



對外經濟貿易大學

University of International Business and Economics

信息学院

School of Information Technology & Management

Python与大数据分析

任课教师：袁石

办公地点：求索楼1006

办公电话：64491064

电子邮箱：ystone1025@uibe.edu.cn



- 大数据的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理和有效管理
- Python语言的简洁、易读以及可扩展性，非常适合用于大数据的处理和管理等工作
- 在国外用Python做科学计算的研究机构日益增多，一些知名大学已经采用Python教授程序设计课程。例如卡耐基梅隆大学的编程基础、麻省理工学院的计算机科学及编程导论就使用Python语言讲授
- Python语言是非常适合初学者学习的编程语言，并且也是大数据处理和首选工具



- 掌握Python的基本语法
- 熟练运用Python进行数据获取、处理与分析
 - ✓数据：数字、文本、图像
 - ✓数据获取：爬虫技术
 - ✓处理和分析方法：Numpy、Pandas、Sklearn、NLTK等
- 培养大数据处理和素养的素养：**发现数据的内在价值**





课程涵盖但不局限于以上内容



- 本课程授课周学时为2学时，总授课学时为32课时（按16周计）
- 理论课与上机实验课的课时比例约为1:1
- 考核方法：
 - ✓ 平时考勤+作业+期中编程考试 (50%)
 - ✓ 期末大作业 (50%)



- 《Python基础教程》 [挪]Magnus Lie Hetland著； 司维，曾军崑， 谭颖华译
- 其他参考书目：
 - ✓ 《Python for Data Analysis:利用Python进行数据分析》
[美]Wes McKinney著； 唐学韬
 - ✓ 《Python核心编程》（第2版） [美] 丘恩（Chun W.J.） 著； 宋吉广 译
 - ✓ 《Python从入门到精通》 明日科技编著
 - ✓ 《Python极简讲义 一本书入门数据分析与机器学习》 张玉宏著



- Python tutorial 3.6
<http://www.pythondoc.com/pythontutorial3/index.html>
- 廖雪峰-Python教程
<https://www.liaoxuefeng.com/wiki/0014316089557264a6b348958f449949df42a6d3a2e542c000>
- 探索 Flask-Python Web教程
<http://www.pythondoc.com/exploreflask/index.html>
- UIBE公开题集
http://lab.uibe.edu.cn/static/Public_PyTestExamples.zip
- Python 编程100例
<https://www.runoob.com/python/python-100-examples.html>



1、访问本课程教学管理平台：

tas1.uibe.edu.cn

可查看通知，在论坛提问，下载相关课件，查看作业要求
并按时上传作业

2、微信群

发布实时通知、与老师进行交流





對外經濟貿易大學

University of International Business and Economics

信息学院

School of Information Technology & Management

走进Python

任课教师：袁石

办公地点：求索楼1006

办公电话：64491064

电子邮箱：ystone1025@uibe.edu.cn



- 1、Python简介
- 2、Python开发环境

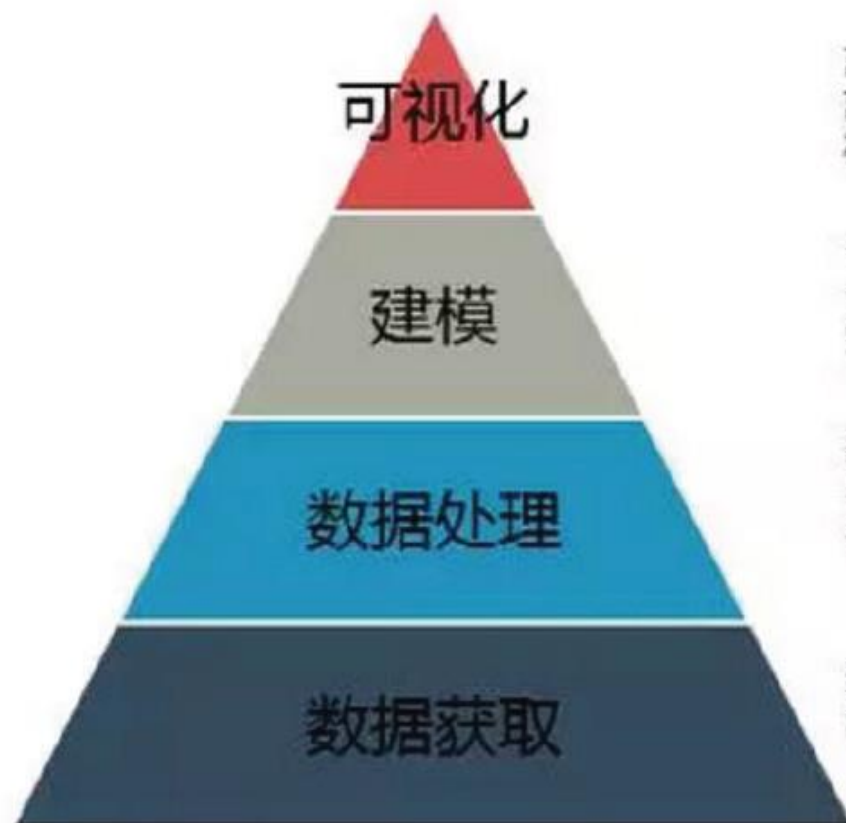




1、Python简介



为什么大数据首选是Python呢？



离线界面型：Excel，Powerpoint，PS，Tableau.....
离线代码型：R，SAS，Python，D3.js，Process.....
在线型：Echarts，Tagxedo.....

专用界面型：SPSS.....
专用代码型：R，SAS.....
泛用代码型：Python

数据库：SQL，Hive.....
界面型：Excel，SPSS.....
代码型：VBA，Python，R，SAS.....

数据库：SQL，Hadoop，Hive.....
爬虫：Python，Java.....

1/1 什么是Python

Python: 1989年由荷兰人Guido van Rossum (吉多·范·罗苏姆) 发明的一种面向对象的解释性高级编程语言

如何发明了Python? ? ?



