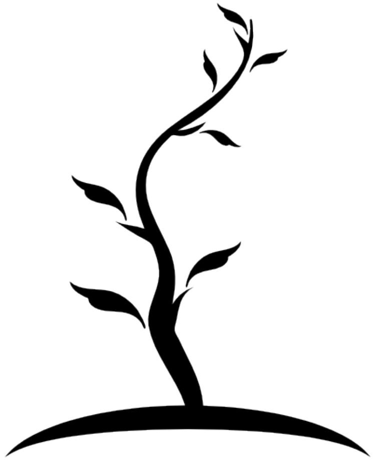


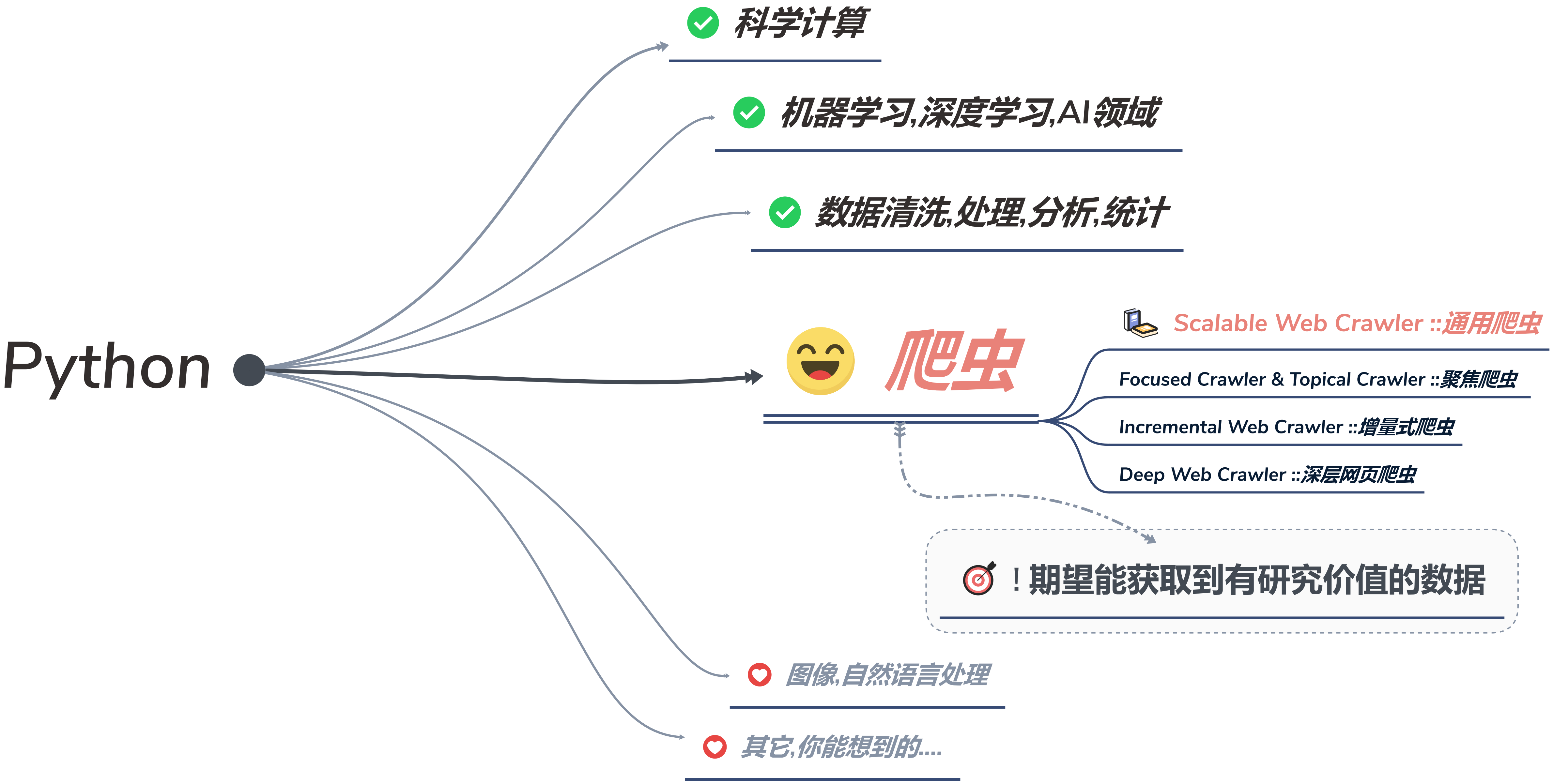
PyShare 03

👤 *lyj8512@126.com*

2018-11-14

Python ≠ 爬虫

Python 语言,工具



Python

- ✓ 科学计算
- ✓ 机器学习,深度学习,AI领域
- ✓ 数据清洗,处理,分析,统计
- 爬虫
  - Scalable Web Crawler ::通用爬虫
  - Focused Crawler & Topical Crawler ::聚焦爬虫
  - Incremental Web Crawler ::增量式爬虫
  - Deep Web Crawler ::深层网页爬虫
  - ！期望能获取到有研究价值的数据
- 图像,自然语言处理
- 其它,你能想到的....

