南科新青年讲堂

基于Python的科研数据分析入门

游正新，张家澍

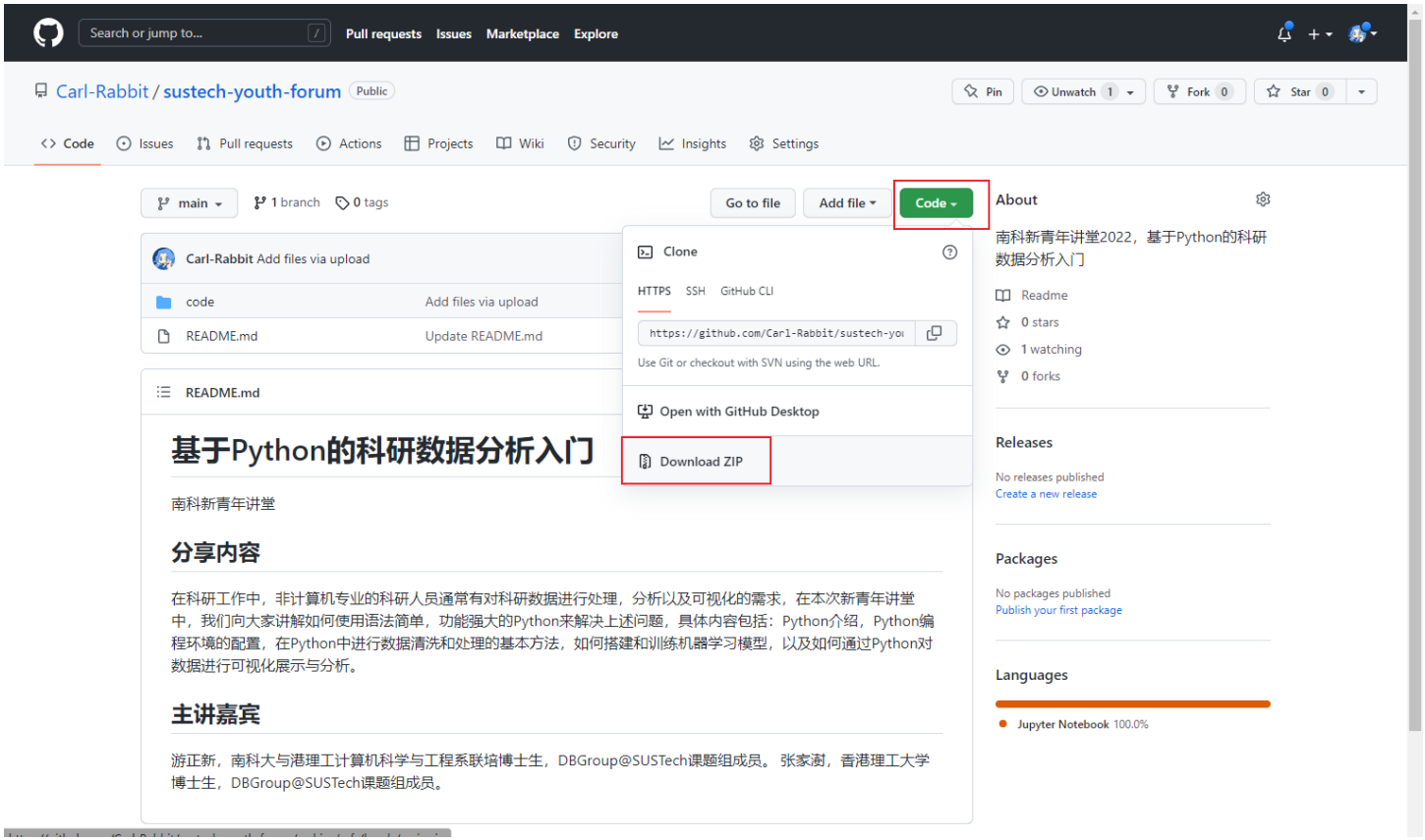
2022/5/8

目录

- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

下载资料

- https://github.com/Carl-Rabbit/sustech-youth-forum



目录

- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

什么是Python

- Python是一种编程语言

- 语法简便，功能强大，广泛应用于科研场景
  - 数据清洗
  - 数据分析
  - 数据可视化

目录

- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

安装Python

- 以Windows环境为例，向大家介绍Python的安装

- 使用Anaconda安装Python
  - 免费的Python环境管理平台
  - 提供简便易用的桌面操作环境
  - 包括Python，常用的软件包以及开发环境

Anaconda

- https://www.anaconda.com/

Data science technology for
a better world.

Anaconda offers the easiest way to perform Python/R data science and machine learning on a single machine. Start working with thousands of open-source packages and libraries today.