中国Python程序员都叫龟叔

龟叔的愿望: 有一种编程语言

- ✓ C语言: 全面调用计算机的功能接口
- ✓ Shell: 可以轻松的编程

ABC语言让龟叔看到希望:

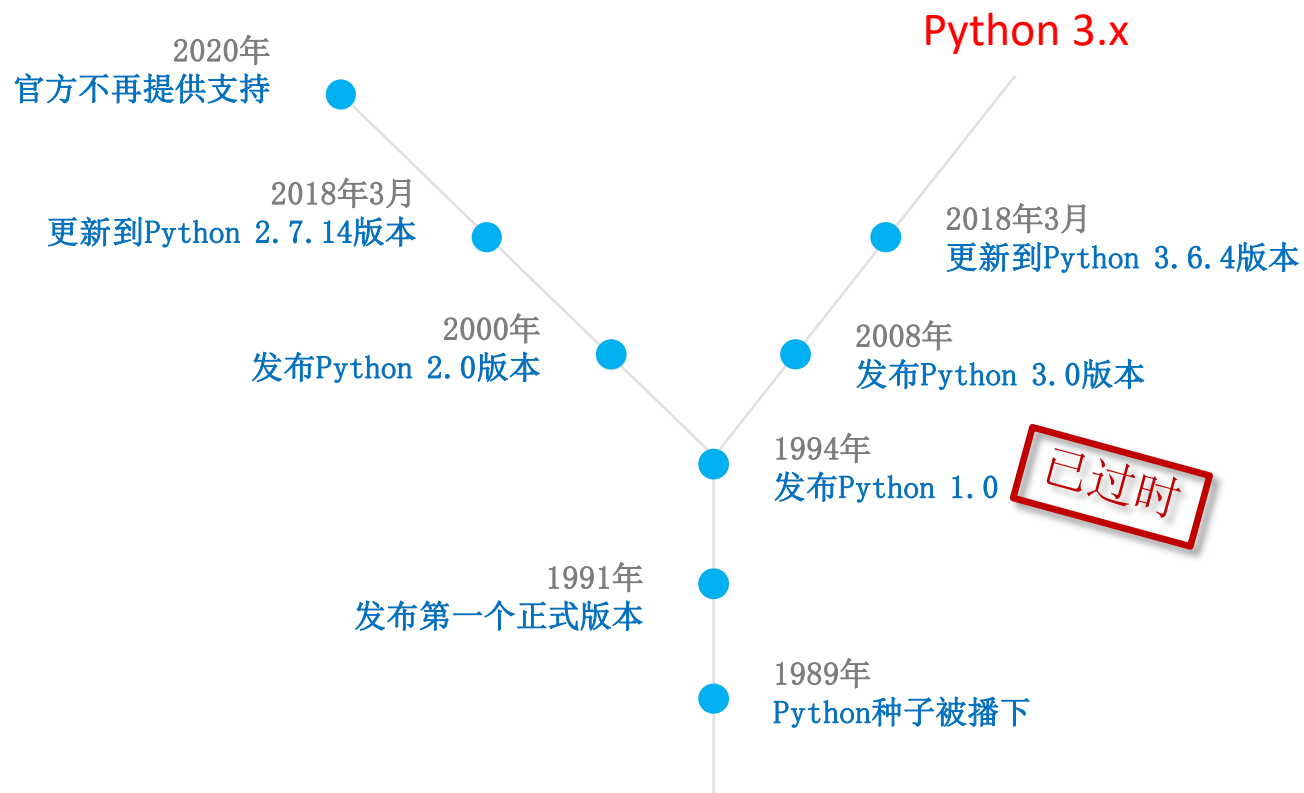
- ✓ ABC是由荷兰的数学和计算机研究所开发的
- ✓ ABC语言以教学为目的, 目标是“让用户感觉更好”
- ✓ ABC语言希望让语言变得容易阅读, 容易使用, 容易记忆, 容易学习, 并以此来激发人们学习编程的兴趣

1989年, 为了打发圣诞节假期, 龟叔开始写Python语言的编译器

- ✓ Python: 来自龟叔所挚爱的电视剧Monty Python's Flying Circus (巨蟒剧团的飞行的马戏团)



