**Python数据杂谈**

1. **Python开发环境**

**1.1python解释器**

Python作为动词，它是解释器，是一种能够执行Python代码的程序。它**读取和解释Python源代码，并将其转换为计算机可以执行的指令**。Python解释器可以理解和执行Python语法和语义，从而实现程序的运行。

Python解释器有多种不同的实现，其中最常见的是CPython，它是官方的Python解释器。CPython是用C语言编写的，它解释和执行Python代码，并**将其转换为字节码形式**，然后通过解释器**循环逐条执行字节码**。

除了CPython之外，还有其他的Python解释器，如Jython、IronPython和PyPy等。这些解释器在实现上有所不同，但它们都能够解释和执行Python代码。

Python解释器的存在使得开发人员能够快速编写和调试代码，并提供交互式编程的能力。开发人员可以通过命令行界面或集成开发环境（IDE）来使用Python解释器。无论使用哪种Python解释器，Python的语法和语义在不同的解释器之间是一致的，因此代码可以在不同的解释器中进行移植和执行。

**1.1.1Python解释器的下载和安装**



**安装路径：C:\Python312\python.exe**

**一台机器可以安装多个不同版本的Python解释器，比如，我可以安装Python2.7也可以安装Python3.10。**

如果在控制台输入python命令后，显示找不到命令或者不是内部命令等情况，大部分的原因是因为在第一步时没有勾选上Add Python 3.X to PATH，这个时候我们需要手动的添加环境变量：

右键单击桌面 我的电脑/此电脑-->属性-->找到高级系统设置-->环境变量-->下面的系统变量-->找到 Path

添加python安装的路径（C:\Python312）和Scripts路径（C:\Python312\Scripts）即可。

**1.1.2Python环境**

当你安装好了一个**特定版本的Python解释器**，你就建立了一个**基础**的Python环境（包括Python解释器以及相关的库和工具），下一步你可以在这个基础环境下安装依赖于这个Python解释器的其他第三方包来完成特定的项目任务。

但是在日常的工作中，我们经常会遇见这样的场景：

**1、各个项目使用的python版本不相同**

由于Python的解释器版本众多，各版本之间差异非常大。特别是python2和python3，互不兼容。有些项目可能用的python2.7，有些项目可能用的是python3.6，有些则使用的3.8等，但是它们却需要运行在同一个服务器环境中。（docker除外，docker容器可以隔离不同的项目环境。）

**2、系统依赖自带的解释器**

系统的一些服务组件一般也会依赖Python环境。不同的Linux发行版自带的Python也不同。如ubuntu16自带2.7和3.5版本, Centos7依赖python2.7。而系统很多组件都依赖自带的解释器，比如yum等，你不能轻易删除这个版本，一旦删除或者更改都可能造成系统出问题。

**3、依赖默认的解释器路径冲突**

比如Centos7系统自带的python是2.7，系统很多组件比如yum依赖的都是2.7这个版本，我们发现这些工具开头使用的都是：#!/usr/bin/python。而一些新的使用python开发的服务组件，它们依赖的却是python3.6以上的版本，但是它们一些代码开头用的也是这个引用：#!/usr/bin/python。它们都是用python这一个引用，却没有使用python2、python3这样分开，这就很容易导致它们的一些python引用冲突。

**4、依赖冲突（最常见）**

我们都知道python的软件包依赖经常是个很头疼的问题，经常因为这个问题导致到家在安装一些python环境或者服务组件时失败。

而不同的python解释器版本，对软件包依赖库的管理也是个问题。

比如sqlalchemy这个包，有些项目使用的python2.7版本，它需要依赖这个库，有些项目使用的python3.6版本，它也需要依赖这个库，有些项目使用的python3.8版本，它同样也需要依赖这个库，

但是头疼的是，这三者它们依赖的这个包版本还不一致。sqlalchemy从0.1-2.0有众多版本。

这时候如果你在系统上直接使用pip install sqlalchemy的话，它只能选择安装一个版本，但是这样其他两个项目是无法使用这个版本，就会出现依赖冲突的问题。由于 Python 的依赖库管理是中心化的，而且大版本上的不兼容且长期并行，就出现了这么一个独特的话题。

**1.1.3Python环境管理**

针对上述问题，我们可以基于这个版本的Python解释器（base ）创建新的环境（虚拟环境）来隔离不同的项目需求，即每一个项目基于base interpreter创建一个python环境（虚拟环境）。这里需要用到的两个工具：

* Venv：

python3.3版本之后自带的模块，只支持3.3版本之后的，不支持2.x。

* Virtualenv：

需要自己安装：pip3 install virtualenv，同时支持 python2 和 python3，并可以选择继承基础版本的包。

终端下：

**（1）进入到项目目录下D:\myProject**

**- Python自3.3版本之后，官方自带了用于创建虚拟环境的venv模块**

**- python -m venv [虚拟环境名称]**

**- Python自3.3版本之前，**

**- virtualenv [虚拟环境名称]**

**- 创建完虚拟环境后，会在项目文件夹下生成一个虚拟环境文件夹D:\myProject\venv，解释器的位置D:\myProject\venv\Scripts\python.exe**

* **激活虚拟环境D:\myProject\venv\Scripts下运行activate.bat**
* **安装依赖pip install <package\_name>，安装的库存在D:\myProject\venv\Lib\site-packages**

**当然，也可以在Pycharm中直接为项目创建一个基于**base interpreter的python环境**。**

**两者共同存在的最大问题：**

Venv/virtualenv可以为每个虚拟环境指定 python 解释器，但是指定的只能是当前系统已经安装好的python解释器，无法使用其他版本的python解释器。

**1.2anaconda**

尽管Python解释器本身就可以执行Python代码，Anaconda是一个用于科学计算和数据科学的Python发行版，它包含了Python解释器以及大量的科学计算和数据科学相关的库、工具和软件包。Anaconda的目标是简化Python环境的设置和管理，并提供一个集成的平台，使得科学计算和数据科学的工作更加方便和高效。