Download

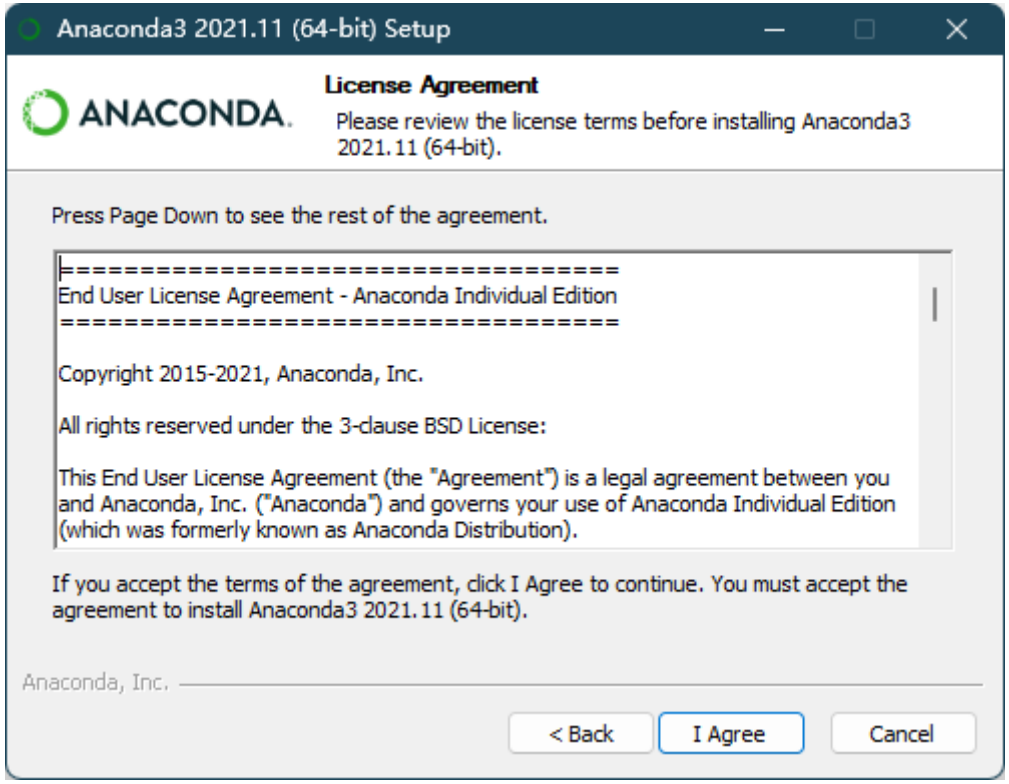
For Windows

Python 3.9 • 64-Bit Graphical Installer • 510 MB

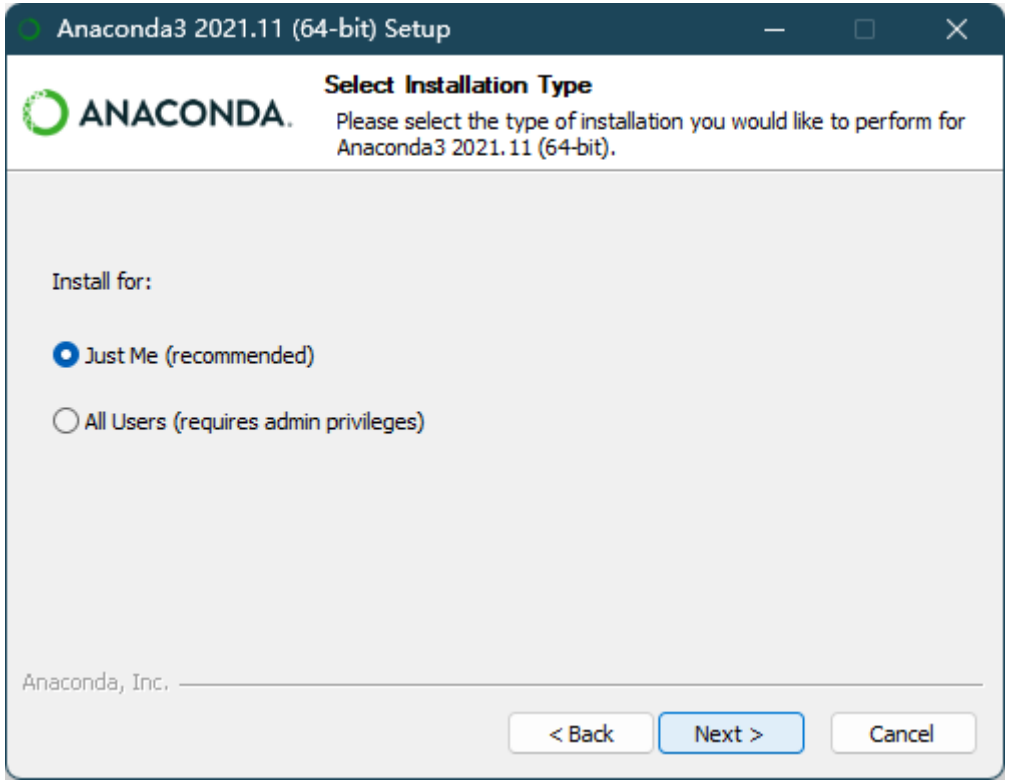
Get Additional Installers

下载和安装Anaconda

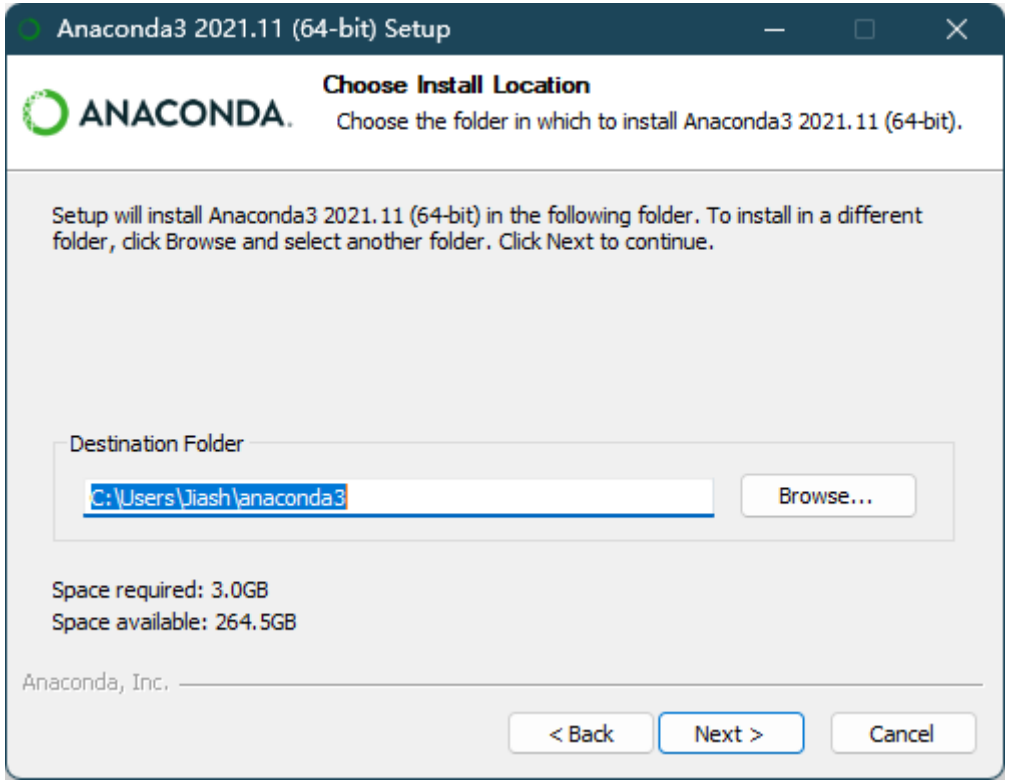
下载和安装Anaconda



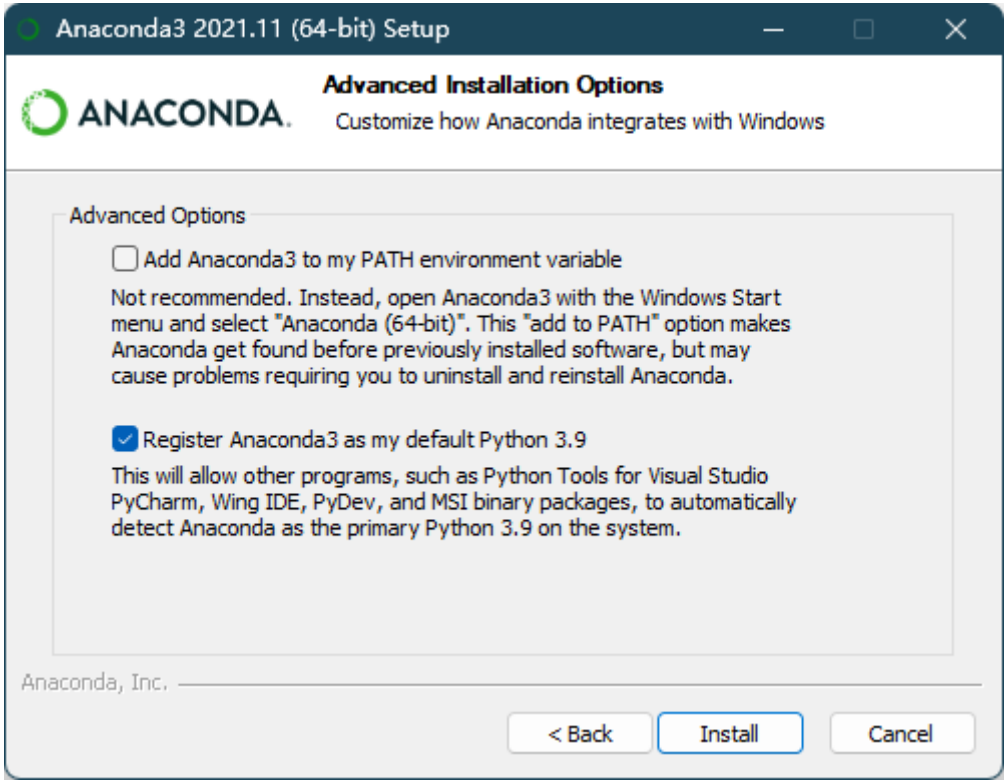
下载和安装Anaconda



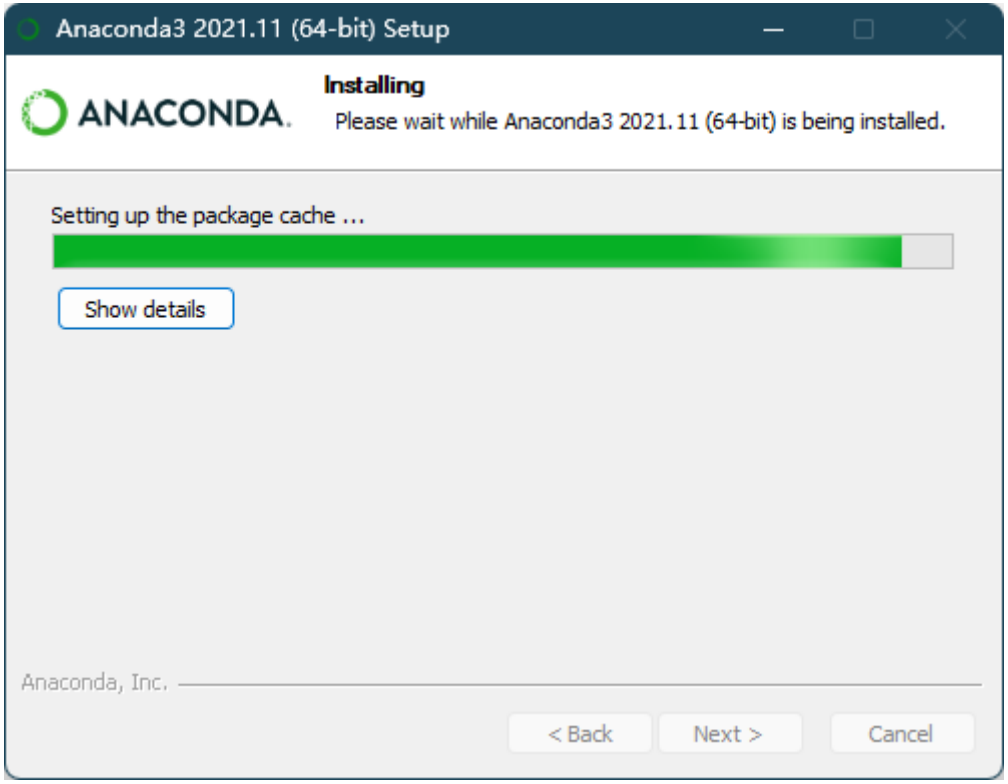
下载和安装Anaconda



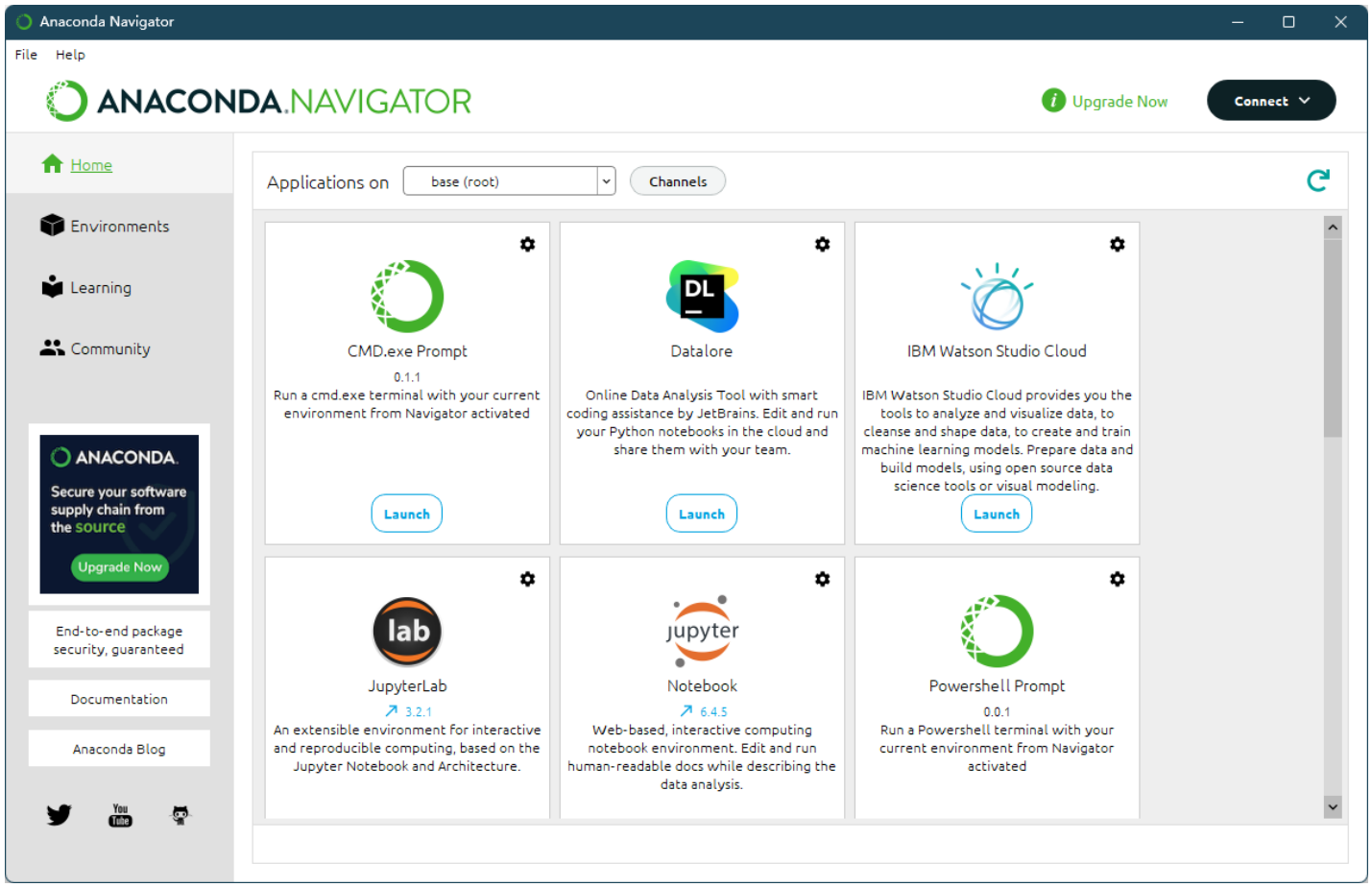
下载和安装Anaconda



下载和安装Anaconda



打开Anaconda

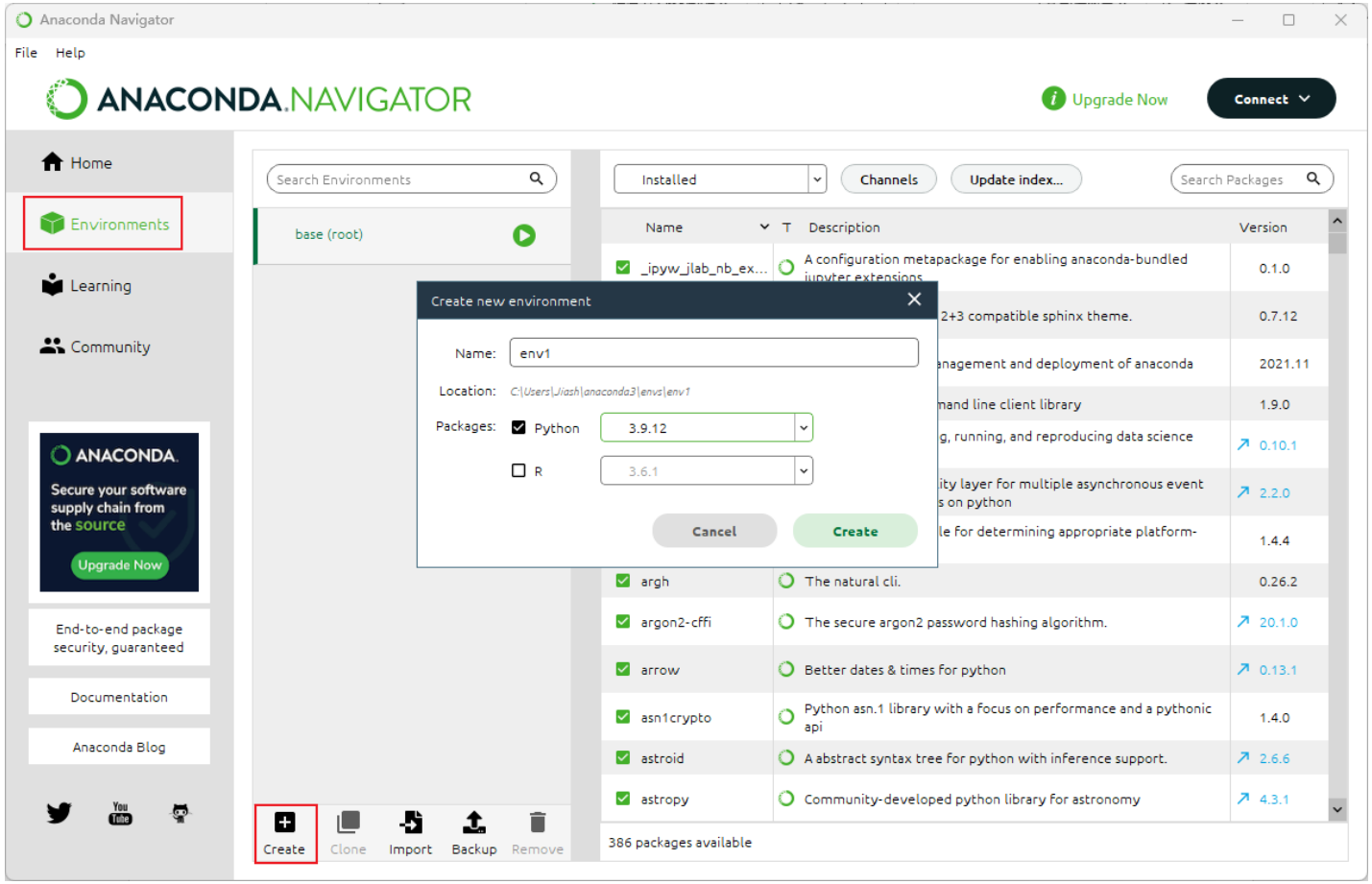

Anaconda Navigator
File
Help
ANACONDA.NAVIGATOR
Upgrade Now
Connect
Home
Environments
Learning
Community
Applications on
base (root)
Channels
CMD.exe Prompt
0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated
Launch
Datalore
Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.
Launch
IBM Watson Studio Cloud
IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.
Launch
ANACONDA.
Secure your software supply chain from the source
Upgrade Now
End-to-end package security, guaranteed
Documentation
Anaconda Blog
JupyterLab
3.2.1
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Notebook
6.4.5
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Powershell Prompt
0.0.1
Run a Powershell terminal with your current environment from Navigator activated