Anaconda Prompt

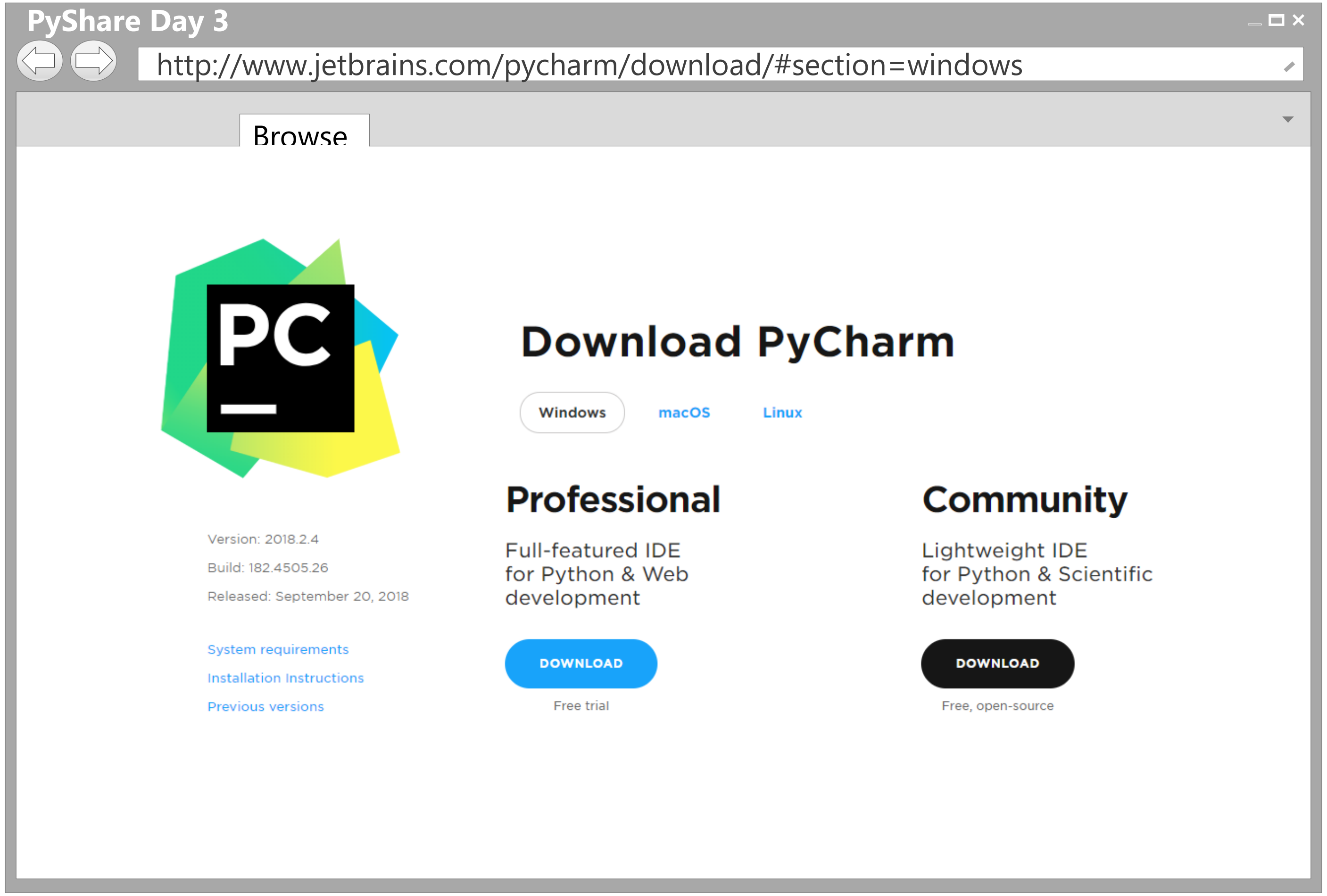
安装第三方库

```
Pip install pyquery
Pip install requests
Pip install aiothhp
```

代码实现,用什么?