Anaconda的主要特点包括：

（1）管理Python环境：Anaconda提供了一个命令行工具conda，可以用于创建、管理和切换**不同的Python环境。这使得开发人员可以在同一台机器上同时使用不同版本的Python和库，而不会相互干扰**。

（2）预安装科学计算库：Anaconda预先安装了许多常用的科学计算和数据科学库，如NumPy、SciPy、Pandas和Matplotlib等。这些库对于数据处理、数值计算、统计分析和可视化等任务非常有用，使用Anaconda可以避免手动安装和配置这些库的麻烦。

（3）跨平台支持：**Anaconda可在多个操作系统上运行，包括Windows、macOS和Linux。这使得开发人员可以在不同的平台上保持一致的开发环境**，并轻松共享代码和项目。

（4）社区支持和扩展：Anaconda拥有庞大的用户社区和生态系统，用户可以从社区中获取支持、分享经验和获取扩展库和工具。Anaconda还提供了一个集成的包管理系统，可以方便地安装、更新和管理库和软件包。

**1.2.1 anaconda的下载与安装**

**1.2.2 conda环境**

Conda环境是指由Anaconda或Miniconda（Anaconda的轻量级版本，只包含最基本的组件）软件包管理器创建和管理的**独立Python环境**。

Conda环境允许你在同一台机器上同时管理和使用多个独立的Python环境。每个环境都可以具有不同的Python版本和不同的包集合，这使得你可以更好地管理和组织你的Python项目，确保项目之间的隔离性和依赖关系的一致性。

通过创建Conda环境，你可以：

（1）管理Python版本：你可以创建一个新的Conda环境，并在其中安装特定版本的Python解释器。这使得你可以在不同的项目中使用不同的Python版本而不会相互干扰。

（2）管理包依赖关系：每个Conda环境都可以拥有自己独立的包集合，你可以在每个环境中安装所需的软件包和库，而不会与其他环境产生冲突。这样可以确保每个项目使用的库版本是一致的，避免了包之间的冲突和版本不兼容的问题。

（3）共享环境配置：你可以通过创建和共享environment.yml文件来轻松重现和共享Conda环境的配置。这个文件记录了环境中安装的所有包及其版本，使得其他人可以轻松地创建相同的环境。

当安装好anaconda之后，你就拥有了一个基础conda环境，目录通常在D:\software\Anaconda3\python.exe，下一步你可以借助conda命令工具在这个环境下安装这个conda环境需要的其他包，同时，你也可以**指定python版本**建立一个新的conda环境，这个环境的目录通常在D:\software\Anaconda3\envs\env\_name\python.exe来隔离不同项目的需求。在这个python版本解释器下，也可以基于这个python版本的解释器进一步为项目创建虚拟环境，在项目文件夹下生成一个虚拟环境文件夹D:\myProject\venv，这个虚拟环境不继承conda，只继承了python解释器的版本。

**1.2.3 利用conda来管理python环境**

Anaconda Prompt 是 Anaconda 发行版中提供的一个命令行界面，用于管理和运行 Python 环境。它基于 Windows 的命令提示符（Command Prompt）或者 macOS/Linux 的终端，提供了一些额外的功能和命令，方便用户在 Anaconda 环境中进行 Python 开发和数据科学工作。

使用 Anaconda Prompt，你可以执行以下操作：

（1）创建和管理 Python 环境：Anaconda Prompt 允许你创建独立的 Python 环境，每个环境可以拥有不同的 Python 版本和依赖包，方便项目之间的隔离和管理。

（2）安装和更新依赖包：通过使用 conda 命令，你可以方便地安装、更新和删除 Python 包和依赖项。conda 是 Anaconda 的包管理工具，类似于 pip，但它可以处理更复杂的环境和依赖关系。

（3）启动 Jupyter Notebook：Anaconda Prompt 提供了命令来启动 Jupyter Notebook，这是一个流行的交互式开发环境，用于编写和运行 Python 代码、创建数据分析报告等。

（4）管理 Anaconda 发行版：你可以使用 Anaconda Prompt 来管理 Anaconda 发行版本身，包括更新 Anaconda、安装额外的包和工具、配置环境变量等。

**谈一谈几种环境的区别：**

1. **Virtualenv Environment**



**b．Conda Environment**



**c．System Interpreter**



**d. Pipenv Environment**



**e. Poetry Environment**



Pycharm版本过低可能无法识别高版本的python。报错pycharm unsupported Python 3.1**。**

**1.3运行python代码的方式**

运行Python程序常见的方式包括交互式运行（Python解释器运行）、命令行运行（脚本运行）、集成开发环境运行。

**1.3.1交互式运行**

Python 是一门解释型语言，在交互式环境中，可以逐行或逐个代码块地执行Python代码，并立即查看结果。这种方式适用于实验、调试和快速尝试代码片段等情况。在命令行中直接运行 python 命令，进入Python解释器的交互式模式。然后，你可以**逐行输入和执行**代码。这种方式适用于简单的代码片段和交互式开发。

**1.3.2命令行运行**

在命令行终端中，通过运行Python解释器并指定**代码文件**来执行Python代码。这种方式适用于运行较大的代码文件或批处理任务。这里的代码文件通常是指脚本文件。接下来说一说，同样都是python的代码文件，**脚本与模块的有区别？如何区分？**

**脚本（Script）：**

* 脚本是一段可执行的代码，通常用于完成特定的任务或解决特定的问题。
* 脚本通常以 .py 扩展名保存，并且可以在命令行或脚本解释器中直接执行。
* 脚本可以包含一系列的语句和函数调用，用于实现特定的功能。
* 脚本可以独立运行，并且可以接受命令行参数作为输入。
* 脚本通常被用于自动化任务、批处理处理、快速原型开发等场景。