创建虚拟环境

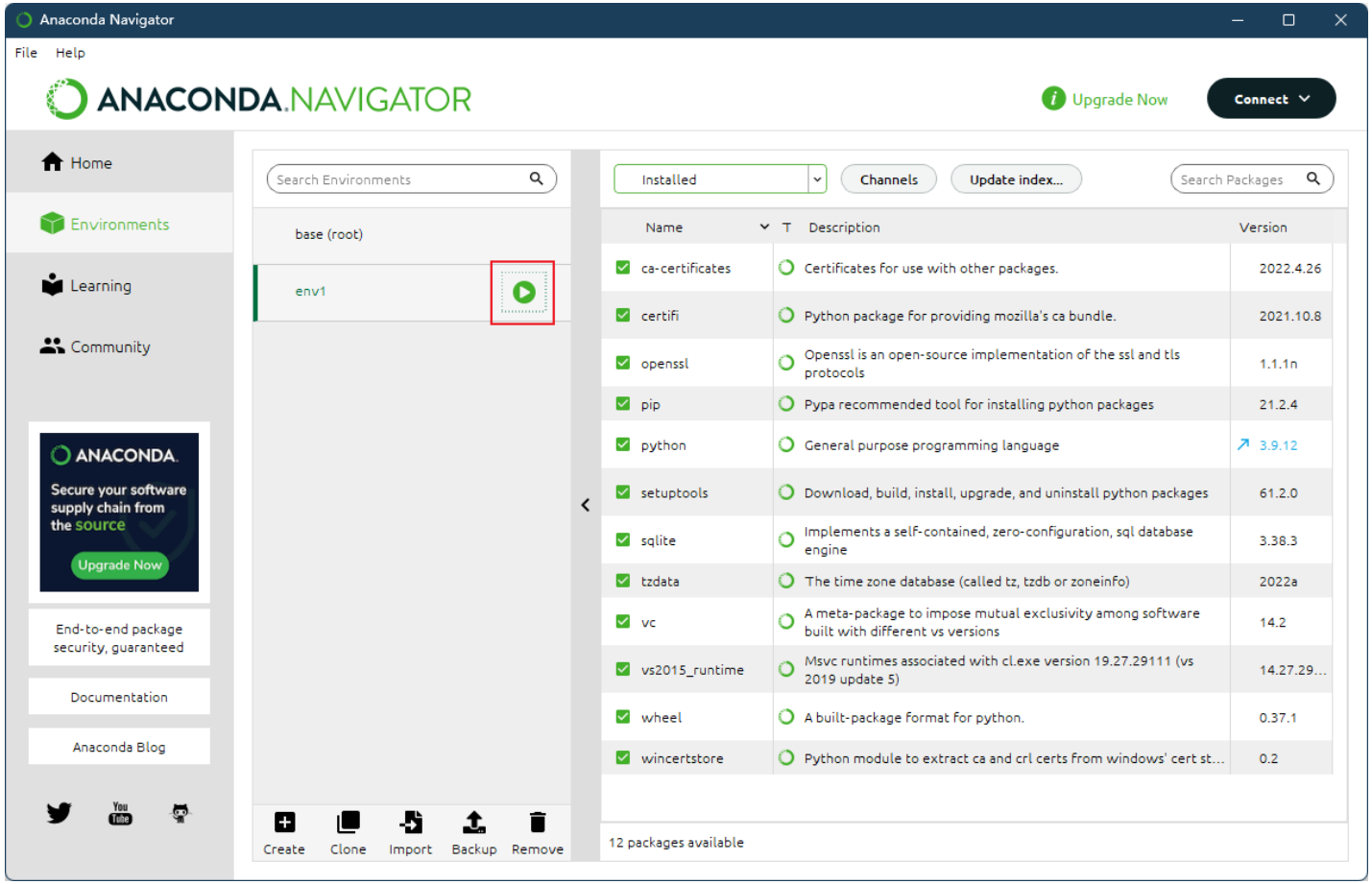
- 什么是虚拟环境？
  - 电脑上安装的Python本身是一个物理环境，我们可以基于这个物理环境创建多个虚拟环境

- 为什么要创建虚拟环境？
  - 隔离不同的环境
  - 解决复杂的环境需求

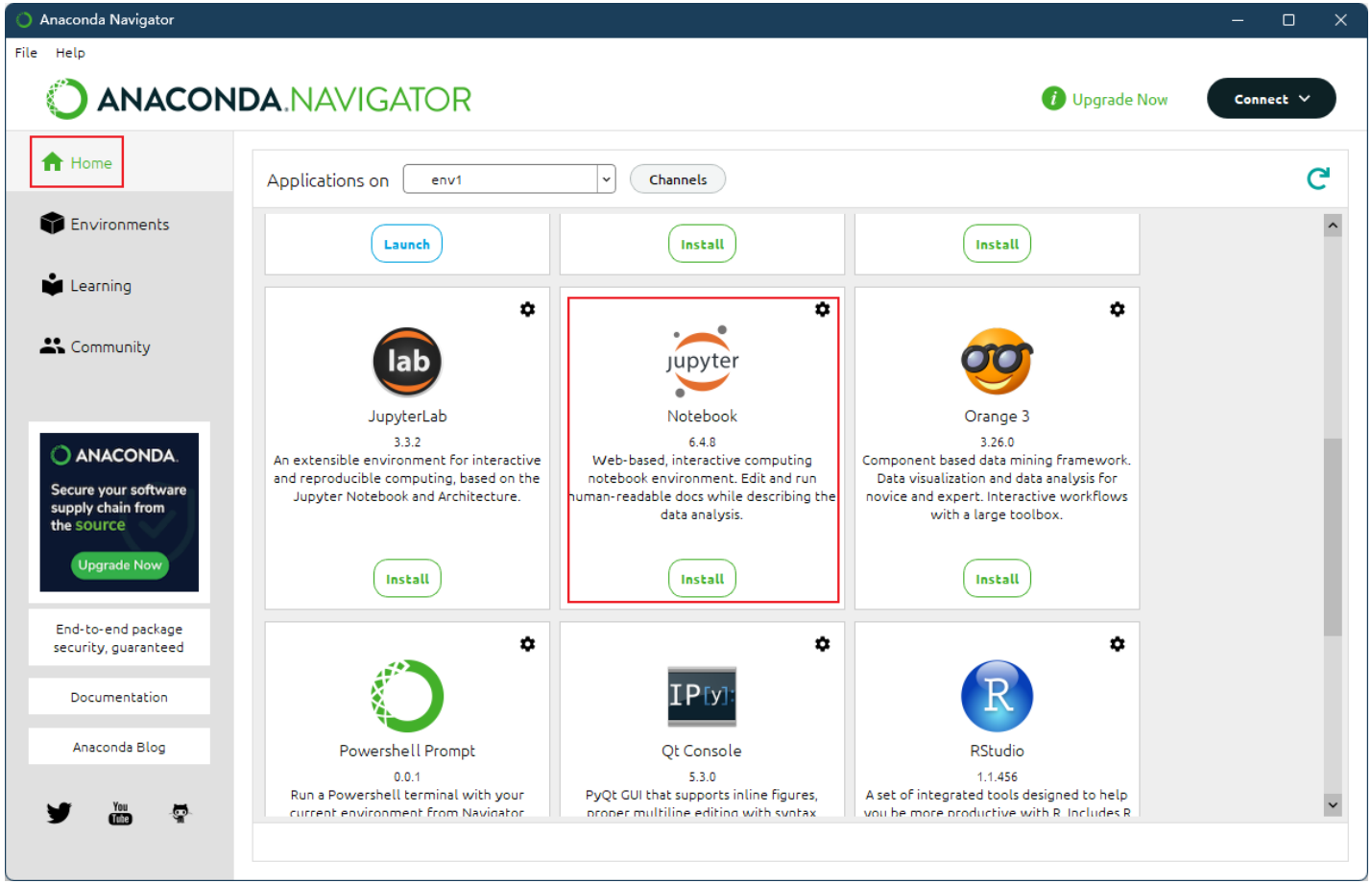
创建虚拟环境


Anaconda Navigator
File
Help
ANACONDA.NAVIGATOR
Upgrade Now
Connect
Home
Environments
Learning
Community
Search Environments
base (root)
Installed
Channels
Update index...
Search Packages
Name
T
Description
Version
_ipyw_jlab_nb_ex...
A configuration metapackage for enabling anaconda-bundled jupyter extensions
0.1.0
2+3 compatible sphinx theme.
0.7.12
anagement and deployment of anaconda
2021.11
mand line client library
1.9.0
g, running, and reproducing data science
0.10.1
ity layer for multiple asynchronous event s on python
2.2.0
le for determining appropriate platform-
1.4.4
argh
The natural cli.
0.26.2
argon2-cffi
The secure argon2 password hashing algorithm.
20.1.0
arrow
Better dates & times for python
0.13.1
asn1crypto
Python asn.1 library with a focus on performance and a pythonic api
1.4.0
astroid
A abstract syntax tree for python with inference support.
2.6.6
astropy
Community-developed python library for astronomy
4.3.1
386 packages available
Create new environment
Name:
env1
Location:
C:\Users\Jiash\anaconda3\envs\env1
Packages:
Python
3.9.12
R
3.6.1
Cancel
Create
Create
Clone
Import
Backup
Remove
ANACONDA.
Secure your software supply chain from the source
Upgrade Now
End-to-end package security, guaranteed
Documentation
Anaconda Blog