编写,调试,运行…

下载安装



PyShare Day 3

http://www.jetbrains.com/pycharm/download/#section=windows

Download PyCharm

Version: 2018.2.4

Build: 182.4505.26

Released: September 20, 2018

System requirements

Installation Instructions

Previous versions

Professional

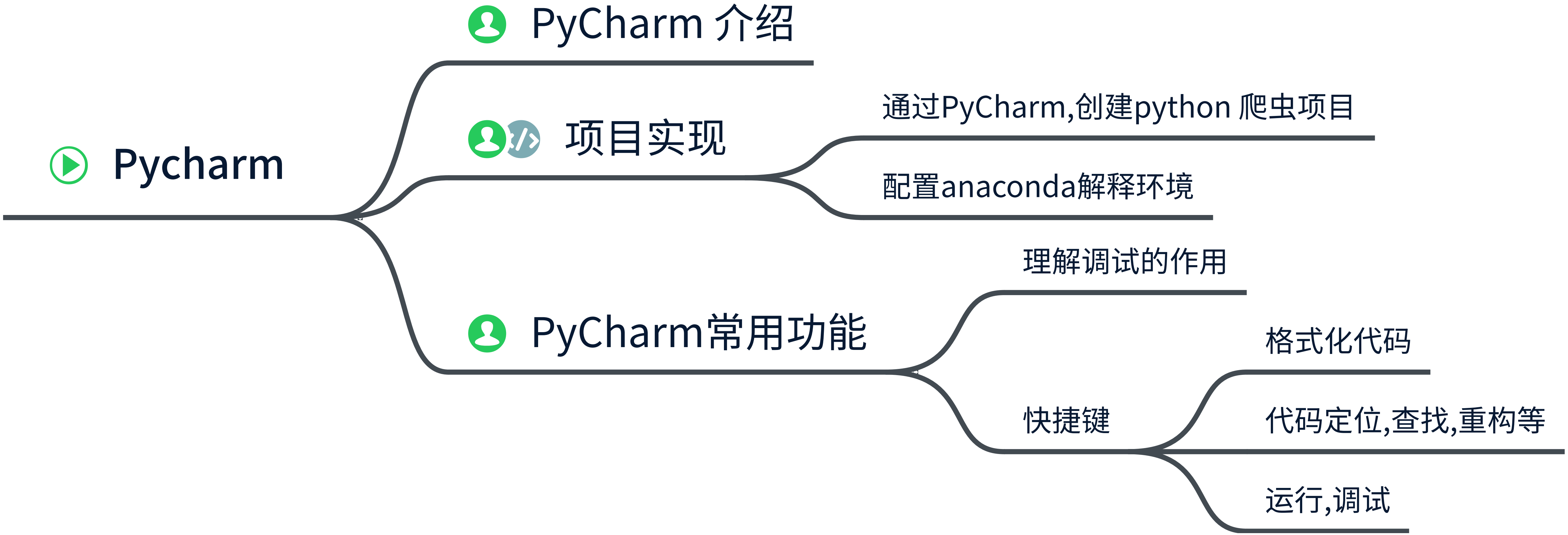
Full-featured IDE
for Python & Web
development