1/2 Python的发展历程



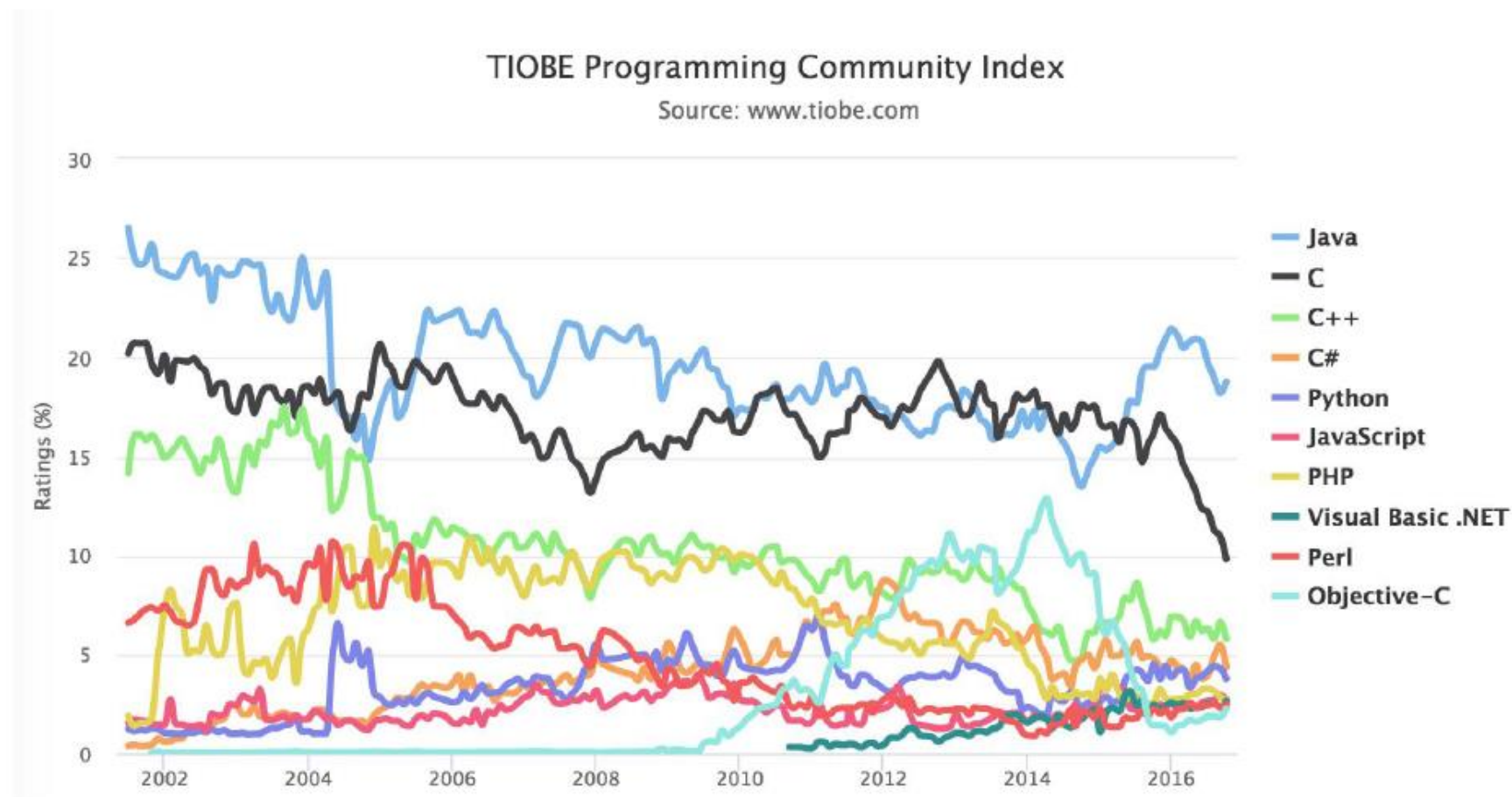
1991年，第一个Python编译器诞生

- ✓ 用C语言实现的，并能够调用C语言的库文件
- ✓ 已经具有了：类、函数、异常处理、包含表和字典在内的核心数据类型、以及模块为基础的拓展系统



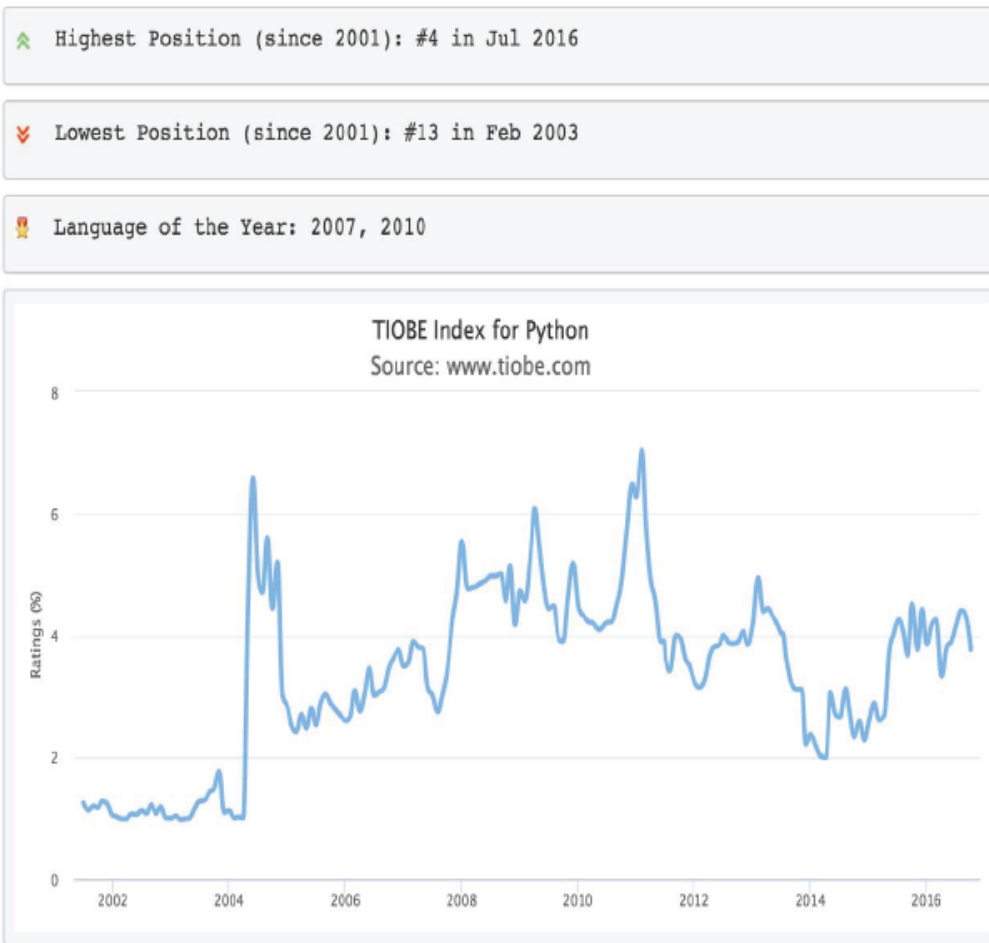
1/3 Python的流行度

- TIOBE INDEX:编程语言流行程度排行榜



1/3 Python的流行度

- Python曾在2007年和2010年两度被TIOBE排行榜评为“年度编程语言”
- 现已成为了第五大流行编程语言（截至2016年10月）



2016 Average Developer Salary in the U.S.

indeed.com
estimations(USD)

Language

#1 115,000

Swift

#2 107,000

Python

Ruby

#3 104,000

C++

#5 102,000

Java

C

#6 99,000

JavaScript

#7 94,000

C

#8 92,000

SQL

#9 89,000

PHP

made by codementor.io



1/3 Python的流行度



★PYPL编程语言排行榜2020年8月:

Worldwide, Aug 2020 compared to a year ago:

Rank	Change	Language	Share	Trend
1		Python	31.59 %	+3.3 %
2		Java	16.9 %	-2.7 %
3		Javascript	8.17 %	+0.0 %
4		C#	6.54 %	-0.7 %
5	↑	C/C++	5.88 %	+0.1 %
6	↓	PHP	5.78 %	-0.7 %
7		R	4.18 %	+0.3 %
8		Objective-C	2.6 %	-0.0 %
9		Swift	2.35 %	-0.0 %
10	↑	TypeScript	1.94 %	+0.2 %
11	↓	Matlab	1.63 %	-0.2 %
12		Kotlin	1.57 %	+0.1 %
13	↑↑	Go	1.39 %	+0.2 %
14	↓	Ruby	1.22 %	-0.2 %
15	↓	VBA	1.19 %	-0.1 %
16		Scala	0.97 %	-0.1 %
17	↑	Rust	0.91 %	+0.3 %
18	↓	Visual Basic	0.82 %	-0.2 %
19	↑↑↑↑↑	Dart	0.57 %	+0.2 %
20	↑↑↑	Ada	0.54 %	+0.2 %
21	↑	Lua	0.52 %	+0.1 %
22	↓↓↓	Perl	0.45 %	-0.1 %
23	↓↓↓	Abap	0.44 %	-0.1 %
24	↑↑↑	Julia	0.43 %	+0.2 %
25		Cobol	0.42 %	+0.1 %
26	↓↓↓↓↓	Groovy	0.41 %	-0.1 %
27	↓	Haskell	0.32 %	+0.0 %
28		Delphi	0.28 %	+0.0 %

© Pierre Carbonnelle, 2020

TIOBE 1 月编程语言：Python 摘得 2020 年度编程语言！



CSDN

发布时间：01-05 06:19 | 北京创新乐知信息技术有限公司

—|| CSDN ||—

【CSDN 编者按】恭喜 Python 荣获 2020 年度编程语言称号，这也是自 TIOBE 榜单发布以来，首款编程语言四次获得该奖项。

整理 | 苏宓

编程语言社区 TIOBE 最新发布了 1 月编程语言排行榜。这次揭晓了 2020 年度最受欢迎的编程语言，其中，**Python 以 2.01% 的正增长荣获 2020 年度 TIOBE 编程语言奖**！C++ 以微弱差距无缘冠军，凭借 1.99% 的增长率获得了亚军。其他编程语言中，C (+1.66%)、Groovy (+1.23%)、R (+1.10%) 分别位居其后。