使用虚拟环境

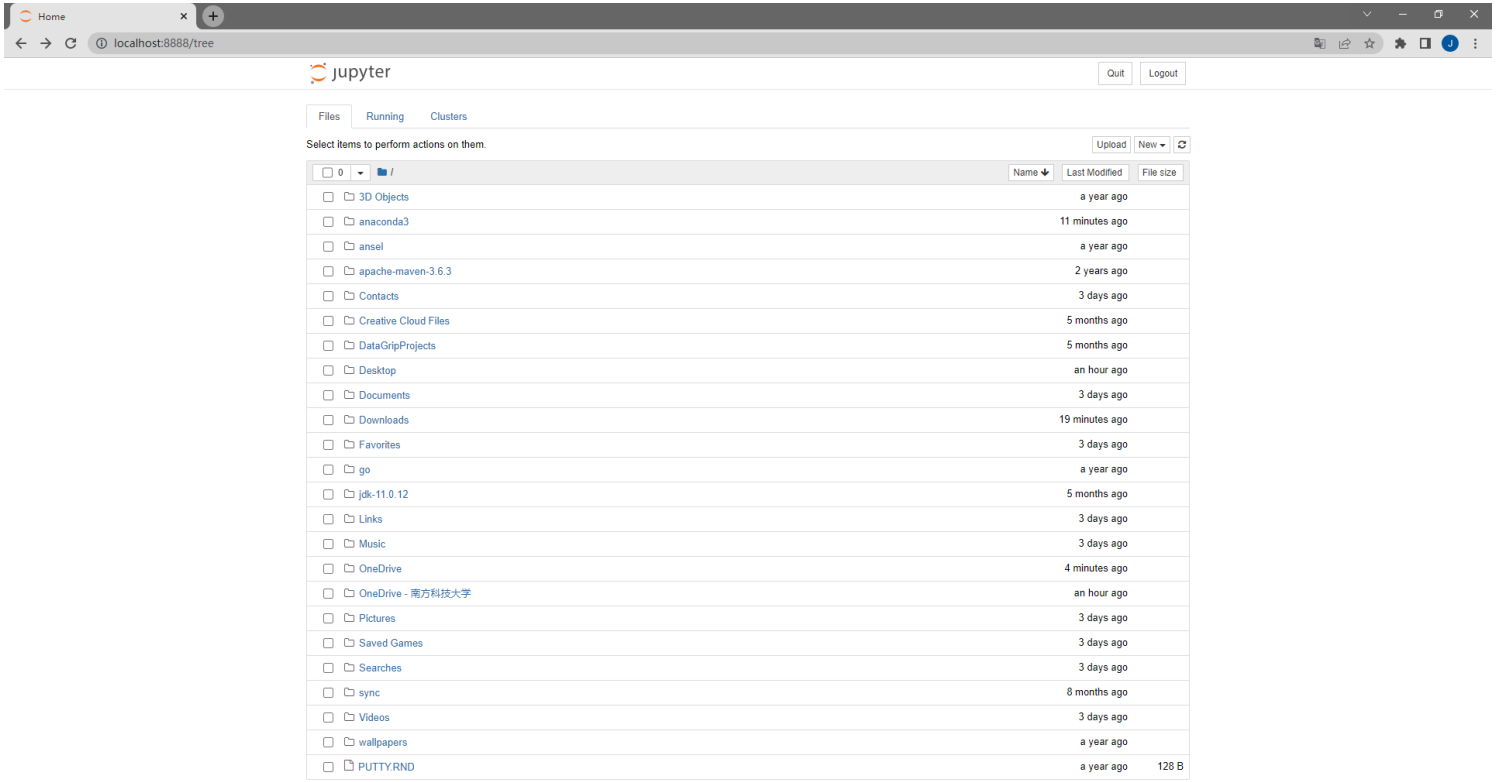

Anaconda Navigator
File
Help
ANACONDA.NAVIGATOR
Upgrade Now
Connect
Home
Environments
Learning
Community
ANACONDA.
Secure your software supply chain from the source
Upgrade Now
End-to-end package security, guaranteed
Documentation
Anaconda Blog
Search Environments
base (root)
env1
Create
Clone
Import
Backup
Remove
Installed
Channels
Update index...
Search Packages
Name
T
Description
Version
ca-certificates
Certificates for use with other packages.
2022.4.26
certifi
Python package for providing mozilla's ca bundle.
2021.10.8
openssl
Openssl is an open-source implementation of the ssl and tls protocols
1.1.1n
pip
Pypa recommended tool for installing python packages
21.2.4
python
General purpose programming language
3.9.12
setuptools
Download, build, install, upgrade, and uninstall python packages
61.2.0
sqlite
Implements a self-contained, zero-configuration, sql database engine
3.38.3
tzdata
The time zone database (called tz, tzdb or zoneinfo)
2022a
vc
A meta-package to impose mutual exclusivity among software built with different vs versions
14.2
vs2015_runtime
Msvc runtimes associated with cl.exe version 19.27.29111 (vs 2019 update 5)
14.27.29...
wheel
A built-package format for python.
0.37.1
wincertstore
Python module to extract ca and crl certs from windows' cert st...
0.2
12 packages available

安装Jupyter Notebook

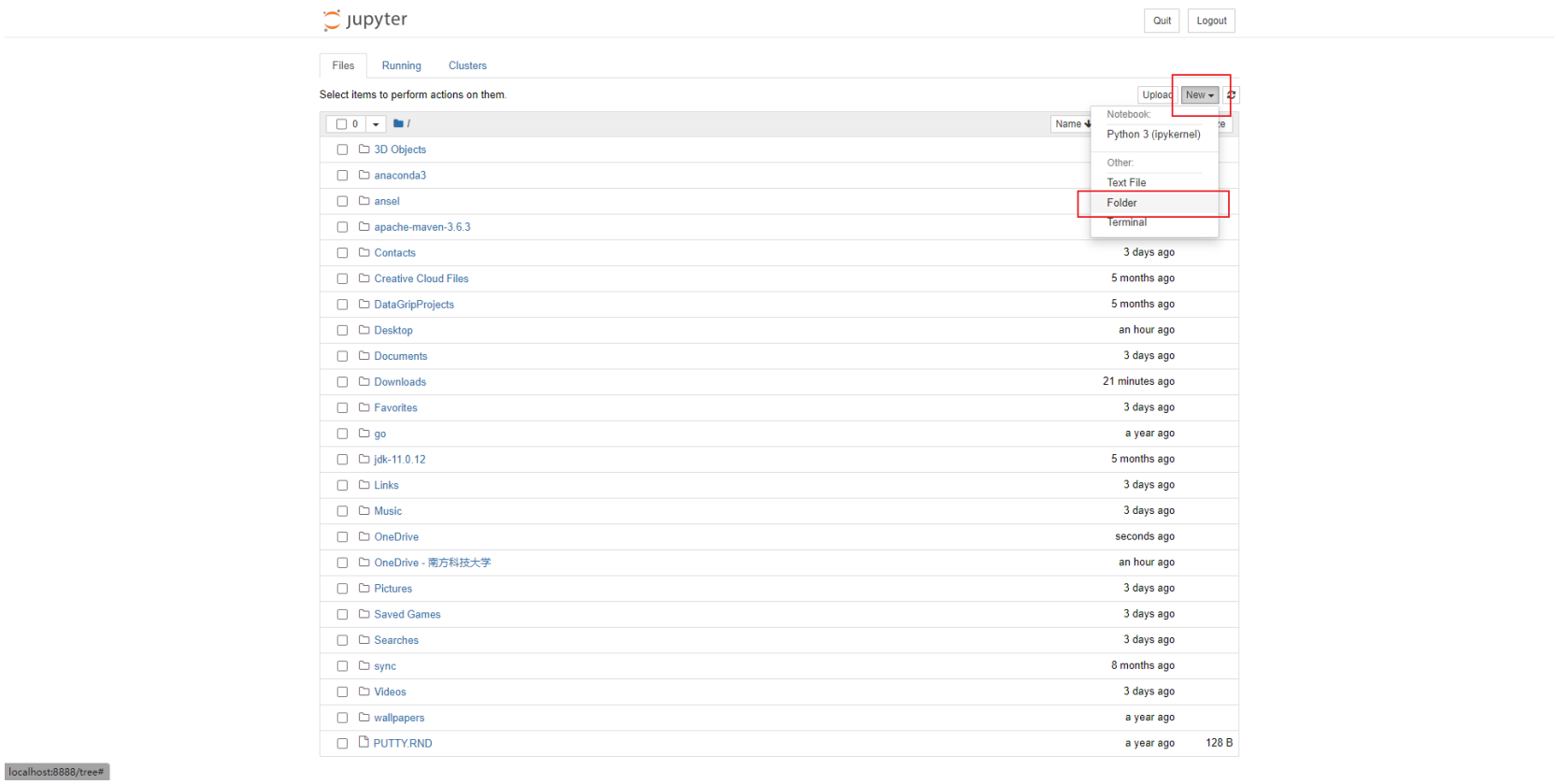


Jupyter Notebook

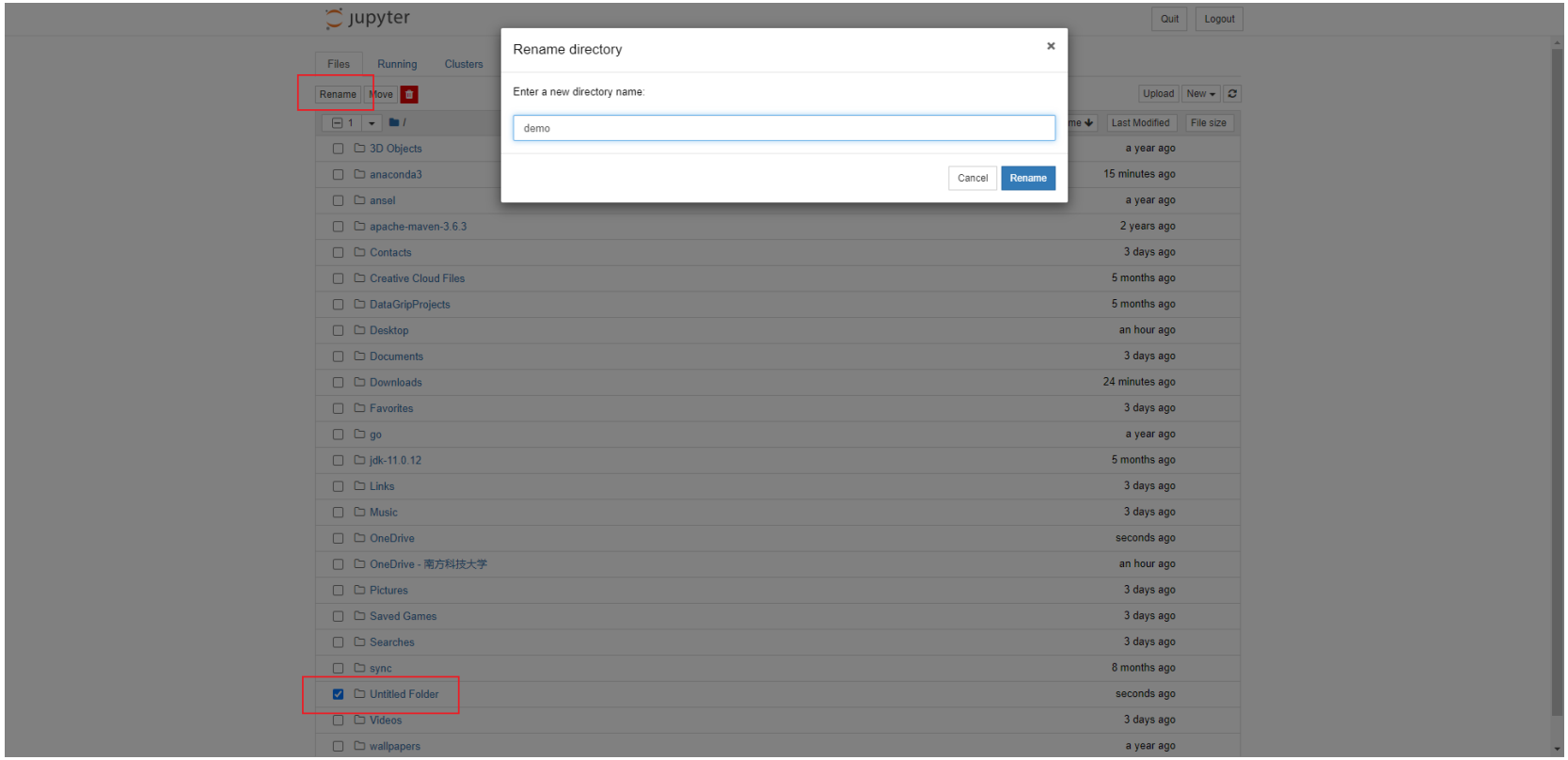
- 在浏览器里写代码，运行代码



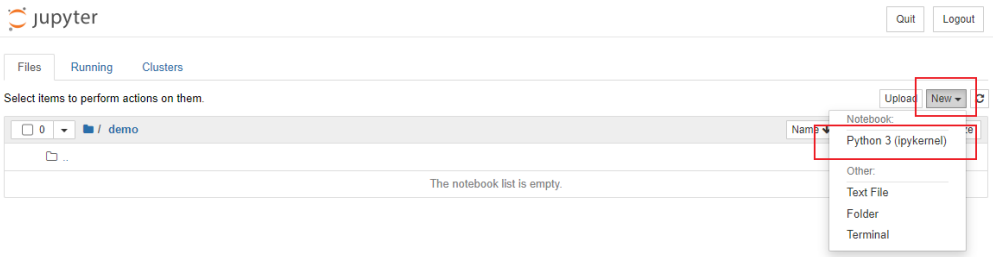
写下第一行代码


jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
Notebook:
Python 3 (ipykernel)
Other:
Text File
Folder
Terminal
Name
3D Objects
anaconda3
ansel
apache-maven-3.6.3
Contacts 3 days ago
Creative Cloud Files 5 months ago
DataGripProjects 5 months ago
Desktop an hour ago
Documents 3 days ago
Downloads 21 minutes ago
Favorites 3 days ago
go a year ago
jdk-11.0.12 5 months ago
Links 3 days ago
Music 3 days ago
OneDrive seconds ago
OneDrive - 南方科技大学 an hour ago
Pictures 3 days ago
Saved Games 3 days ago
Searches 3 days ago
sync 8 months ago
Videos 3 days ago
wallpapers a year ago
PUTTY.RND a year ago 128 B
localhost:8888/tree#