Free trial

Community

Lightweight IDE
for Python & Scientific
development

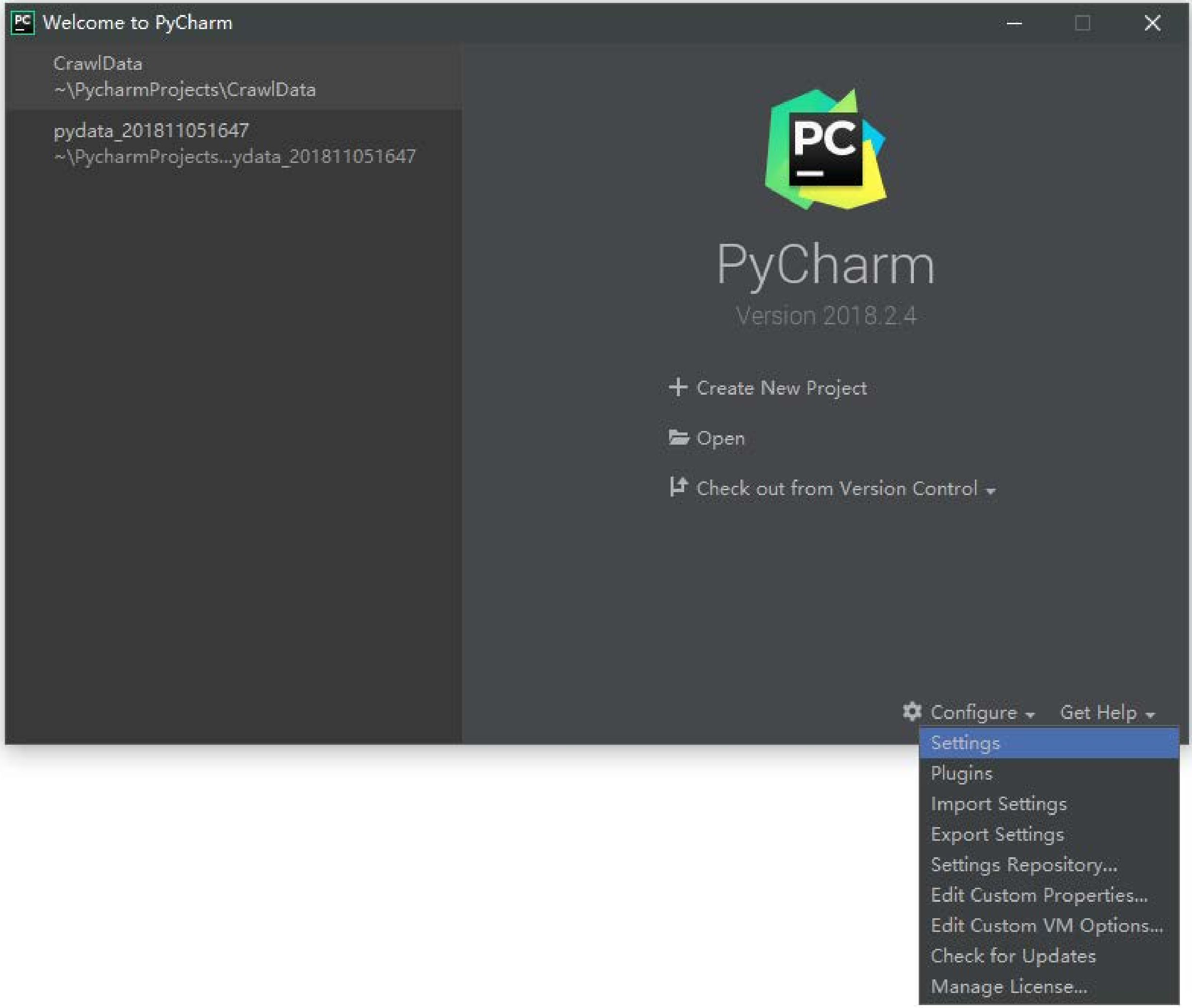
Free, open-source

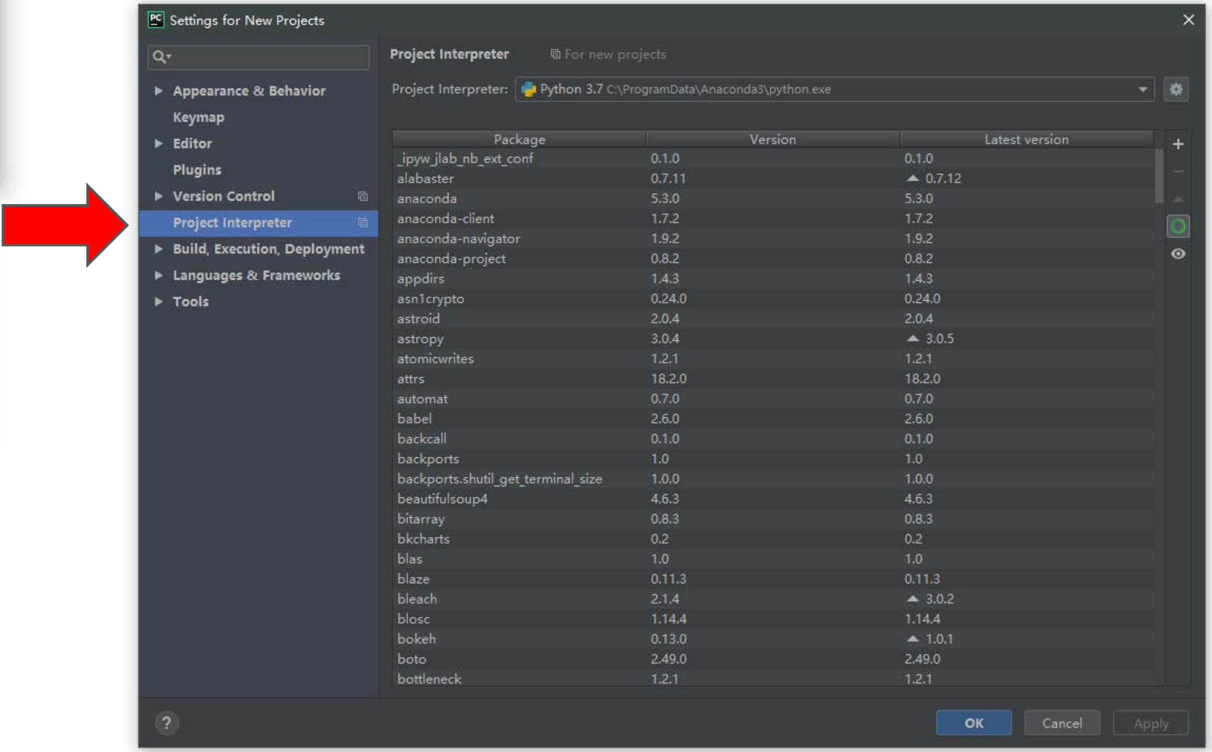


- Pycharm
  - PyCharm 介绍
  - 项目实现
    - 通过PyCharm,创建python 爬虫项目
    - 配置anaconda解释环境
  - PyCharm常用功能
    - 理解调试的作用
    - 快捷键
      - 格式化代码
      - 代码定位,查找,重构等
      - 运行,调试