**模块（Module）：**

* 模块是一个包含了一组相关函数、类和变量的文件，它可以被其他程序导入和使用。
* 模块通常以 .py 扩展名保存，并且可以在其他Python程序中通过 import 语句导入。
* 模块可以提供可重用的代码，可以被多个程序共享和调用。
* 模块可以定义函数、类、变量等，并且可以通过命名空间进行访问。
* 模块可以被组织成包（Package），形成更大的代码组织结构。

脚本和模块之间并没有严格的界限，一个脚本文件也可以作为一个模块被导入和使用。这取决于文件的用途和使用方式。可以通过 if \_\_name\_\_ == "\_\_main\_\_": 判断语句，我们可以确定代码是作为脚本执行还是作为模块导入。

在Python中，每个模块都有一个特殊的全局变量\_\_name\_\_，它存储了模块的名称。当模块直接运行时，\_\_name\_\_的值为\_\_main\_\_；当模块被导入时，\_\_name\_\_的值为模块的名称。

**1.3.3IDE / 代码编辑器运行**

**1.3.3.1 Pycharm**

PyCharm是一款由JetBrains开发的集成开发环境（IDE），专门用于Python开发。它提供了丰富的功能和工具，以提高开发效率并简化Python**项目的管理**。

**1、新建一个项目**

**2、两种运行模式**

在PyCharm中，你可以选择两种运行模式：Run模式和Debug模式。这两种模式在运行和调试代码时有一些区别。

**Run模式**：

Run模式是默认的运行模式。当你选择Run模式时，PyCharm会执行你的代码，并在控制台或输出窗口显示结果。

在Run模式下，你可以通过设置断点来暂停代码的执行，以便检查变量的值和程序的状态。

在Run模式下，你不能逐行调试代码，也不能查看代码的执行流程。

**Debug模式**：

Debug模式用于逐行调试代码，以便更详细地检查代码的执行过程和变量的值。

在Debug模式下，你可以设置断点，并逐步执行代码，观察每一行的执行结果。

在Debug模式下，你可以查看变量的值、观察表达式的结果，并使用调试工具来分析代码的执行流程。

Debug模式通常用于解决代码中的错误、调试复杂的逻辑或查找程序的性能问题。

**3、Terminal（终端命令行控制台）**

**4、Python console（交互式的Python解释器环境）**

**5、常用的快捷键**

**1.3.3.2 Jupyter Notebook**

Jupyter 是一个开源的交互式计算环境，它支持多种编程语言，包括 Python、R、Julia 等。Jupyter通过 Jupyter Notebook 提供了一个基于 Web 的界面，可以创建、编辑和运行 Notebooks。Notebooks是一个包含可执行代码、富文本元素（如说明文档、图表和公式）以及展示结果的交互式环境，你可以在 Jupyter Notebook 中编写和运行 Python 代码。

**启动 Jupyter Notebook**

**修改路径**

**快捷键**

**导出PDF**

1. **Python高阶语法**

Python作为名词，是一种高级编程语言，由Guido van Rossum于1991年创建。它被设计成易于阅读和理解的语言，具有清晰简洁的语法和强大的功能。Python具有广泛的应用领域，包括软件开发、数据分析、人工智能、科学计算、网络编程等。

Python具有许多特点，其中包括：

（1）简洁易读：Python的语法简单明了，使用空格缩进来表示代码块，使得代码易于阅读和理解。

（2）跨平台性：Python可以在多个操作系统上运行，包括Windows、macOS和Linux等。

（3）大量的库和框架：Python拥有丰富的标准库和第三方库，提供了各种功能和工具，使开发人员能够更快速地开发应用程序。

（4）动态类型：Python是一种动态类型语言，不需要显式声明变量的类型，可以根据赋值自动推断类型。

（5）面向对象编程：Python支持面向对象编程，可以使用类和对象来组织和管理代码。

（6）可扩展性：Python可以通过C/C++编写扩展模块，提高性能并与其他语言进行交互。

**2.1数据类型**

**2.1.1 基本数据类型**

Python常用的基本数据类型有7种：数字(Number)、布尔(Booleans)、字符串(String)、列表(List)、元组(Tuple)、字典(Dictionary)、集合(Set)。

**2.1.1.1数字(Numbers)**

在Python中，有几种常见的数字数据类型，包括：

* 整数（int）：整数是没有小数部分的数字，可以是正数、负数或零。例如：-5、0、10等。
* 浮点数（float）：浮点数是带有小数部分的数字，可以是正数、负数或零。浮点数使用小数点来表示。例如：3.14、-0.5、1.0等。
* 复数（complex）：复数是由实数部分和虚数部分组成的数字。复数的虚数部分用 "j" 或 "J" 表示。例如：2+3j、-1+2j等。

下面是一些使用数字数据类型的示例：

1. # 整数
2. x = 10
3. y = -5
4. z = 0
6. # 浮点数
7. a = 3.14
8. b = -0.5
9. c = 1.0
11. # 复数
12. d = 2 + 3j
13. e = -1 + 2j
15. # 数值计算
16. sum = x + y  # 加法
17. difference = x - y  # 减法
18. product = x \* y  # 乘法
19. quotient = x / y  # 除法
20. power = x \*\* 2  # 幂运算
22. # 内置数学函数
23. import math
25. sqrt = math.sqrt(x)  # 平方根
26. abs\_value = abs(y)  # 绝对值
27. ceil\_value = math.ceil(a)  # 向上取整
28. floor\_value = math.floor(a)  # 向下取整

这些数字数据类型在Python中是内置的，你可以直接使用它们进行数值计算和操作。Python还提供了丰富的数学运算和函数库，可以对数字进行各种数值计算、转换和处理。