作者最新文章

程序员为教师妻子开发专属应用；2020 最佳开源项目出炉；中国构建全球量子通信网 | 开发者周刊

中国芯片能不能弯道超车，就看量子计算了

为什么 Netflix 这么强？网飞 CEO 哈斯廷斯跟陆奇摊牌了

Python已经成为编程界的“明星”



1/4 Python的特点



- 可扩充的、具有丰富和强大的库
- “胶水”语言：把其他语言（尤其是C/C++）复杂的应用程序粘合在一起

优点

- ✓ 规范的代码、简单、易学、高层语言、跨平台（可移植性强）
- ✓ 支持面向对象编程、丰富的库等优点
- ✓ 开发效率较高

缺点

- ✓ Python语言的缺点就是执行效率慢，这个是解释型语言所通有的，同时这个缺点也将被计算机越来越强大的性能所弥补



1/4 Python的特点

- 优点一：优雅、简单、明确
(减少花哨、晦涩或以“炫技”为目的的代码)
- 让数据分析师们摆脱了程序本身语法规则的泥潭，更快的进行数据分析

C语言

```
#include<stdio.h>
main()
{
    printf("Hello World");
}
```

Python语言

```
print "Hello World"
```

Hello World



1/4 Python的特点

- 优点二：强大的标准库
- 完善的基础代码库，覆盖了网络通信、文件处理、数据库接口、图形系统、XML处理等大量内容，被形象地称为“内置电池”
(batteries included)
- Python使用者——“调包侠”

```
: import this
```

The Zen of Python, by Tim Peters

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one-- and preferably only one --obvious way to do it.
Although that way may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than *right* now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking great idea -- let's do more of those!

快速上手Python? ! 使用Python调包即可，轻松又愉快



1/4 Python的特点

- 优点三：良好的可扩展性
- 大量的第三方模块，覆盖了科学计算、Web开发、数据接口、图形系统等众多领域，开发的代码通过很好的封装，也可以作为第三方模块给别人使用。如Pandas、Numpy、Seaborn、Scikit-learn等等
- 优点四：免费、开源



1/4 Python的特点

- 缺点一：运行速度慢
- 缺点二：加密难
- 缺点三：缩进规则
- 缺点四：多线程灾难

```
if True:  
print "right"
```

```
File "<ipython-input-1-ab7d6f176ce6>", line 2
```

```
print "right"
```

```
IndentationError: expected an indented block
```

```
if True:  
    print "right"
```

```
right
```

```
junyi lu ... > misc > python > gil  
junyi lu ... > misc > python > gil  
junyi lu ... > misc > python > gil python2.7 ./single_thread.py  
Total time: 11.4724829197  
junyi lu ... > misc > python > gil python2.7 ./multi_thread.py  
Total time: 16.1935360432  
junyi lu ... > misc > python > gil
```