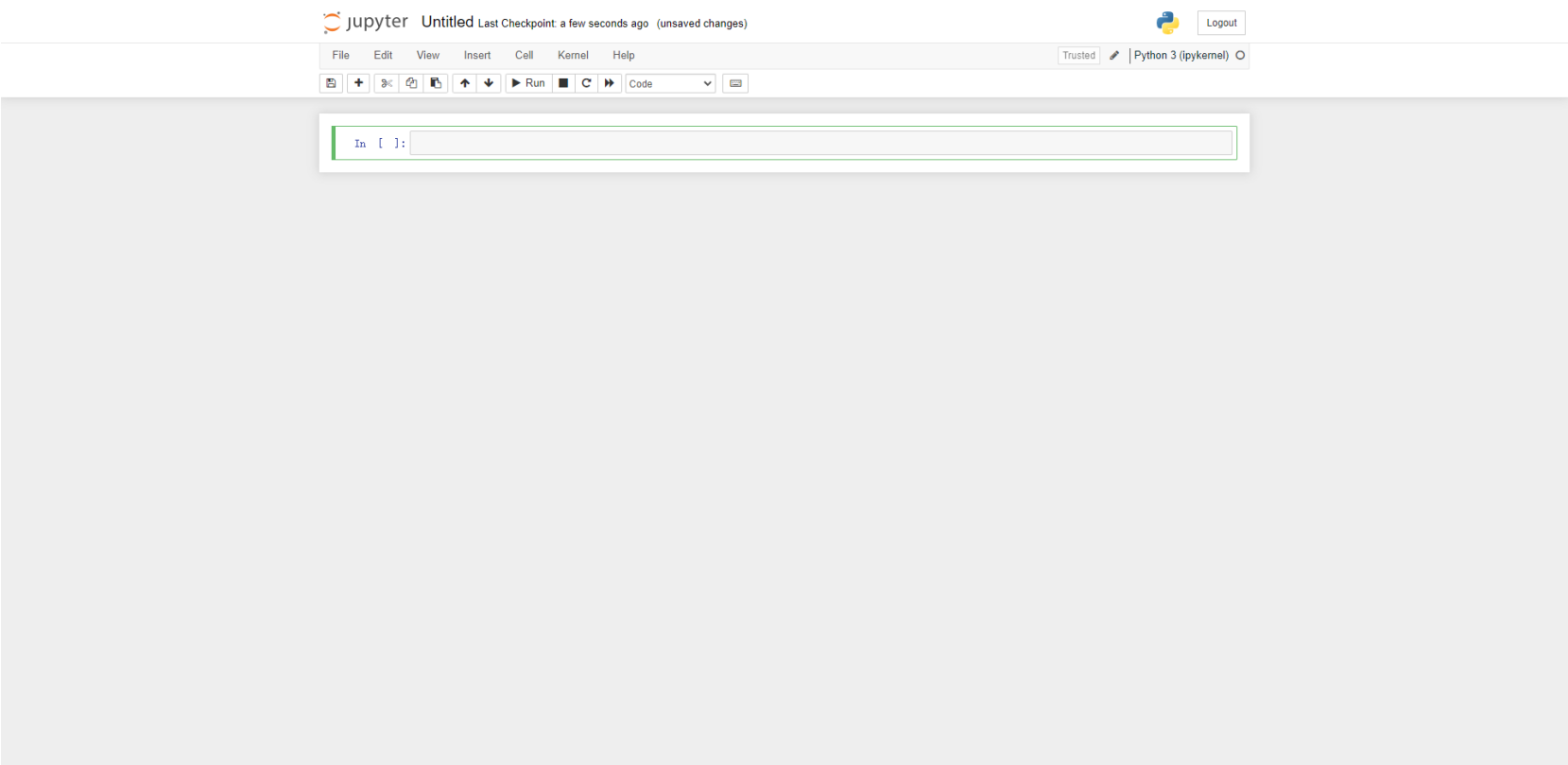
写下第一行代码


jupyter
Quit
Logout
Files
Running
Clusters
Rename
Move
Upload
New
Last Modified
File size
Rename directory
Enter a new directory name:
demo
Cancel
Rename
3D Objects
anaconda3
ansel
apache-maven-3.6.3
Contacts
Creative Cloud Files
DataGripProjects
Desktop
Documents
Downloads
Favorites
go
jdk-11.0.12
Links
Music
OneDrive
OneDrive - 南方科技大学
Pictures
Saved Games
Searches
sync
Untitled Folder
Videos
wallpapers
a year ago
15 minutes ago
a year ago
2 years ago
3 days ago
5 months ago
5 months ago
an hour ago
3 days ago
24 minutes ago
3 days ago
a year ago
5 months ago
3 days ago
3 days ago
seconds ago
an hour ago
3 days ago
3 days ago
3 days ago
8 months ago
seconds ago
3 days ago
a year ago

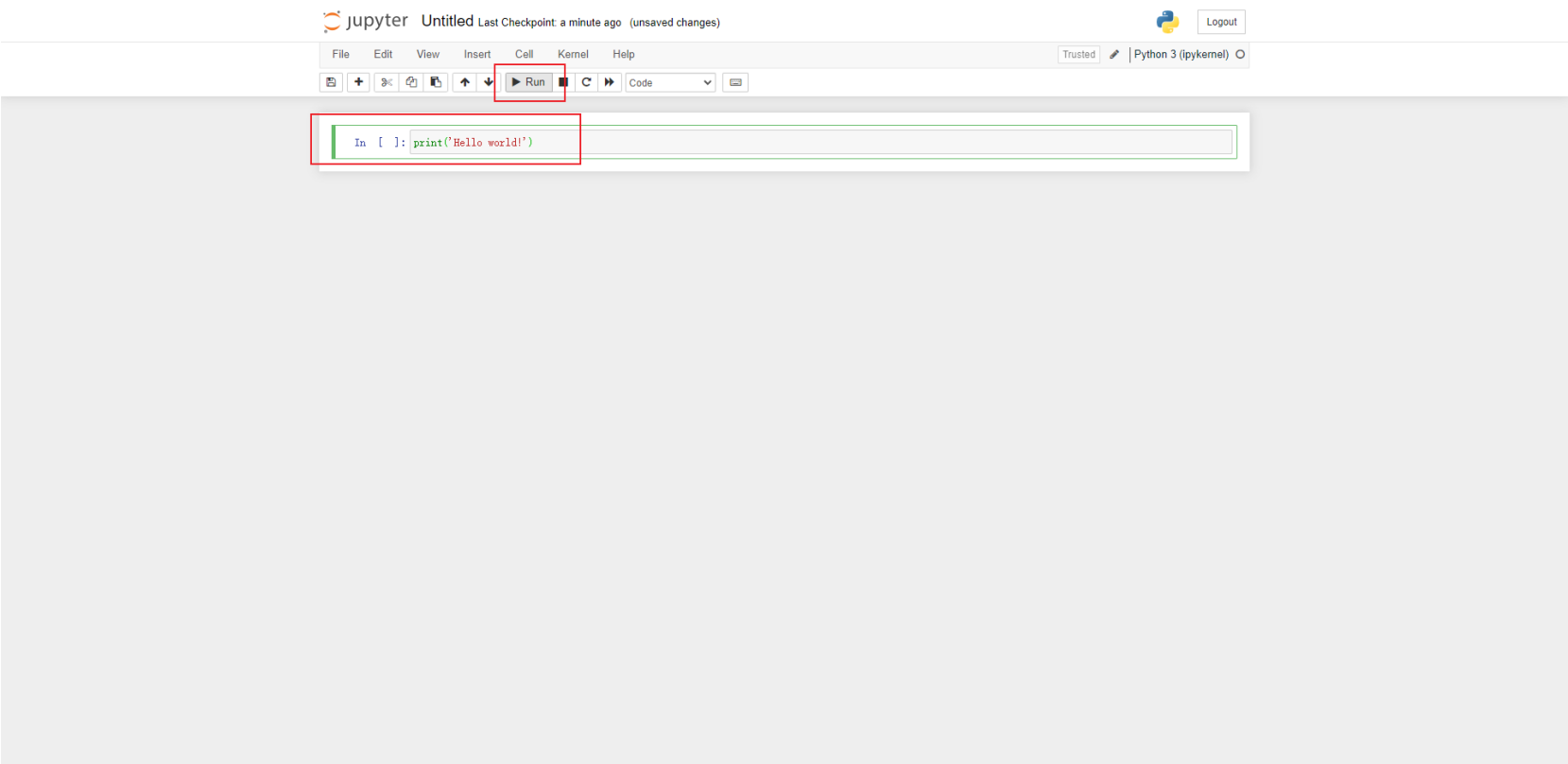
写下第一行代码


jupyter
Quit
Logout
Files
Running
Clusters
Select items to perform actions on them.
Upload
New
0
/ demo
Name
..
The notebook list is empty.
Notebook:
Python 3 (ipykernel)
Other:
Text File
Folder
Terminal