**2.1.1.2布尔(Booleans)**

在Python中，布尔（Booleans）是一种表示真（True）或假（False）的数据类型。布尔值用于条件判断和逻辑运算。

在Python中，布尔类型有两个预定义的值：True（真）和False（假）。注意，首字母必须大写，小写的true和false是无效的。

下面是一些使用布尔类型的示例：

1. x = True
2. y = False
4. print(x)  # 输出：True
5. print(y)  # 输出：False
7. # 布尔运算
8. a = 5
9. b = 10
11. print(a > b)  # 输出：False，因为5不大于10
12. print(a < b)  # 输出：True，因为5小于10
14. # 逻辑运算
15. p = True
16. q = False
18. print(p and q)  # 输出：False，因为p和q都是False
19. print(p or q)  # 输出：True，因为p是True
20. print(not p)  # 输出：False，因为p是True，取反后为False

**2.1.1.5元组(Tuples)**

在Python中，元组（tuple）是一种**有序、不可变**的数据类型。**元组允许重复元素，每个元素可以是任意类型**。元组一般使用圆括号 () 将元素括起来，并用逗号分隔。

有序：元组的元素可以通过索引访问，索引从0开始。也可以使用切片（slice）操作获取元组的子集

不可变：不能对元组的元素进行赋值操作，不能动态地添加、删除元素。

**创建方法**

1. # 创建一个空元组
2. empty\_tuple = ()
4. # 创建一个包含多个元素的元组
5. fruits = ('apple', 'banana', 'orange')
7. # 创建一个包含不同类型元素的元组
8. person = ('John', 25, 'USA')
10. # 创建只包含一个元素的元组
11. single\_element\_tuple = (42,)
13. # 创建元组时可以省略括号
14. coordinates = 1, 2, 3

常规操作

1. fruits = ('apple', 'banana', 'orange')
3. # 访问元组的元素
4. print(fruits[0])  # 输出：'apple'
5. print(fruits[1])  # 输出：'banana'
7. # 使用切片获取元组的子集
8. print(fruits[1:])  # 输出：('banana', 'orange')
10. # 尝试修改元组的元素（会引发错误）
11. fruits[0] = 'pear'  # 引发 TypeError: 'tuple' object does not support item assignment

应用场景

* 当你希望创建一个不可变的序列时，可以使用元组。
* 当你希望将多个值作为一个整体传递给函数时，可以使用元组。

**2.1.1.6集合(Sets)**

在Python中，集合（Sets）是一种**无序、可变**的数据集合。它是由一组唯一的元素组成（不可重复）。**集合中的元素必须是可哈希的，也就是说，集合中的元素必须是不可变的类型。**可以进行集合运算，如并集、交集、差集等**。**集合一般使用大括号{}将元素括起来，并用逗号分隔。

**创建方法**

**常规操作**

**应用场景**

**2.1.1.4列表(list)：**

在Python中，列表（list）是一种**有序、可变**的数据集合。**列表允许重复元素，可以包含任意类型的元素。**创建列表可以使用方括号 [] 或者使用 list() 函数。

创建方法

常规操作

应用场景

**2.1.1.7字典(Dictionaries)： 无序的键值对的组合{k:v}**

在Python中，字典（dictionary）是一种**无序的键值对**数据结构，用于存储和管理**具有唯一键的元素**。字典中的每个元素由一个键（key）和一个对应的值（value）组成。字典中的键必须是不可变的类型，通常使用字符串作为键。值可以是任意类型，包括数字、字符串、列表、字典等。创建字典可以使用花括号 {key:value} 或者使用 dict() 函数。

创建方法

常规操作

应用场景

**字典与json的区别**

- python中，json和dict非常类似，都是key-value的形式，而且json和dict也可以非常方便的通过dumps和loads互转，但是它们之间还是有区别的：

- json：是一种数据格式，是纯字符串。可以被解析成python的dict或其他形式。

- dict：是一个完整的数据结构，是对hash table这一数据结构的实现，是一套从存储到提取都封装好的方案。它使用内置的hash函数来规划key对应的value的存储位置，从而获得O(1)的数据读取速度。

**2.1.1.3字符串(Strings)**

在Python中，字符串是由字符组成的序列，可以包含字母、数字、符号和空格等字符，是一种表示文本数据的数据类型。创建字符串可以使用单引号 ' ' 或双引号 " " 将文本括起来。字符串也可以使用三重引号 ''' ''' 或 """ """ 来创建多行字符串。

创建方法

常规操作

应用场景

可哈希性：可哈希性是一个对象的特性，并且与对象的值相关。指一个对象具有哈希值（hash value），并且在其生命周期中保持不变。哈希值是一个整数，用于唯一标识一个对象。**可哈希的对象是不可变的**，常见的可哈希对象包括整数、浮点数、字符串、元组等。这些对象都具有唯一的哈希值，且在其生命周期中不会发生变化，两个具有相同值的字符串会有相同的哈希值。**不可哈希的对象是可变的**，不可哈希的对象包括列表（list）、字典（dictionary）和集合（set）等可变对象。这些对象在其生命周期中可以发生变化，因此它们没有固定的哈希值，两个具有相同值的列表会有不同的哈希值。

**2.1.2引申数据类型**

**2.1.2.1索引序列(Series)**

- 由index和value组成

- index是索引，是一个一维列表

- value是值，是一个一维列表

**2.1.2.2数据框(Dataframe)**

- 由index（索引）、colums（字段）、values（值）三部分组成

- index和colums是一个一维列表，values是一个二维列表

**2.1.2.3数组(np.array)**

**2.1.2.4张量**

**2.2魔法函数**

魔法属性和方法是Python内置的一些属性和方法，有着特殊的含义。命名时前后加上两个下划线，在执行系统特定操作时，会自动调用。用于定义对象的行为和特性。它们允许你在自定义类中重写和定义特定的行为，使你的类具有类似内置类型的行为，例如迭代、比较、字符串表示等。那么接下来，我们就列举一些常见的魔法属性和方法。