- ✓ 可以赋值给一个变量
- ✓ 可以添加到集合对象中
- ✓ 可以作为参数传递给函数
- ✓ 可以当做函数的返回值

- 动态类型和静态类型
- Python中一切皆对象
- 括号与缩进
- 应用领域





- 机器学习的一把利器
- 可读性强，便于上手
- 灵活性强：可与其他如Web应用程序进行整合



- 以统计推断为导向
- 数据分析之外的领域有所限制
- 包凌乱且一致性较差





- 网络爬虫
- 连接数据库
- 内容管理系统
- API构建



- 统计分析
- 交互式图标/面板

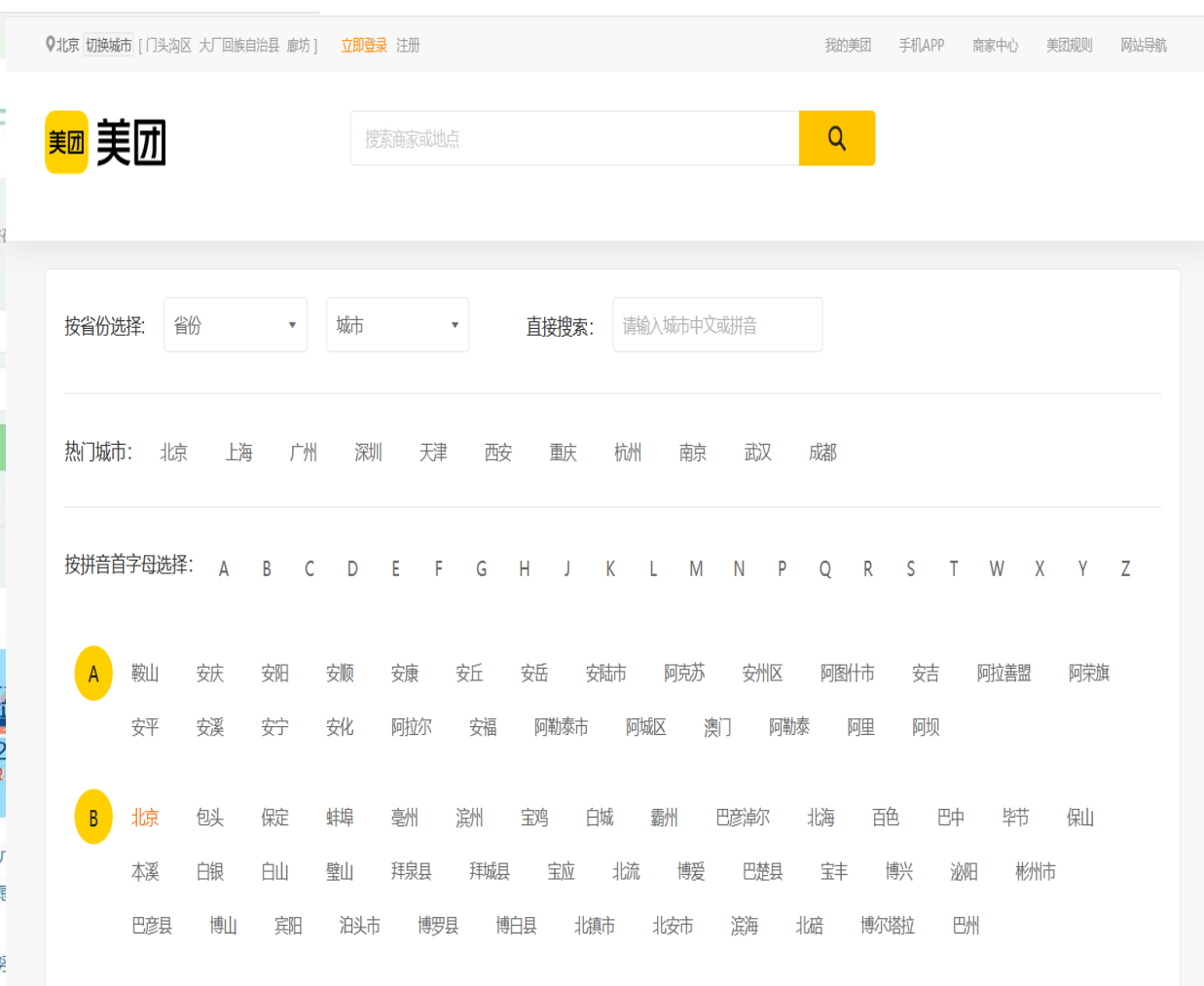
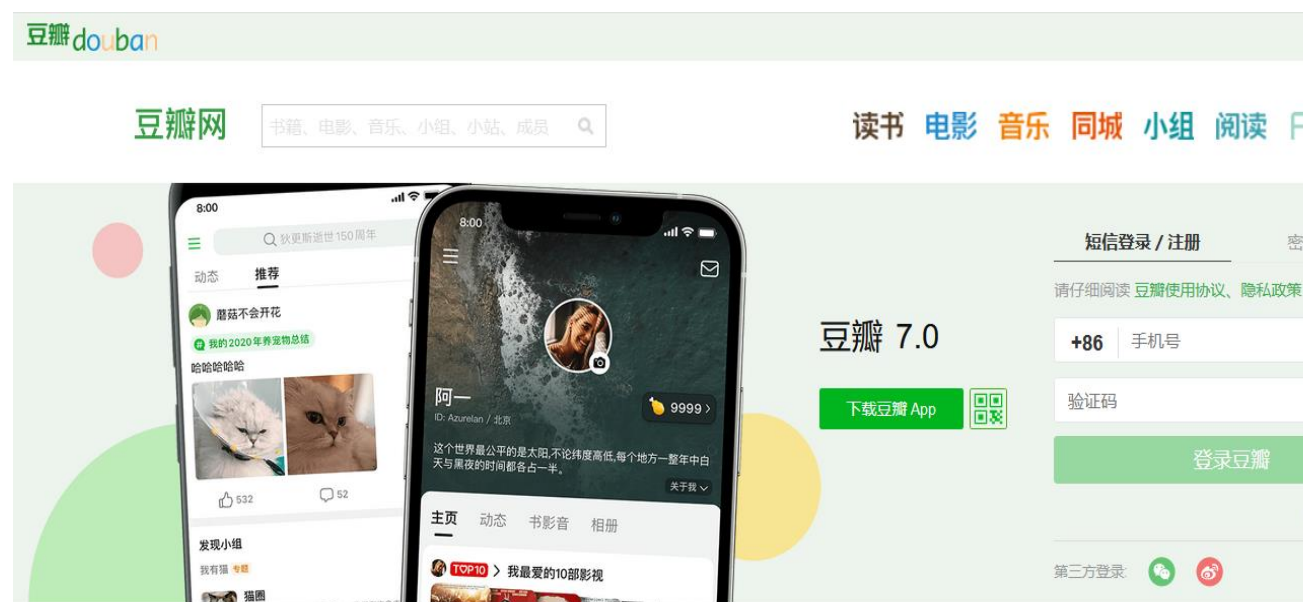


1/7 Python的应用



1、Web开发

■ 豆瓣、知乎、美团、饿了么、搜狐



1/7 Python的应用



2、大数据处理、人工智能

- 大数据处理平台：金融、量化交易
- 简单、快速、可应用多个人工智能框架



达内人工智能培训班:【Python+人工智能+大数据分析】



培训类型: **人工智能** 特色: 免费试学 班型: 业余/培优 适合人群: 零基础
达内**人工智能**培训班,Python+AI**人工智能**讲师8对一辅导,小班授课,**人工智能**培训班,可择业后付款模式,学习-认证-实训-择业一站式..

达内IT培训学校 2021-03 广告 保障

为您推荐: 学人工智能培训班 人工智能培训机构哪个好 python+人工智能

python报班一般多少钱 python在哪里学 python学费大约多少

千锋-人工智能培训班-中国高端IT人才培训机构-官方首页



人工智能培训班-中国高端IT人才培训机构-点击进入,互联网IT教育0元入学,2周免费试听,0基础总监级讲师授课!1对1就业服务跟踪就业服务!