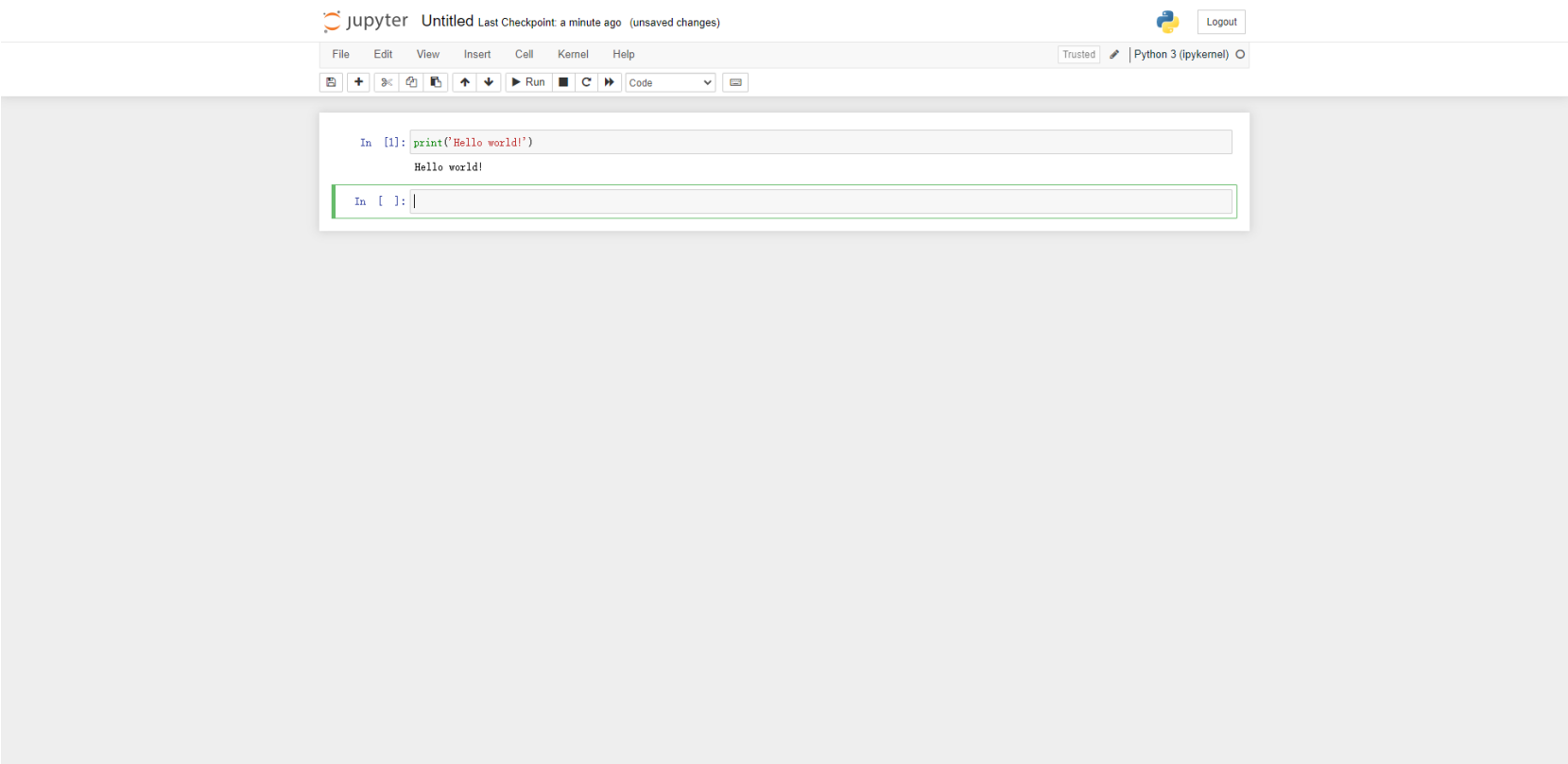
写下第一行代码


jupyter
Untitled Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 3 (ipykernel)
Run
Code
In [ ]:

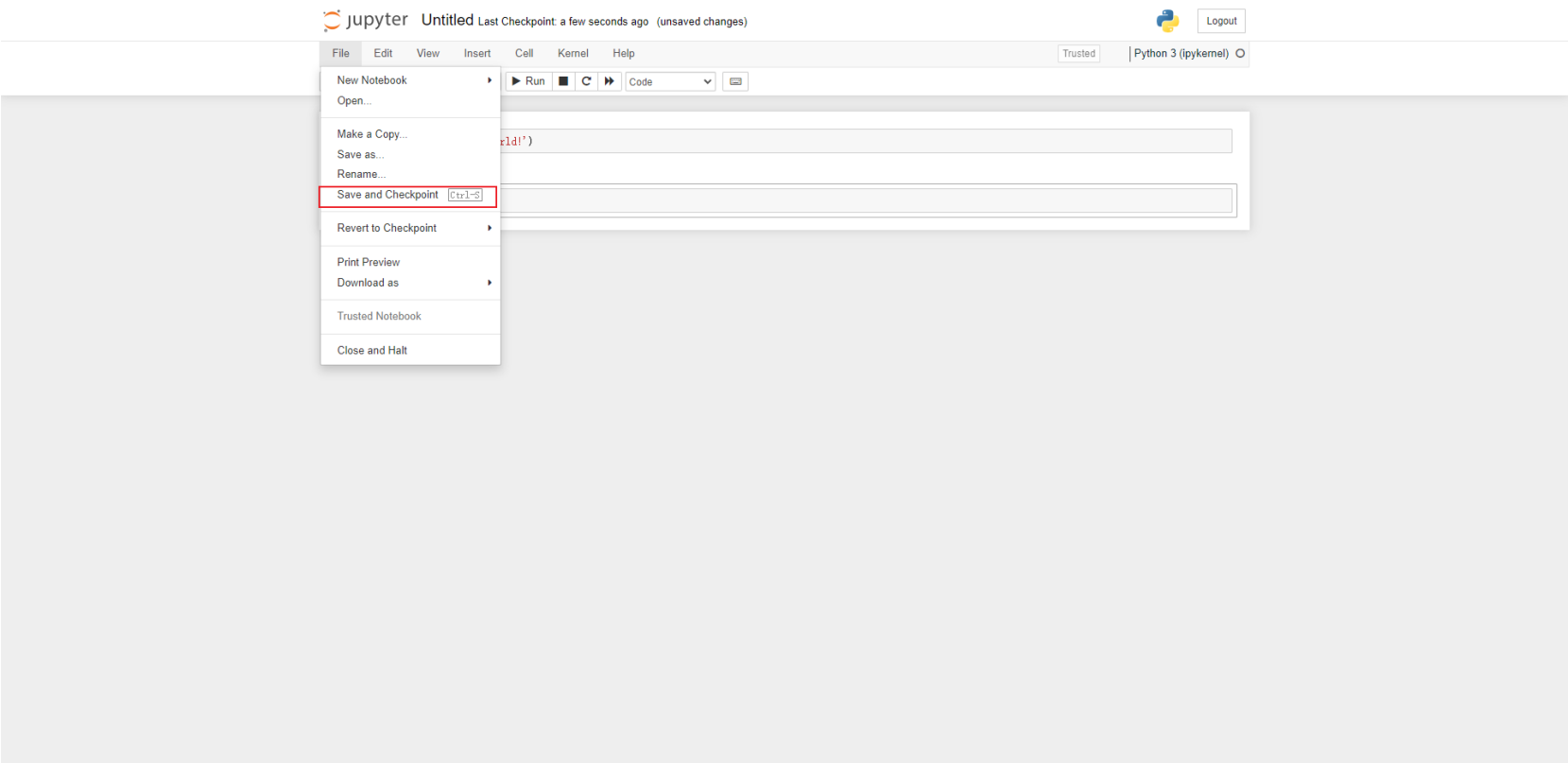
写下第一行代码


jupyter Untitled Last Checkpoint: a minute ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3 (ipykernel)
Run
Code
In [ ]: print('Hello world!')

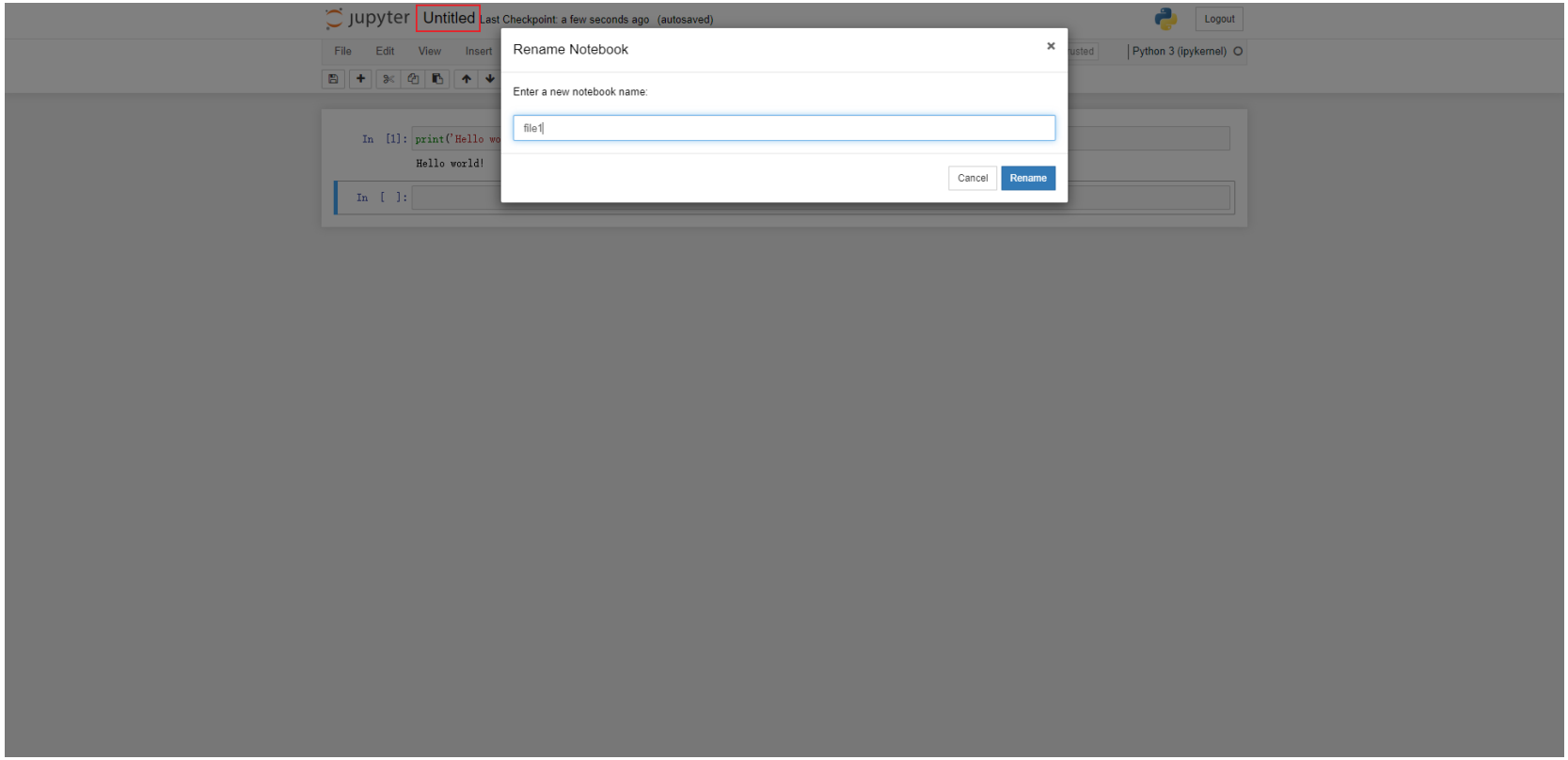
写下第一行代码


jupyter Untitled Last Checkpoint: a minute ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Help
Trusted
Python 3 (ipykernel)
Run
Code
In [1]: print('Hello world!')
Hello world!
In [ ]:

保存代码


jupyter
Untitled Last Checkpoint: a few seconds ago (unsaved changes)
Logout
File
Edit
View
Insert
Cell
Kernel
Help
Trusted
Python 3 (ipykernel)
New Notebook
Open...
Make a Copy...
Save as...
Rename...
Save and Checkpoint Ctrl-S
Revert to Checkpoint
Print Preview
Download as
Trusted Notebook
Close and Halt
Run
Code