**2.3内置函数**

内置函数（Built-in functions）是 Python 语言提供的一组预定义的函数，可以直接在代码中使用，而无需导入其他模块。这些函数可以完成各种常见的操作，例如类型转换、数学计算、字符串处理、文件操作等。例如，len()、str()、print()、type() 等都是内置函数。

魔法属性和魔法方法是用于自定义类行为的特殊方法。它们是与对象相关的，而内置函数是独立于对象的通用函数。需要注意的是，有些内置函数可能与魔法方法的名称相同，例如len()函数与\_\_len\_\_魔法方法。但它们的作用和用法是不同的。内置函数是全局可用的函数，而魔法方法是在类定义中定义的特殊方法。

内置的 zip() 函数

any() 和 all() 函数

内置的 enumerate() 函数

**2.4 标准库**

**2.4.1 os**

OS（Operating System）是Python标准库中的一个模块，它提供了一种与**操作系统**进行交互的方式，可以进行文件和目录操作、进程管理、环境变量访问等。

**2.4.2 sys**

“sys”即“system”，“系统”之意。sys 模块主要负责与 **Python 解释器进行交互**，该模块提供了一系列用于控制 Python 运行环境的函数和变量。该模块提供了一些接口，用于访问 Python 解释器自身使用和维护的变量，同时模块中还提供了一部分函数，可以与解释器进行比较深度的交互。

**2.4.3 re**

Python的标准库re（正则表达式）模块提供了处理正则表达式的功能。它允许您在文本中进行模式匹配、搜索、替换和分割等操作。

下面是re模块中一些常用函数和方法的概述：

1. re.compile(pattern, flags=0)：将正则表达式模式编译为正则表达式对象，以便在后续的匹配操作中重复使用。
2. re.match(pattern, string, flags=0)：尝试从字符串的起始位置进行匹配。如果匹配成功，返回**一个**匹配对象；否则返回None。
3. re.search(pattern, string, flags=0)：在整个字符串中搜索，返回第一个符合要求的匹配。如果匹配成功，返回一个匹配对象；否则返回None。
4. re.findall(pattern, string, flags=0)：在整个字符串中搜索，并以列表形式返回满足要求的所有匹配。
5. re.sub(pattern, repl, string, count=0, flags=0)：使用替换字符串替换字符串中匹配模式的非重叠部分。
6. re.split(pattern, string, maxsplit=0, flags=0)：根据模式的匹配来分割字符串，并返回分割后的子串列表。

**- 案例**

PATTERN=r'(?:https?:\/\/)?(?:www\.)?(?:github|gitee|gitlab)\.com\/[\w\-]+\/[\w\-]+\/(?:pulls|pull|merge\_requests)\/[0-9]+'

+ ：匹配前面的字符一次或多次

?：匹配前面的字符出现0次或1次

\*：匹配前面的字符出现0次或多次

\w：匹配任意一个字母或数字字符

\s : 匹配任何空白字符

\d ： 匹配任何数字字符

(..)：匹配括号中的字符串

(?:) ：是一个非捕获组，表示匹配但不捕获括号中的内容

[...]：匹配方括号中的任意一个字符

|：或操作符

**2.5高阶用法**

**2.5.1 类的继承**

首先提一下，经典类和新式类。在Python2中，如果定义类的方式是 class MyClass: 那么该类叫做经典类，如果定义类的方式为class MyClass(object): 那么该类为新式类。在Python3中，上面两种方式定义出来的类都叫新式类。本文是基于新式类来进行讲解的。

**单继承**

**代码示例：**

1. class FooParent(object):
2. def \_\_init\_\_(self):
3. self.parent = 'I\'m the parent.'
4. print ('Parent')
6. def bar(self,message):
7. print ("%s from Parent" % message)
9. class FooChild(FooParent):
10. def \_\_init\_\_(self):
12. super(FooChild,self).\_\_init\_\_()
13. print ('Child')
15. def bar(self,message):
16. super(FooChild, self).bar(message)
17. print ('Child bar fuction')
18. print (self.parent)
20. if \_\_name\_\_ == '\_\_main\_\_':
21. fooChild = FooChild()
22. fooChild.bar('HelloWorld')

**代码解读：**

super(FooChild,self) 首先找到 FooChild 的父类（就是类 FooParent），然后把类 FooChild 的对象转换为类 FooParent 的对象

**多继承**

**1.4.1 装饰器**

Python装饰器（Decorators）是一种函数或类，用于修改其他函数或类的行为。装饰器提供了一种简洁的方式来对函数或类进行包装、扩展或修改，而无需修改它们的源代码。

在Python中，装饰器使用@符号紧跟在函数或类定义之前，并将装饰器函数或类应用于目标函数或类。装饰器可以接受目标函数或类作为参数，并返回一个新的函数或类，或者在内部修改目标函数或类。

（1）假设我们有一个原函数（目标函数）

（2）定义一个装饰器（是一个嵌套函数），使得原函数能够有处理异常的功能

（3）使用@符号+装饰器的名字来使用装饰器，可以在不改变原函数（divide）的情况下增加异常处理的功能

1. # 假设我有一个函数（原函数），可以实现两个数相除
2. def divide(x, y):
3. return x / y
5. res = divide(10, 2)
6. print(res)
8. # 定义一个装饰器，使得原函数能够有处理异常的功能
9. def handle\_exceptions(func):    # 外层函数：接受一个函数（原函数）作为参数，并返回一个函数（新函数）。
10. def wrapper(\*args, \*\*kwargs):  # 里层函数：接受和原函数相同的参数
11. try:
12. # 调用被装饰的函数
13. return func(\*args, \*\*kwargs)     # 将原来的函数包装在自己功能中，同时添加新的功能
14. except Exception as e:
15. # 处理异常的逻辑
16. # 返回异常信息
17. return e
18. return wrapper   # 新函数