课程价格: 千锋10周年活动 盛大开启 千万学费优惠 申请条件查询 更多>

课程简介: javascript Node.js React框架 小程序开发 原生APP开发 更多>



千锋10周年活动
报名优惠

查看更多相关信息>>



1.8万家合作企业
学长答疑



300讲师授课教学
校企合作

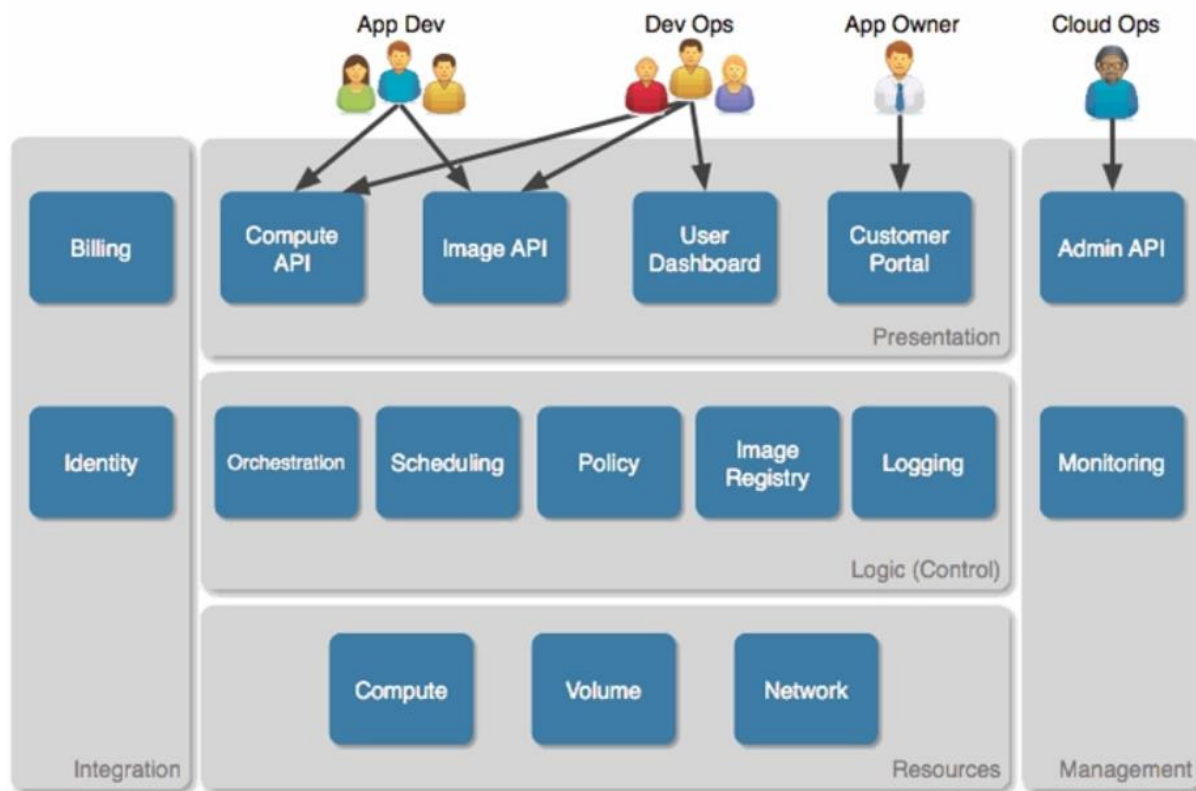


10周年学费优惠入口
14天试听



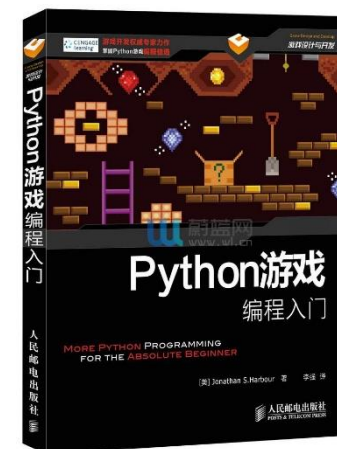
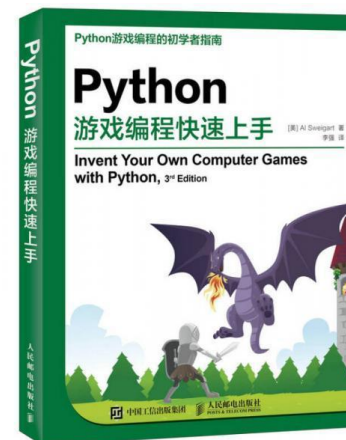
3、云计算、自动化运维开发、爬虫

- 云计算软件：OpenStack（开放协议栈，NASA公司开发）
- 满足大部分自动运维的需求：高级运维工程师的必备技能
- 网络爬虫：获取不同来源的数据集合



4、游戏开发

- 游戏程序
- 游戏脚本








2、Python开发环境



2/1 Anaconda介绍

- Anaconda是一个用于科学计算的Python发行版，支持Windows、maxOS和Linux三大系统
- Anaconda提供了强大且方便的类库管理（拥有超过1000个数据处理包）与环境（即包的依赖）管理功能，可以很方便地解决多版本Python并存、切换及各种第三方库的安装问题

Anaconda Installers

Windows 	MacOS 	Linux 
Python 3.8	Python 3.8	Python 3.8
64-Bit Graphical Installer (466 MB)	64-Bit Graphical Installer (462 MB)	64-Bit (x86) Installer (550 MB)
32-Bit Graphical Installer (397 MB)	64-Bit Command Line Installer (454 MB)	64-Bit (Power8 and Power9) Installer (290 MB)



类库的管理：conda

- ✓ 安装类库：conda install <类库名>
- ✓ 卸载类库：conda uninstall <类库名>
- ✓ 查看已安装类库：conda list

Python运行

1) Python解释器 输入“python”，启动解释器

```
(base) C:\Users\yston>python
Python 2.7.13 |Anaconda 4.3.1 (64-bit)| (default, Dec 19 2016, 13:29:36) [MSC v.1500 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
>>>
```

输入python语句

2) Python脚本文件：“.py”文件

- ✓ 命令行执行：python <文件名>.py

3) Jupyter Notebook Jupyter

- ✓ 代码的运行和结果展示不需要离开当前文档描述的平台



2/2 Anaconda使用



Anaconda Navigator

File Help

ANACONDA NAVIGATOR

Upgrade Now

Sign in to Anaconda.org

Home

Environments

Learning

Community

Applications on

base (root)

Channels

Refresh



console_shortcut

0.1.1

Console shortcut creator for Windows (using menuinst)

Launch



JupyterLab

0.35.4

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Launch



Notebook

5.7.8

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Launch



powershell_shortcut

0.0.1

Launch



Qt Console

4.4.3

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Launch



Spyder

3.3.3

Scientific PYTHON Development Environment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

Launch



Glueviz

1.0.0

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Install



Orange 3

3.26.0

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

Install



RStudio

1.1.456

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

Install



VS Code

1.53.2

Streamlined code editor with support for development operations like debugging, task running and version control.