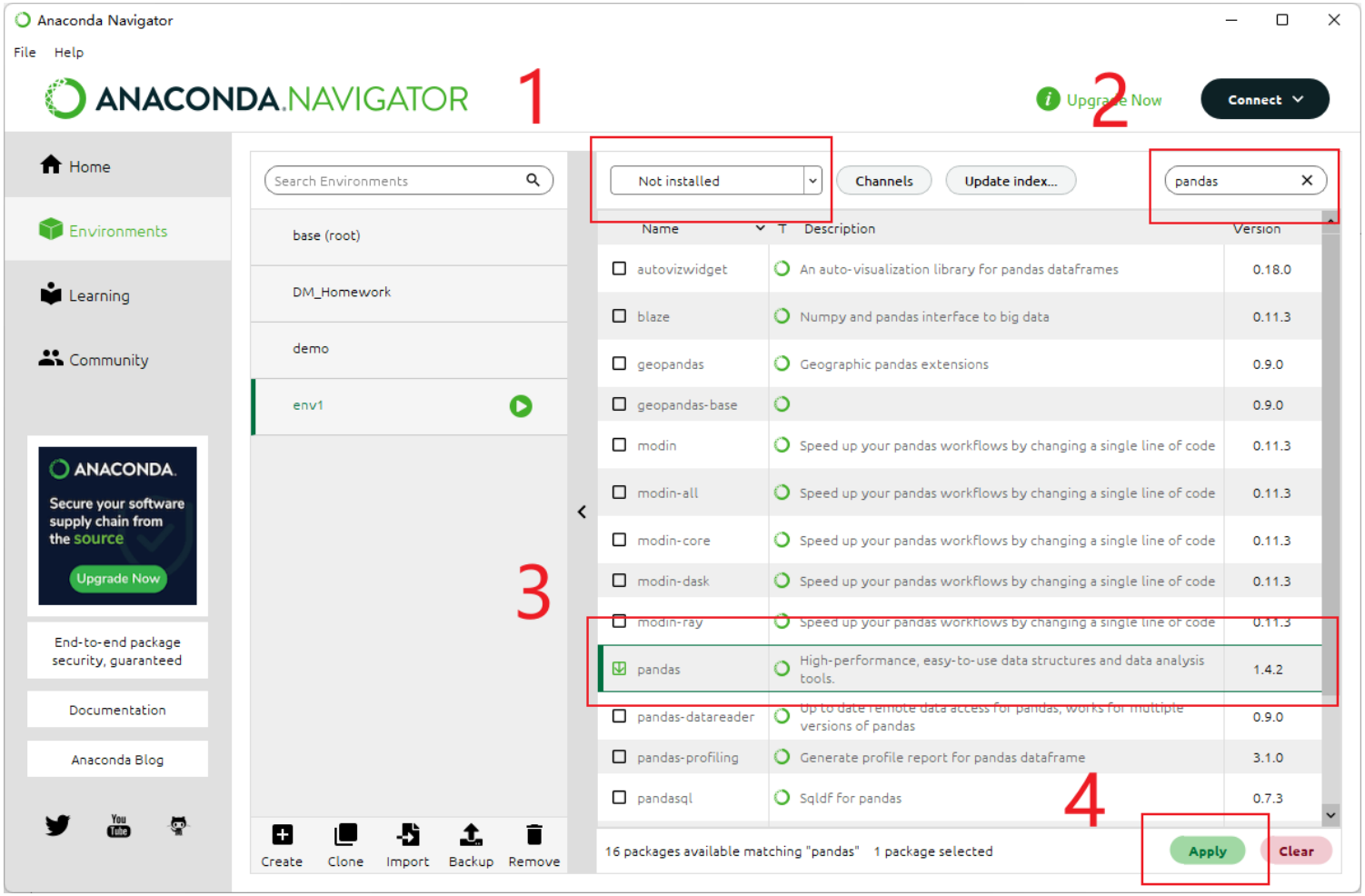
保存代码


jupyter
Untitled
Last Checkpoint: a few seconds ago (autosaved)
Logout
File
Edit
View
Insert
Python 3 (ipykernel)
Rename Notebook
Enter a new notebook name:
file1
Cancel
Rename
In [1]: print('Hello wo
Hello world!
In [ ]:

目录

- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

安装pandas


Anaconda Navigator
File Help
ANACONDA.NAVIGATOR
1
Upgrade Now
2
Connect
Home
Environments
Learning
Community
Search Environments
base (root)
DM_Homework
demo
env1
3
Not installed
Channels
Update index...
pandas
Name
T
Description
Version
autovizwidget
An auto-visualization library for pandas dataframes
0.18.0
blaze
Numpy and pandas interface to big data
0.11.3
geopandas
Geographic pandas extensions
0.9.0
geopandas-base
0.9.0
modin
Speed up your pandas workflows by changing a single line of code
0.11.3
modin-all
Speed up your pandas workflows by changing a single line of code
0.11.3
modin-core
Speed up your pandas workflows by changing a single line of code
0.11.3
modin-dask
Speed up your pandas workflows by changing a single line of code
0.11.3
modin-ray
Speed up your pandas workflows by changing a single line of code
0.11.3
pandas
High-performance, easy-to-use data structures and data analysis tools.
1.4.2
pandas-datareader
Up to date remote data access for pandas, works for multiple versions of pandas
0.9.0
pandas-profiling
Generate profile report for pandas dataframe
3.1.0
pandasql
Sqldf for pandas
0.7.3
4
16 packages available matching "pandas"
1 package selected
Apply
Clear
Create
Clone
Import
Backup
Remove
ANACONDA.
Secure your software supply chain from the source
Upgrade Now
End-to-end package security, guaranteed
Documentation
Anaconda Blog

加载数据

```python
import pandas as pd

# 注意需要设置dtype，否则date列会被识别为浮点数
data = pd.read_csv('./data/covid-19-raw.csv', dtype={'date': str})
data
```

	date	year	newConfirm	newSuspect	newImportedCase	newInfect	newRecovered	newDeath	nowConfirm
0	04.05	2022.0	2263.0	0.0	32.0	19199.0	3357.0	87.0	280083
1	04.06	2022.0	2097.0	4.0	39.0	21784.0	2420.0	111.0	279649
2	04.07	2022.0	3096.0	4.0	36.0	22648.0	3358.0	97.0	279290
3	04.08	NaN	NaN	NaN	NaN	NaN	NaN	NaN	277812
4	04.09	2022.0	2049.0	0.0	33.0	25111.0	3491.0	63.0	276307

清理数据

- 删除不需要的数据列

- 对于数据中缺失的数据，有4种常用的填补方法:
  - 直接删除包含缺失数据的行
  - 用预设值填充
  - 用前一个或后一个合法值填充
  - 用均值或中值填充

安装Scikit-learn

scikit-learn	A set of python modules for machine learning and data mining	1.0.2

- Scikit-learn是Python的一个免费机器学习库
- 它包含多个常用的机器学习模型，包括:
  - 支持向量机
  - 随机森林
  - 决策树
  - 神经网络
  - …

训练线性回归模型

- 以新增确诊（newConfirm）为例

```
from sklearn.model_selection import train_test_split
from sklearn import linear_model

# sklearn的输入是二维数组
x = [[d] for d in range(len(data))]
y = [[d] for d in data['newConfirm'].tolist()]

# 取7天用于预测，剩下的数据用于训练
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=7/len(data), shuffle=False)

# 创建并训练模型
model = linear_model.LinearRegression()
model.fit(x_train, y_train)

# 用模型进行预测
y_test_pred = model.predict(x_test)

print(f'Expression: y = {model.coef_[0][0]}x + {model.intercept_[0]}')
print(y_test_pred)
```

可视化结果

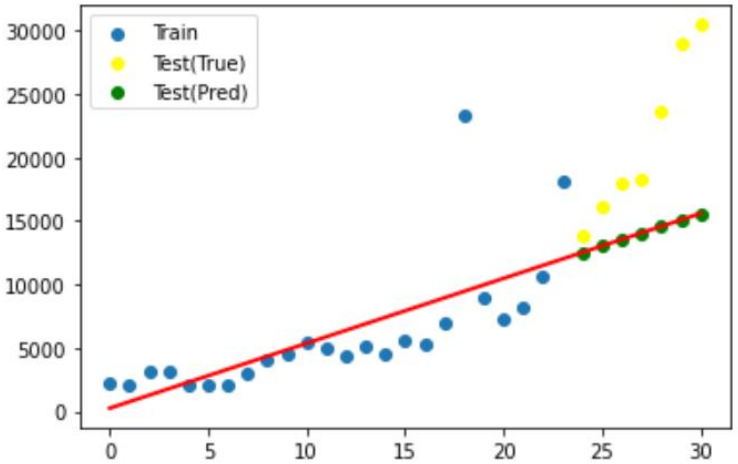
matplotlib	Publication quality figures in python	3.5.1

```python
import matplotlib.pyplot as plt

# 绘制训练数据
plt.scatter(x_train, y_train, label='Train')
# 绘制测试数据的真实值
plt.scatter(x_test, y_test, color='yellow', label='Test(True)')
# 绘制测试数据的预测值
plt.scatter(x_test, [d[0] for d in y_test_pred],
            color='green', label='Test(Pred)')
# 绘制线性回归模型
end_points_x = [[0],[len(data) - 1]]
plt.plot(end_points_x, model.predict(end_points_x),
         color="red", linewidth=2, linestyle="-")

plt.legend(loc = 'best')

plt.show()
```