21. # 使用@符号+装饰器的名字来使用装饰器，可以在不改变原函数（divide）的情况下增加异常处理的功能
22. @handle\_exceptions
23. def divide(x, y):    # 这是想要装饰的函数（原函数），即实现两个数相除
24. return x / y
26. # 也可以使用divide = handle\_exceptions(divide)来使用装饰器，可以在不改变原函数（divide）的情况下增加异常处理的功能
27. # divide = handle\_exceptions(divide)
29. res = divide(10, 0)
30. print(res)

**1.4.2 并发编程**

**计算密集型任务与I/O密集型任务：**

- IO密集型，涉及到网络、磁盘IO的任务都是IO密集型任务，这类任务的特点是CPU消耗很少，任务的大部分时间都在等待IO操作完成（因为IO的速度远远低于CPU和内存的速度）。对于IO密集型任务，任务越多，CPU效率越高，但也有一个限度。常见的大部分任务都是IO密集型任务。

- CPU密集型任务的特点是要进行大量的计算，消耗CPU资源，比如计算圆周率、对视频进行高清解码等等，全靠CPU的运算能力。这种计算密集型任务虽然也可以用多任务完成，但是任务越多，花在任务切换的时间就越多，CPU执行任务的效率就越低，所以，要最高效地利用CPU，计算密集型任务同时进行的数量应当等于CPU的核心数。

**多任务协同处理**

多任务在代码里面，就是一个程序有多个地方同时执行，多任务分为并行和并发，并行（真的多任务）,比如：有两个任务要同时进行，而cpu内核是4核的时候，这两个任务就会真正的同时分配给其中两个内核来完成。当任务数量大于内核数量的时候就不能这么干了，因为cpu内核为n时最多同时运行n个任务。这时候就变成了并发（假的多任务）由于cpu处理的速度很快，它采用一种轮转的方式。一个任务执行一段时间后扔一边又去执行另一个任务。这样来回转，雨露均沾。以达到类似于同时进行的效果

**并发(行)编程是实现多任务协同处理，改善系统性能的方式**。Python中实现并发编程主要依靠

* **进程(Process)**：进程是计算机中的程序关于某数据集合的一次运行实例，是操作系统进行资源分配的最小单位
  + 多进程编程：使用多个进程同时执行任务。每个进程都是独立的执行环境，可以并发地执行多个任务。在Python中，可以使用multiprocessing模块来创建和管理进程。
* **线程(Thread)**：线程被包含在进程之中，是操作系统进行程序调度执行的最小单位，同一个进程内的多个线程可以并发执行，分享进程的资源。
  + 多线程编程：使用多个线程同时执行任务。每个线程都可以独立执行代码，并且可以并发地执行多个任务。在Python中，可以使用threading模块来创建和管理线程。
* **协程(Coroutine)**：协程是用户态执行的轻量级编程模型，由单一线程内部发出控制信号进行调度
  + 基于协程的异步编程：使用协程来实现异步编程，通过挂起和恢复执行来实现并发。Python提供了asyncio库来支持协程异步编程。**协程可以在需要时挂起，让其他协程执行，从而实现并发执行**

## 二、Python中的线程

在Python中，我们可以使用threading模块创建和管理线程。以下是一个创建线程的例子：

**import** threading  
  
**def** **print\_numbers**():  
    **for** i **in** range(10):  
        print(i)  
  
**def** **print\_letters**():  
    **for** letter **in** 'abcdefghij':  
        print(letter)  
  
thread1 = threading.Thread(target=print\_numbers)  
thread2 = threading.Thread(target=print\_letters)  
  
thread1.start()  
thread2.start()  
  
thread1.join()  
thread2.join()

在这个例子中，我们创建了两个线程：thread1和thread2。thread1的任务是打印数字0到9，thread2的任务是打印字母'a'到'j'。 我们使用threading.Thread类创建了两个线程对象，并指定了每个线程的目标函数。然后，我们调用了start方法来启动线程。 join方法用于等待线程结束。这是阻塞调用，也就是说，调用join方法的线程（在这个例子中是主线程）会被阻塞，直到被调用join方法的线程（在这个例子中是thread1和thread2）结束。

## 三、线程同步

在多线程环境中，多个线程可能需要共享一些资源。例如，多个线程可能需要写入同一个文件或修改同一个内存位置。这可能会引发所谓的竞态条件，其中一个线程的操作可能会影响其他线程的结果。 为了避免竞态条件，我们需要进行线程同步。Python的threading模块提供了多种线程同步的工具，包括锁（Lock）、条件（Condition）、事件（Event）、信号量（Semaphore）等。以下是一个使用锁进行线程同步的例子：

**import** threading  
  
lock = threading.Lock()  
counter = 0  
  
**def** **increase\_counter**():  
    **global** counter  
    **with** lock:  
        counter += 1  
        print(counter)  
  
threads = []  
**for** \_ **in** range(10):  
    thread = threading.Thread(target=increase\_counter)  
    threads.append(thread)  
    thread.start()  
  
**for** thread **in** threads:  
    thread.join()

在这个例子中，我们创建了10个线程，每个线程都会增加共享变量counter的值。为了避免竞态条件，我们使用了一个锁来保护对counter的修改。只有获得了锁的线程才能修改counter，其他试图修改counter的线程将被阻塞，直到锁被释放。 with lock:语句用于自动获取和释放锁。当进入with语句时，锁被自动获取；当退出with语句时，锁被自动释放。这确保了即使在发生异常的情况下，锁也能被正确释放。

## 四、Python中的线程局限性

需要注意的是，由于Python的全局解释器锁（GIL）的存在，Python的线程并不能实现真正的并行计算。全局解释器锁是Python解释器级别的锁，用于同步Python字节码的执行。由于这个锁的存在，同一时间只有一个线程能够执行Python字节码。这意味着在计算密集型任务中，多线程可能并不会带来速度的提升。 然而，在I/O密集型任务中，多线程仍然可以有效地提高程序的性能。当一个线程在等待I/O操作完成（例如读取文件或网络请求）时，其他的线程可以继续执行，从而提高了程序的整体效率。