Install

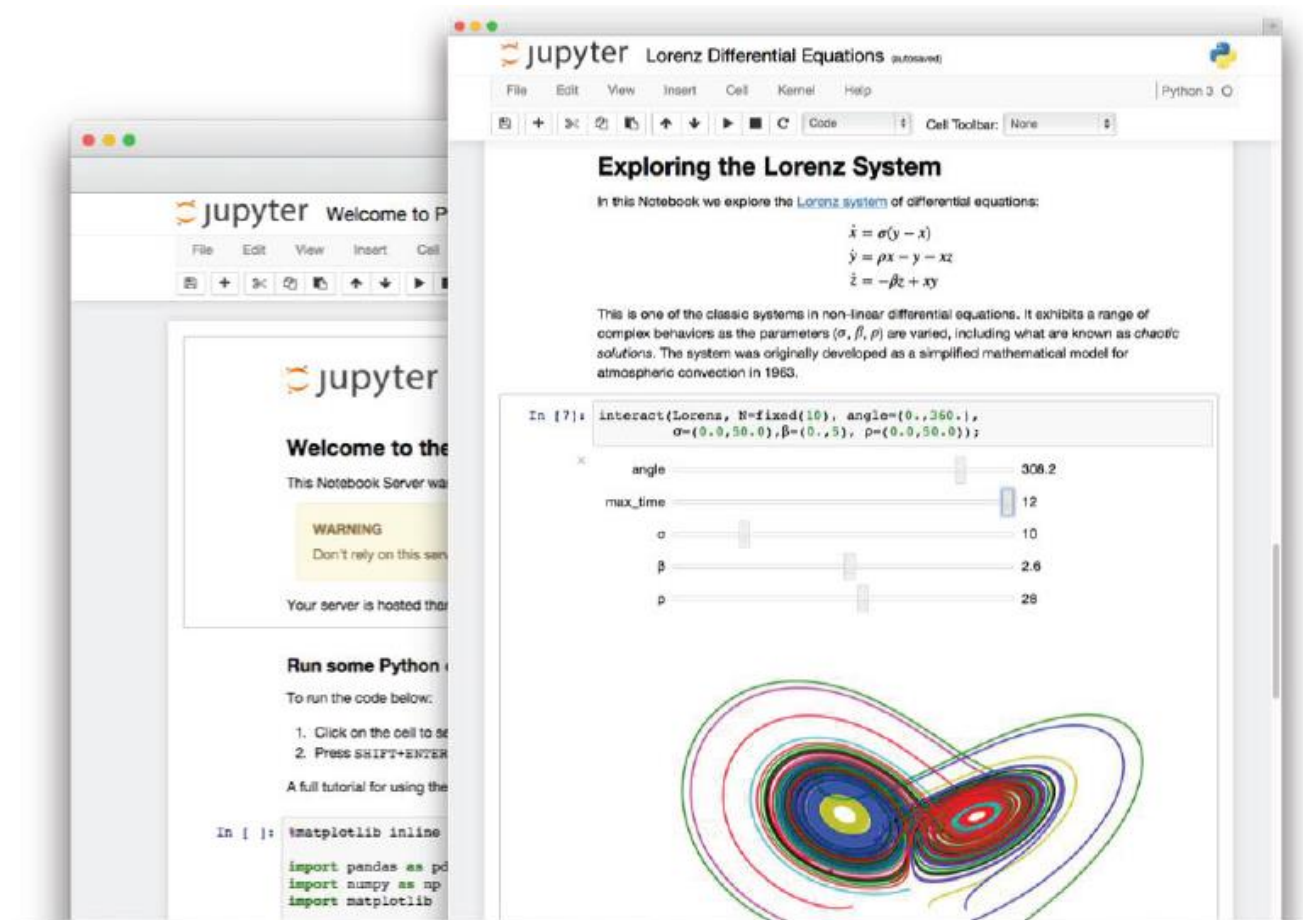
Documentation

Developer Blog



2/3 Jupyter Notebook

- Julia+Python+R = Jupyter
- 基于Web技术的交互式计算文档格式
- 支持Markdown和Latex语法
- 支持代码运行、文本输入、数学公式编辑、内嵌式画图和其他如图片文件的插入，是一个对代码友好的笔记本



2/3 Jupyter Notebook

Anaconda命令行输入 “jupyter notebook” ，即可打开Jupyter Notebook

文件名: .ipynb



The screenshot shows the Jupyter Notebook interface. The top bar includes the Jupyter logo, the file name "1-5elements (2)" (highlighted with a red box), and "(自动保存)". The right side of the top bar has a "Logout" button and a "Python 3" dropdown menu (highlighted with a red box). Below the top bar is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help" (collectively labeled as "菜单栏" with a red box). Below the menu bar is a toolbar with icons for saving, creating new files, opening recent files, undo, redo, running code, and other functions (collectively labeled as "工具栏" with a red box). The main area contains code cells. The first cell is highlighted with a red box and contains a list of Python keywords and operators: `int`, `float`, `+ - * / % **`, `bool True False (0 0.0 None "" [] {} set() ())`, `== > >= !=`, `not or and`, `| &` 按照位或与运算, `NoneType None`, and `str + * split join format`. To the right of this cell, there are two checkmarks and text: "✓ 代码块: 编写代码、添加说明" and "✓ 每个代码块可直接运行". The second cell contains the code: `a = 123454345454235223555545423453523333333121212`, `b = 2`, and `print(a, b, type(b))`. Below the second cell, the output is shown: `123454345454235223555545423453523333333 2 <class 'int'>` (highlighted with a red box). To the right of the output, there is text: "代码块运行结果".

1-5elements (2) (自动保存)

Logout Python 3

File Edit View Insert Cell Kernel Widgets Help

不可信

Python 3

菜单栏

工具栏

版本信息

```
In [ ]: int
float

+ - * / % **

bool True False (0 0.0 None "" [] {} set() ())
== > >= !=
not or and
| & 按照位或与运算

NoneType None

str + * split join format
```

✓ 代码块: 编写代码、添加说明

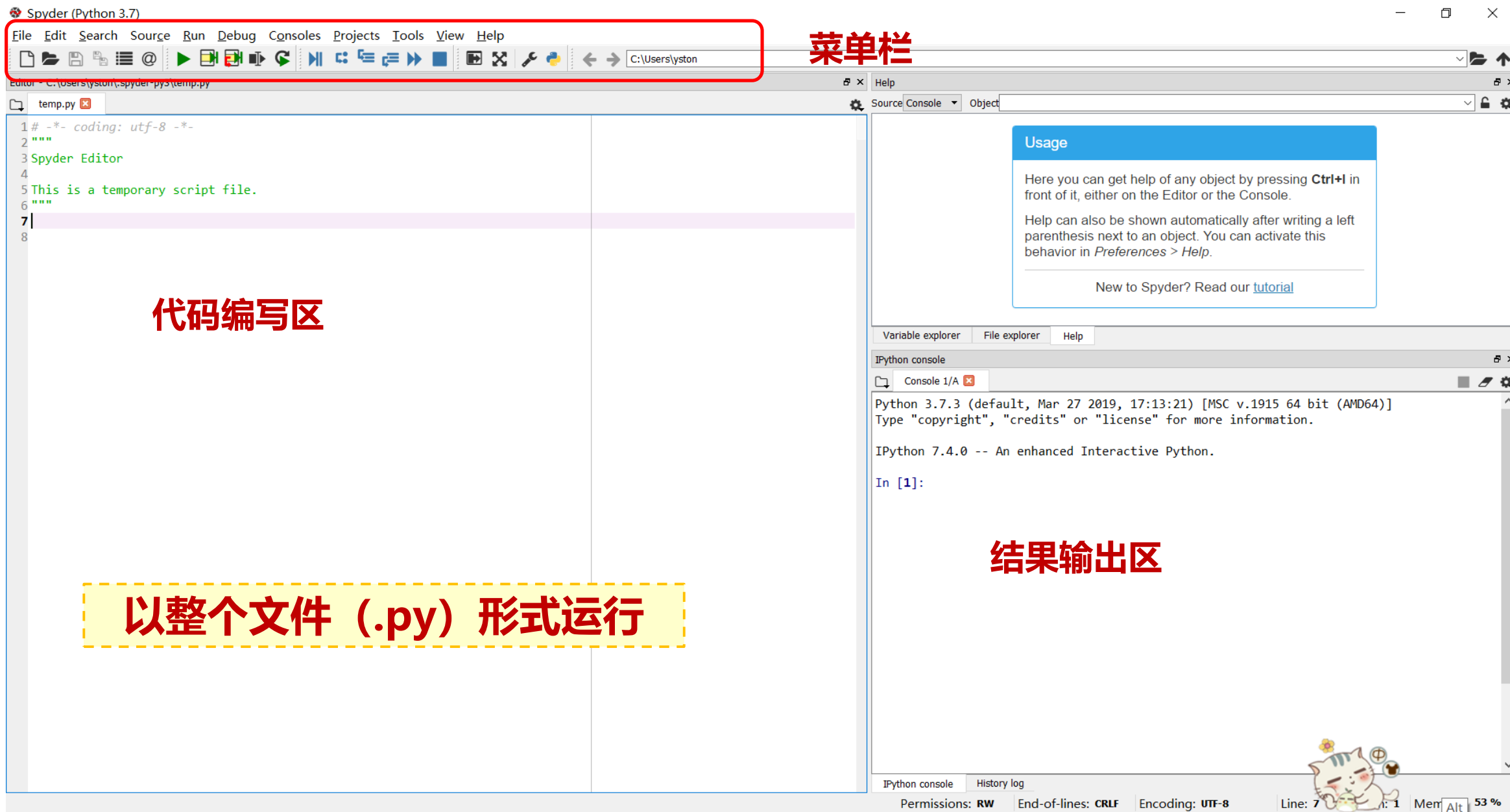
✓ 每个代码块可直接运行

```
In [4]: a = 123454345454235223555545423453523333333121212
b = 2
print(a, b, type(b))
```