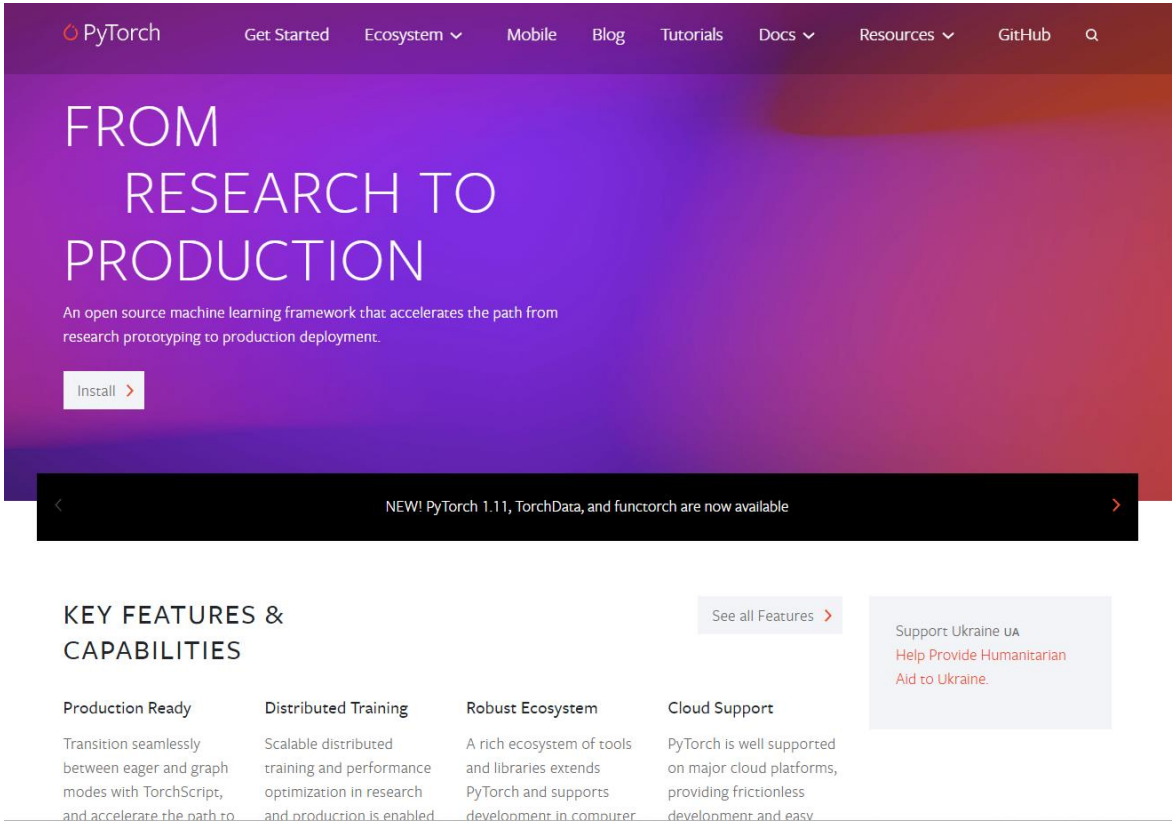


目录

- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

使用PyTorch训练神经网络

- https://pytorch.org/

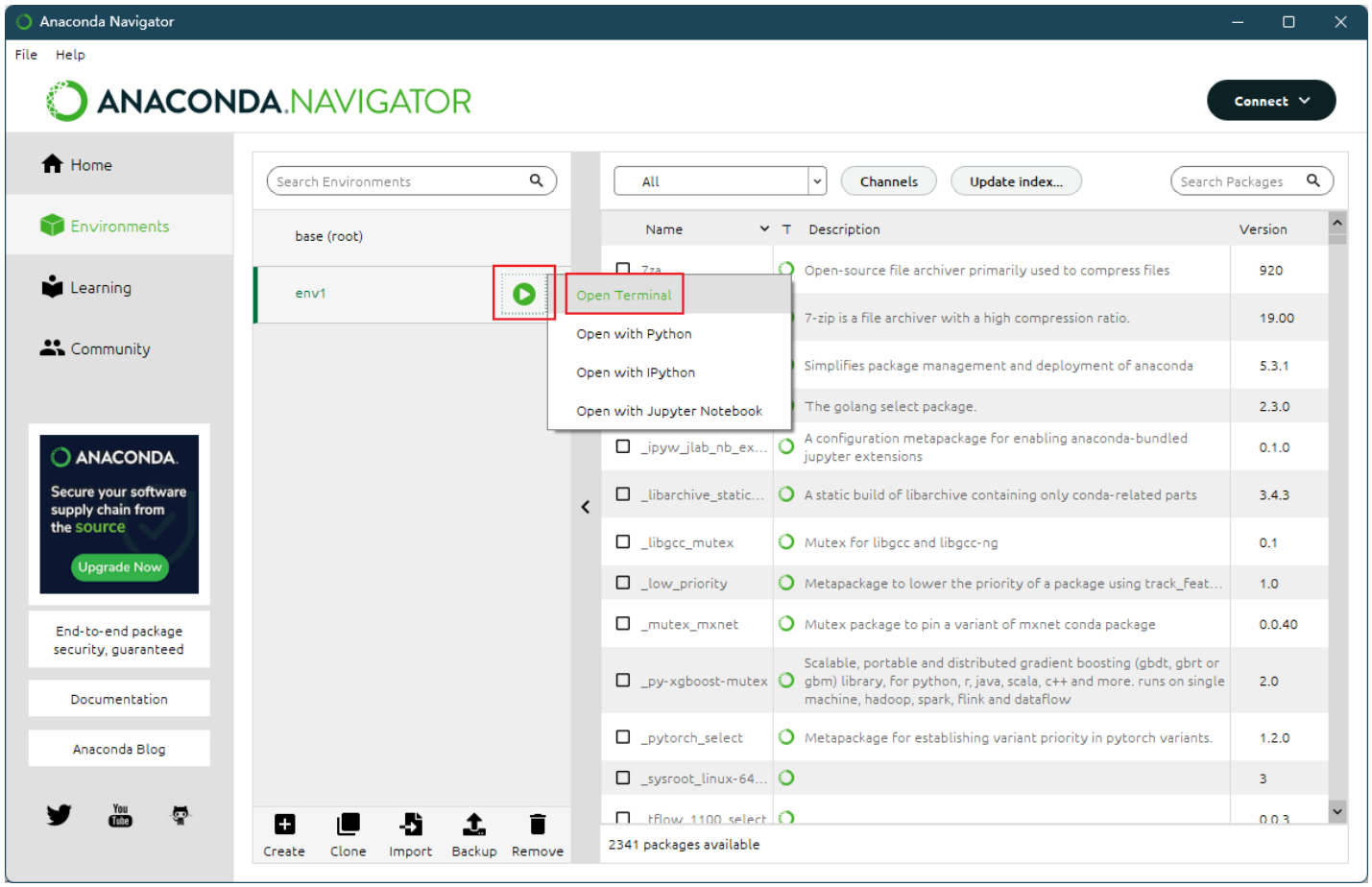


安装PyTorch

- 参考PyTorch的官方文档
- https://pytorch.org/get-started/locally

PyTorch Build	**Stable (1.11.0)**	Preview (Nightly)	LTS (1.8.2)	
Your OS	Linux	Mac	**Windows**	
Package	**Conda**	Pip	LibTorch	Source
Language	**Python**	C++ / Java		
Compute Platform	CUDA 10.2	CUDA 11.3	~~ROCM 4.5.2 (beta)~~	**CPU**
Run this Command:	`conda install pytorch torchvision torchaudio cpuonly -c pytorch`			

安装PyTorch


Anaconda Navigator
File
Help
ANACONDA.NAVIGATOR
Connect
Home
Environments
Learning
Community
ANACONDA.
Secure your software supply chain from the source
Upgrade Now
End-to-end package security, guaranteed
Documentation
Anaconda Blog
Search Environments
base (root)
env1
Open Terminal
Open with Python
Open with IPython
Open with Jupyter Notebook
Create
Clone
Import
Backup
Remove
All
Channels
Update index...
Search Packages
Name
T
Description
Version
Open-source file archiver primarily used to compress files
920
7-zip is a file archiver with a high compression ratio.
19.00
Simplifies package management and deployment of anaconda
5.3.1
The golang select package.
2.3.0
_ipyw_jlab_nb_ex...
A configuration metapackage for enabling anaconda-bundled jupyter extensions
0.1.0
_libarchive_static...
A static build of libarchive containing only conda-related parts
3.4.3
_libgcc_mutex
Mutex for libgcc and libgcc-ng
0.1
_low_priority
Metapackage to lower the priority of a package using track_feat...
1.0
_mutex_mxnet
Mutex package to pin a variant of mxnet conda package
0.0.40
_py-xgboost-mutex
Scalable, portable and distributed gradient boosting (gbdt, gbrt or gbm) library, for python, r, java, scala, c++ and more. runs on single machine, hadoop, spark, flink and dataflow
2.0
_pytorch_select
Metapackage for establishing variant priority in pytorch variants.
1.2.0
_sysroot_linux-64...
3
2341 packages available

安装PyTorch

C:\Windows\system32\cmd.e:

```
(env1) C:\Users\Jiash>conda install pytorch torchvision torchaudio cpuonly -c pytorch
Collecting package metadata (current_repodata.json): \
```

安装PyTorch

```
C:\Windows\system32\cmd.e:

  idna                pkgs/main/noarch::idna-3.3-pyhd3eb1b0_0
  intel-openmp        pkgs/main/win-64::intel-openmp-2021.4.0-haa95532_3556
  jpeg                pkgs/main/win-64::jpeg-9e-h2bbff1b_0
  libpng              pkgs/main/win-64::libpng-1.6.37-h2a8f88b_0
  libtiff             pkgs/main/win-64::libtiff-4.2.0-hd0e1b90_0
  libuv               pkgs/main/win-64::libuv-1.40.0-he774522_0
  libwebp             pkgs/main/win-64::libwebp-1.2.2-h2bbff1b_0
  lz4-c               pkgs/main/win-64::lz4-c-1.9.3-h2bbff1b_1
  mkl                 pkgs/main/win-64::mkl-2021.4.0-haa95532_640
  mkl-service         pkgs/main/win-64::mkl-service-2.4.0-py39h2bbff1b_0
  mkl_fft             pkgs/main/win-64::mkl_fft-1.3.1-py39h277e83a_0
  mkl_random          pkgs/main/win-64::mkl_random-1.2.2-py39hf11a4ad_0
  numpy               pkgs/main/win-64::numpy-1.21.5-py39h7a0a035_2
  numpy-base          pkgs/main/win-64::numpy-base-1.21.5-py39hca35cd5_2
  pillow              pkgs/main/win-64::pillow-9.0.1-py39hdc2b20a_0
  pyopenssl           pkgs/main/noarch::pyopenssl-22.0.0-pyhd3eb1b0_0
  pysocks             pkgs/main/win-64::pysocks-1.7.1-py39haa95532_0
  pytorch             pytorch/win-64::pytorch-1.11.0-py3.9_cpu_0
  requests            pkgs/main/noarch::requests-2.27.1-pyhd3eb1b0_0
  tk                  pkgs/main/win-64::tk-8.6.11-h2bbff1b_0
  torchaudio          pytorch/win-64::torchaudio-0.11.0-py39_cpu
  torchvision         pytorch/win-64::torchvision-0.12.0-py39_cpu
  urllib3             pkgs/main/win-64::urllib3-1.26.9-py39haa95532_0
  win_inet_pton       pkgs/main/win-64::win_inet_pton-1.1.0-py39haa95532_0
  xz                  pkgs/main/win-64::xz-5.2.5-h8cc25b3_1
  zlib                pkgs/main/win-64::zlib-1.2.12-h8cc25b3_2
  zstd                pkgs/main/win-64::zstd-1.4.9-h19a0ad4_0


Proceed ([y]/n)? y
```

基于PyTorch编写代码

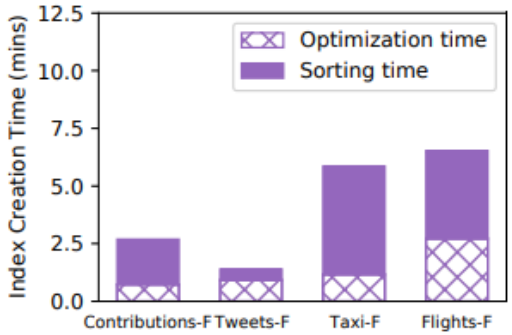
- 见neural-network-tutorial.ipynb

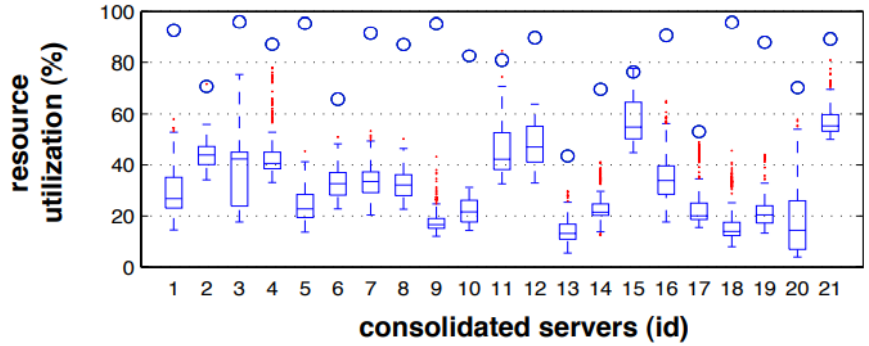
目录

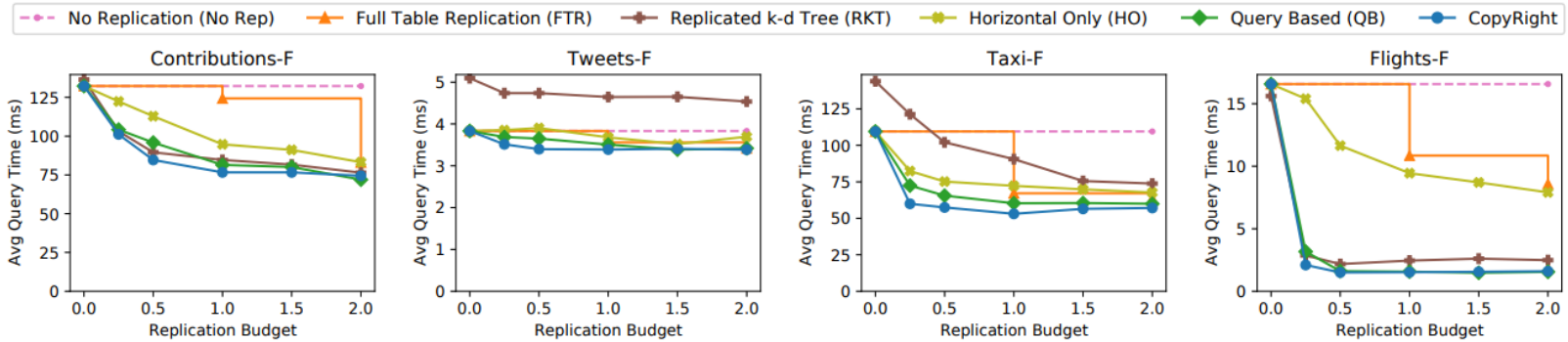
- 简介
- 安装Python环境
- 案例1：使用scikit-learn分析新冠数据
- 案例2：使用PyTorch搭建神经网络
- 案例3：使用Matplotlib进行数据可视化

科研中的绘图需求

- 大量的点图，线图和柱图







Python绘图

- 在Python中，我们可以使用Matplotlib和Seaborn等第三方数据可视化库

- 见plot-tutorial.ipynb

谢谢大家