Pycharm 配置

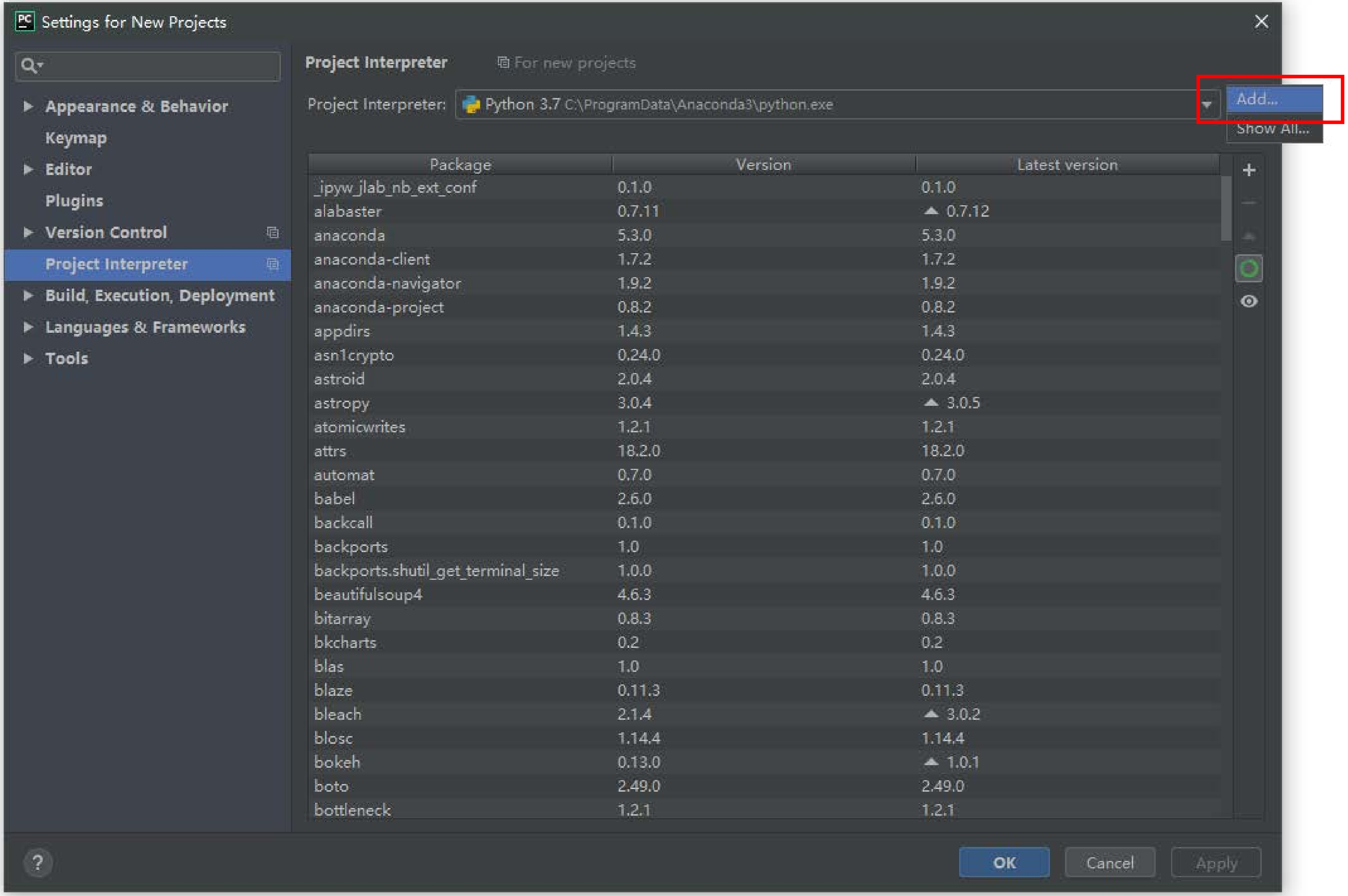
配置加载需要时间,…





Project Interpreter: Python 3.7 C:\ProgramData\Anaconda3\python.exe

Package	Version	Latest version
_ipyw_jlab_nb_ext_conf	0.1.0	0.1.0
alabaster	0.7.11	▲ 0.7.12
anaconda	5.3.0	5.3.0
anaconda-client	1.7.2	1.7.2
anaconda-navigator	1.9.2	1.9.2
anaconda-project	0.8.2	0.8.2
appdirs	1.4.3	1.4.3
asn1crypto	0.24.0	0.24.0
astroid	2.0.4	2.0.4
astropy	3.0.4	▲ 3.0.5
atomicwrites	1.2.1	1.2.1
attrs	18.2.0	18.2.0
automat	0.7.0	0.7.0
babel	2.6.0	2.6.0
backcall	0.1.0	0.1.0
backports	1.0	1.0
backports.shutil_get_terminal_size	1.0.0	1.0.0
beautifulsoup4	4.6.3	4.6.3
bitarray	0.8.3	0.8.3
bkcharts	0.2	0.2
blas	1.0	1.0
blaze	0.11.3	0.11.3
bleach	2.1.4	▲ 3.0.2
blosc	1.14.4	1.14.4
bokeh	0.13.0	▲ 1.0.1
boto	2.49.0	2.49.0
bottleneck	1.2.1	1.2.1