123454345454235223555545423453523333333 2 <class 'int'>

代码块运行结果

2/4 集成开发环境Spyder

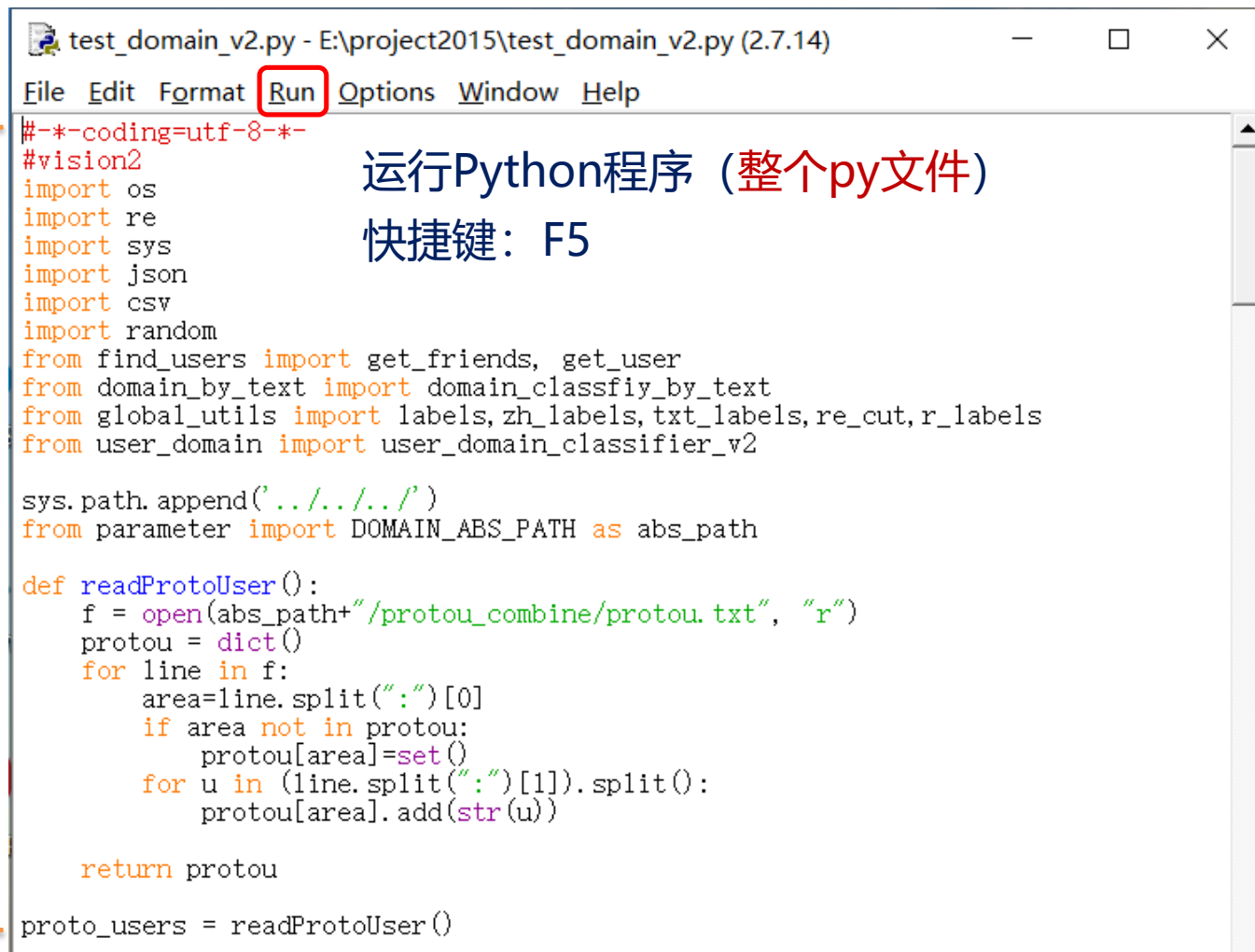


2/5 集成开发环境IDLE

- IDLE是开发Python 程序的基本IDE（集成开发环境），具备基本的IDE的功能
- 当安装好Python以后，IDLE就会自动安装，不需要额外安装

菜单栏

Python代码



```
#-*-coding=utf-8-*-
#vision2
import os
import re
import sys
import json
import csv
import random
from find_users import get_friends, get_user
from domain_by_text import domain_classfiy_by_text
from global_utils import labels, zh_labels, txt_labels, re_cut, r_labels
from user_domain import user_domain_classifier_v2

sys.path.append('../...')
from parameter import DOMAIN_ABS_PATH as abs_path

def readProtoUser():
    f = open(abs_path+"/protou_combine/protou.txt", "r")
    protou = dict()
    for line in f:
        area=line.split(":")[0]
        if area not in protou:
            protou[area]=set()
        for u in (line.split(":")[1]).split():
            protou[area].add(str(u))

    return protou

proto_users = readProtoUser()
```

运行Python程序（整个py文件）

快捷键：F5





谢谢



- 1、使用Jupyter Notebook编写并运行程序
- 2、使用Spyder编写并运行程序

