the Web Application http://your\_ip\_address/index/