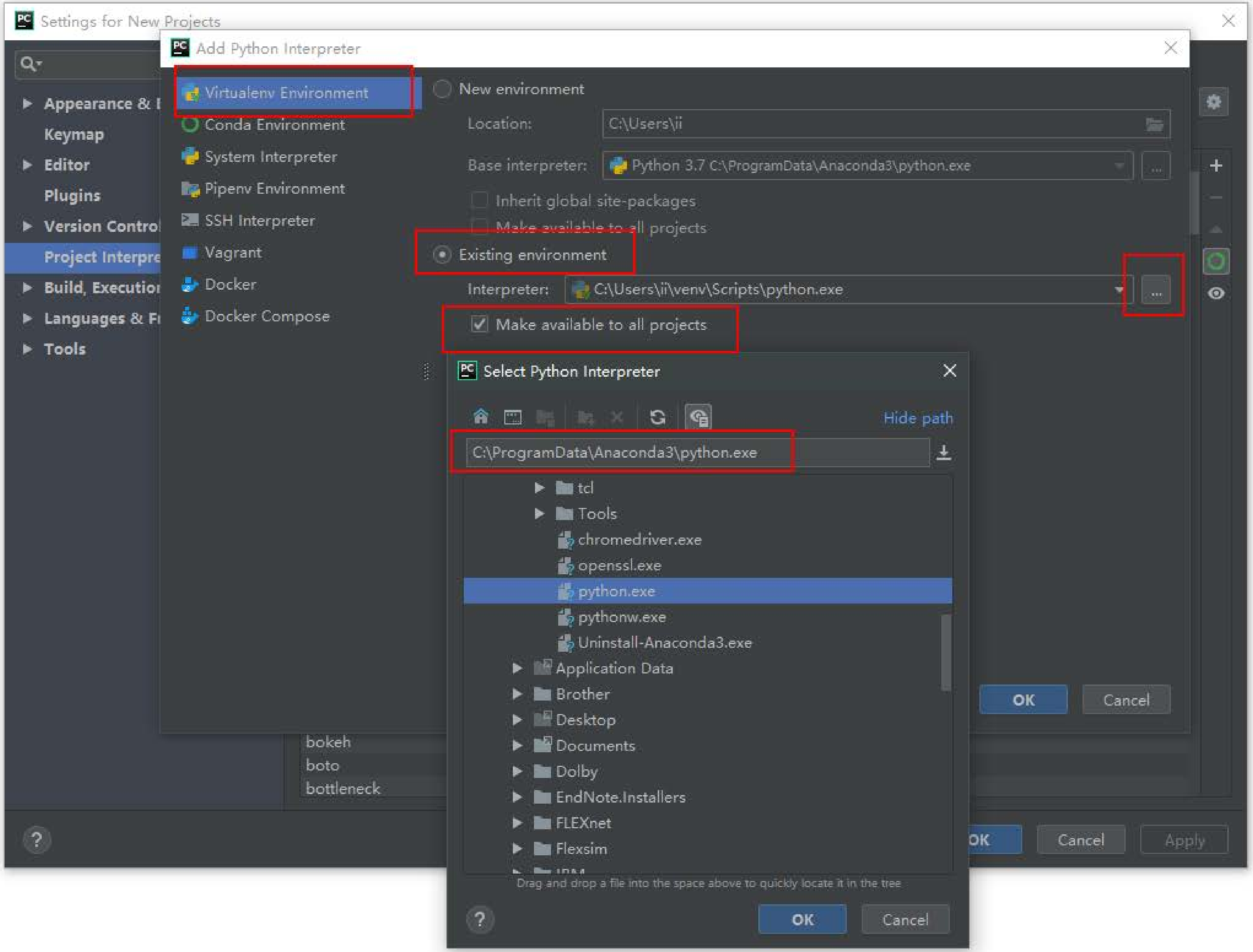


Settings for New Projects

Project Interpreter For new projects

Project Interpreter: Python 3.7 C:\ProgramData\Anaconda3\python.exe

Package	Version	Latest version
_ipyw_jlab_nb_ext_conf	0.1.0	0.1.0
alabaster	0.7.11	▲ 0.7.12
anaconda	5.3.0	5.3.0
anaconda-client	1.7.2	1.7.2
anaconda-navigator	1.9.2	1.9.2
anaconda-project	0.8.2	0.8.2
appdirs	1.4.3	1.4.3
asn1crypto	0.24.0	0.24.0
astroid	2.0.4	2.0.4
astropy	3.0.4	▲ 3.0.5
atomicwrites	1.2.1	1.2.1
attrs	18.2.0	18.2.0
automat	0.7.0	0.7.0
babel	2.6.0	2.6.0
backcall	0.1.0	0.1.0
backports	1.0	1.0