## 五、结论

Python的多线程是一种强大的工具，可以帮助我们编写出更高效的程序。然而，与所有并发编程一样，多线程编程也有其复杂性。为了避免竞态条件和其他并发问题，我们需要仔细地设计我们的程序，并正确地使用线程同步工具。 以上就是对Python多线程的基本介绍，希望对你有所帮助。

**1.4.3常用的语法糖**

Python 中的语法糖是指一些特殊的语法结构，它们并没有引入新的功能，但能够使代码更加简洁易读。

**with 语句：**

with open("file.txt", "r") as file:

    data = file.read()

这个语法糖可以自动管理文件等资源的打开和关闭，无需手动调用 file.close()。

上下文管理器类：

class MyContextManager:

    def \_\_enter\_\_(self):

        # 在进入 with 代码块前执行的操作

        return self

    def \_\_exit\_\_(self, exc\_type, exc\_val, exc\_tb):

        # 在离开 with 代码块后执行的操作

        pass

with MyContextManager() as obj:

    # 在 with 代码块中可以安全地使用 obj 对象

    pass

这个语法糖可以通过定义 \_\_enter\_\_() 和 \_\_exit\_\_() 方法，实现自己的上下文管理器，用于资源的安全管理。

with 语句的扩展：

from contextlib import contextmanager

@contextmanager

def open\_file(filename):

file = open(filename)

yield file

file.close()

with open\_file("file.txt") as file:

data = file.read()

这个语法糖可以通过使用 contextlib 模块中的 @contextmanager 装饰器将函数转换为上下文管理器，用于创建自定义的上下文管理器。

**try-except 块与异常处理：**

try:

    # 可能引发异常的代码块

    pass

except SomeException:

    # 处理某个特定异常

    pass

except AnotherException:

    # 处理另一个特定异常

    pass

else:

    # 当没有引发任何异常时执行

    pass

finally:

    # 无论是否引发异常都会执行

    pass

**可变参数与关键字参数：**

def func(\*args, \*\*kwargs):

    # 处理可变位置参数 args

    # 处理可变关键字参数 kwargs

    pass

func(1, 2, name="Alice", age=25)

这个语法糖可以在定义函数时接受可变数量的位置参数和关键字参数，方便灵活的函数调用。

**2.6编码规范**

1. **Python数据分析**

**3.1数据概述**

**3.1.1什么是数据？**

入门数据学科，首先第一步要认识数据什么，可能大多数人都无法对数据做一个准确的定义，在我们印象中，提到数据首先头脑浮现的是数据表格，是一堆堆数字，那么数据就是数字吗？显然二者不能完全划等号，我们来看一下度娘对数据定义的解读。

数据（data）是事实或观察的结果，是对客观事物的逻辑归纳，是用于表示客观事物的未经加工的的原始素材。

数据是信息的表现形式和载体，可以是文字、符号、数字、语音、图像、视频等。数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵，数据本身没有意义，数据只有对实体行为产生影响时才成为信息。

数据可以是连续的值，比如声音、图像，称为模拟数据（连续数据）。也可以是离散的，如符号、文字，称为数字数据（离散数据）。

在计算机系统中，数据以二进制信息单元0、1的形式表示。

通过以上解读，我们可以发现，我们印象中的数字只是数据的一种形式，广义的数据包含的内容很多，文字，语音，图像，视频都是数据定义的范畴，因为无论哪种形式，它们到计算机里面都是0和1，只是编码的形式有差别，这样说来，数据挖据，计算机视觉，自然语言处理等都是数据科学的范畴。

**3.1.2什么是数据类型？**

在学任何一门编程语言的开始，我们都要学习相关的数据类型，那么数据类型究竟又是如何定义的呢？

数据类型在数据结构中的定义是一组性质相同的值的集合以及定义在这个值集合上的一组操作的总称。

变量是用来存储值的所在处，它们有名字和数据类型。变量的数据类型决定了如何将代表这些值的位存储到计算机的内存中。在声明变量时也可指定它的数据类型，所有变量都具有数据类型，以决定能够存储哪种数据。

在学python语言时，有一个问题一直困扰着我，列表元组这些到底是数据类型还是数据结构呢？后来读了一些博客，发现数据类型主要包括原子类型和结构类型：

原子类型是指一种值的集合以及定义在值上的一组操作，比如在python中，有整数型(int)，浮点数型(float)，布尔型(bool)，字符串(str)等，这是我们印象中的常规表现形式；

结构类型（复合类型）是指一种数据结构以及定义在结构上的一种操作，是数据类型的另一种表现形式，是通过原子类型封装的更复杂的数据类型，比如在面向对象的编程语言python中，列表(List)，字典(Dict)，元组(Tuple)等都是这种类型。

**3.1.3什么是数据结构？**

从我的困惑可以看出，数据类型和数据结构之间存在联系但也有一定的区别。

数据结构是指相互之间存在着一种或多种关系的数据元素的集合和该集合中数据元素之间的关系组成 。包括逻辑结构和物理结构。

逻辑结构描述数据之间的相互关系，按照逻辑结构划分，主要分为线性结构和非线性结构，线性结构数据元素是一对一的关系，非线性结构数据元素是一对多和多对多的关系。

常见的线性结构有数组(Array)、链表(Linked List)、栈(Stack)、队列(Queue)，常见的非线性结构有树(Tree)，堆(Heap)，图(Graph)，散列表(Hash)。

物理结构描述数据具体在内存中的存储（如：顺序结构、链式结构、索引结构、哈希结构）等。

最后需要指出的是，数据类型和数据结构是基于某种编程语言的，不同的计算机语言（C、python、C#等）它们的数据类型和数据结构略有不同，但本质差异不大。

入门数据科学，第一步就是认识数据，了解数据类型和数据结构，数据类型和数据结构是依赖于编程语言的，对数据类型和数据结构有充分的认识，对以后数据分析和挖掘非常有帮助。