backports.shutil_get_terminal_size	1.0.0	1.0.0
beautifulsoup4	4.6.3	4.6.3
bitarray	0.8.3	0.8.3
bkcharts	0.2	0.2
blas	1.0	1.0
blaze	0.11.3	0.11.3
bleach	2.1.4	▲ 3.0.2
blosc	1.14.4	1.14.4
bokeh	0.13.0	▲ 1.0.1
boto	2.49.0	2.49.0
bottleneck	1.2.1	1.2.1



Settings for New Projects

Add Python Interpreter

○ New environment

Location: C:\Users\ii

Base interpreter: Python 3.7 C:\ProgramData\Anaconda3\python.exe

□ Inherit global site-packages

□ Make available to all projects

● Existing environment

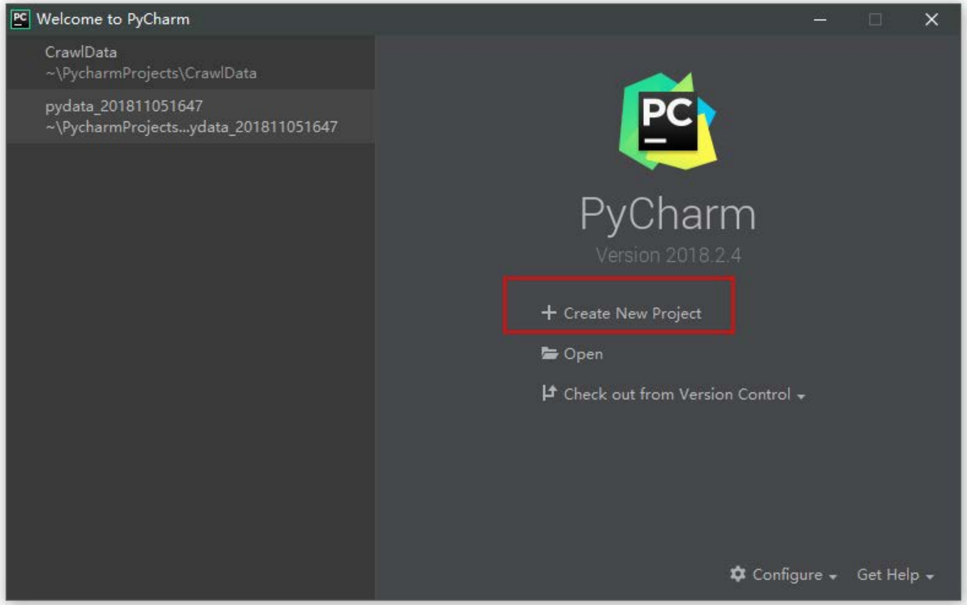
Interpreter: C:\Users\ii\venv\Scripts\python.exe