**3.2数据读取**

**3.2.1爬虫技术**

通过编写程序模拟人类浏览网页的行为，从网页上抓取数据。爬虫技术需要一定的编程能力和网络知识，但可以采集各种类型的数据，包括结构化数据和非结构化数据。

**3.2.2API接口调用**

许多网站和应用程序提供API接口，可以通过编程调用API接口获取数据。API接口调用需要一定的编程能力和对API接口的理解，但可以获得高质量的数据，且通常比爬虫技术更加稳定和可靠。

**3.2.3数据库查询**

许多数据存储在数据库中，可以通过查询数据库获取数据。数据库查询需要一定的SQL语言知识和对数据库的理解，但可以获得高质量的结构化数据。

**3.2.4文件导入**

许多数据存储在各种类型的文件中，可以通过导入文件的方式获取数据。文件导入需要对文件格式的理解和相应的工具支持，但可以获得各种类型的数据，包括结构化数据和非结构化数据。

导入数据

pd.read\_csv(filename)：从CSV文件导入数据

pd.read\_table(filename)：从限定分隔符的文本文件导入数据

pd.read\_excel(filename)：从Excel文件导入数据

pd.read\_sql(query, connection\_object)：从SQL表/库导入数据

pd.read\_json(json\_string)：从JSON格式的字符串导入数据

pd.read\_html(url)：解析URL、字符串或者HTML文件，抽取其中的tables表格

pd.read\_clipboard()：从你的粘贴板获取内容，并传给read\_table()

导出数据

df.to\_csv(filename)：导出数据到CSV文件

df.to\_excel(filename)：导出数据到Excel文件

df.to\_sql(table\_name, connection\_object)：导出数据到SQL表

df.to\_json(filename)：以Json格式导出数据到文本文件

**3.2.5传感器技术**

许多物理量和环境数据可以通过传感器等硬件设备采集。传感器技术需要相应的硬件设备和技术支持，但可以获得实时的、高精度的数据。

不同的数据采集方式适用于不同的数据来源和采集目的。在选择数据采集方式时，需要考虑数据来源、数据类型、数据质量、采集效率和采集成本等因素。

**3.1数据创建与插入**

**创建dataframe的几种方式**

**3.2数据删除**

**3.3数据修改**

**3.3.1替换**

**3.4数据查找**

**3.5数据聚合**

**3.6数据分组**

**3.7数据合并**

**3.8数据可视化**

1. **Python特征工程**

特征工程是一门将原始数据转换为有意义的变量并输入模型的艺术。它涉及选择、转换和创建新特征，为学习算法提供相关信息。领域知识在此过程中起着至关重要的作用，因为它指导最能指示目标变量的特征的选择。

**4.1缺失值处理**

**4.2异常值处理**

**4.3重复值处理**

**4.4标准化**

**4.5特征筛选（特征重要性分析）**

**4.6特征构造**

1. **Python机器学习**

机器学习（Machine Learning）是一种人工智能（Artificial Intelligence）的基础，它研究和开发使计算机能够自动学习和改进的算法和模型。机器学习算法通过从大量的训练数据中学习，自动发现数据中的模式和规律，并使用这些学习到的知识来进行预测和决策。机器学习算法的分类有多种，本文将其分为以下几个主要类型，并对该类型下常见的算法做一个简单的应用展示。



**5.1有监督模型**

有监督学习万能模板V1.0



有监督学习万能模板V2.0



有监督学习万能模板V3.0



**5.1.1 单模型**

**5.1.1.1线性模型**

- 线性回归（只能回归任务）

- 逻辑回归（只能分类任务）

- Lasso

- Ridge

- LDA

**5.1.1.2决策树模型**

**5.1.1.3支持向量机模型**

**5.1.1.4K近邻模型**

**5.1.1.5朴素贝叶斯（Naive Bayes）：只能用于分类任务**

**5.1.1.5多层感知机模型（放在下一章解释）**

**5.1.2 集成学习模型**

* **Boosting模型**
  + **AdaBoost**
    - **基于KNN**
    - **基于决策树**
  + **GBDT**
    - **Xgbboost**
    - **LightGBM**
* **Bagging模型**
  + **随机森林**

**5.2无监督模型**

**5.3概率模型**

sklearn安装：

pip install scikit-learn

1. **Python深度学习**

**6.1神经网络**

调用model.evaluate，可以直接得到loss与accuracy

调用model. predict，可以直接得到预测出的每个类的概率值

案例：结构化表格数据分类

**6.2卷积神经网络CNN**

**案例：图像识别**

**6.3循环神经网络RNN**

**案例：时序数据预测**

在现实世界的案例中，我们主要有两种类型的时间序列分析：

* **单变量时间序列**

对于单变量时间序列数据，我们将使用单列进行预测。只有一列，因此即将到来的未来值将仅取决于它之前的值。

* **多元时间序列**

在多元时间序列数据的情况下，将有不同类型的特征值并且目标数据将依赖于这些特征。，目标数据不仅取决于它以前的值，还取决于其他特征。因此，要预测即将到来的目标数据值，我们必须考虑包括目标列在内的所有列来对目标值进行预测。

**6.4 生成对抗网络GAN**

1. **模型部署**
2. **Python项目学习**

**8.1 学习别人的优秀GitHub项目**

7.1.1认识项目中的一些配置文件

7.1.2下载软件包（项目已经发布到Pypi了）

这种方式比较简单，只需要pip install 就可以跑这个项目，实现一些功能

7.1.1 git下载源文件

这种方法需要创建环境，然后安装项目需要的环境，

Requirement.txt，认识**Setup.py**

修改一些配置文件就可以运行

7.1.3 下载镜像文件（项目已经打包成docker镜像）

通过docker跑这个项目，认识**Dockerfile**

**8.2 打造自己的优秀GitHub项目**