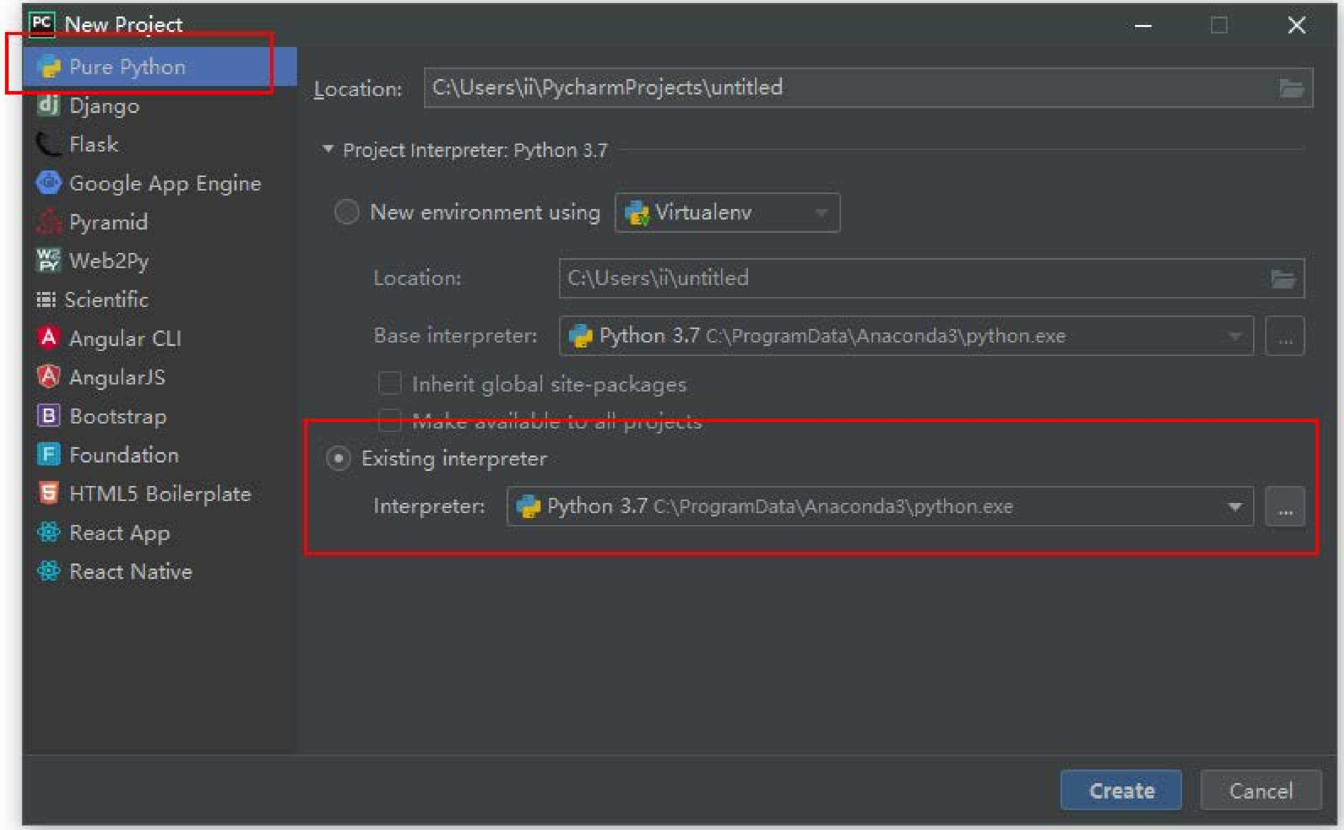
☑ Make available to all projects

Select Python Interpreter

C:\ProgramData\Anaconda3\python.exe

Drag and drop a file into the space above to quickly locate it in the tree





Location: C:\Users\ii\PycharmProjects\untitled

Project Interpreter: Python 3.7

○ New environment using Virtualenv

Location: C:\Users\ii\untitled

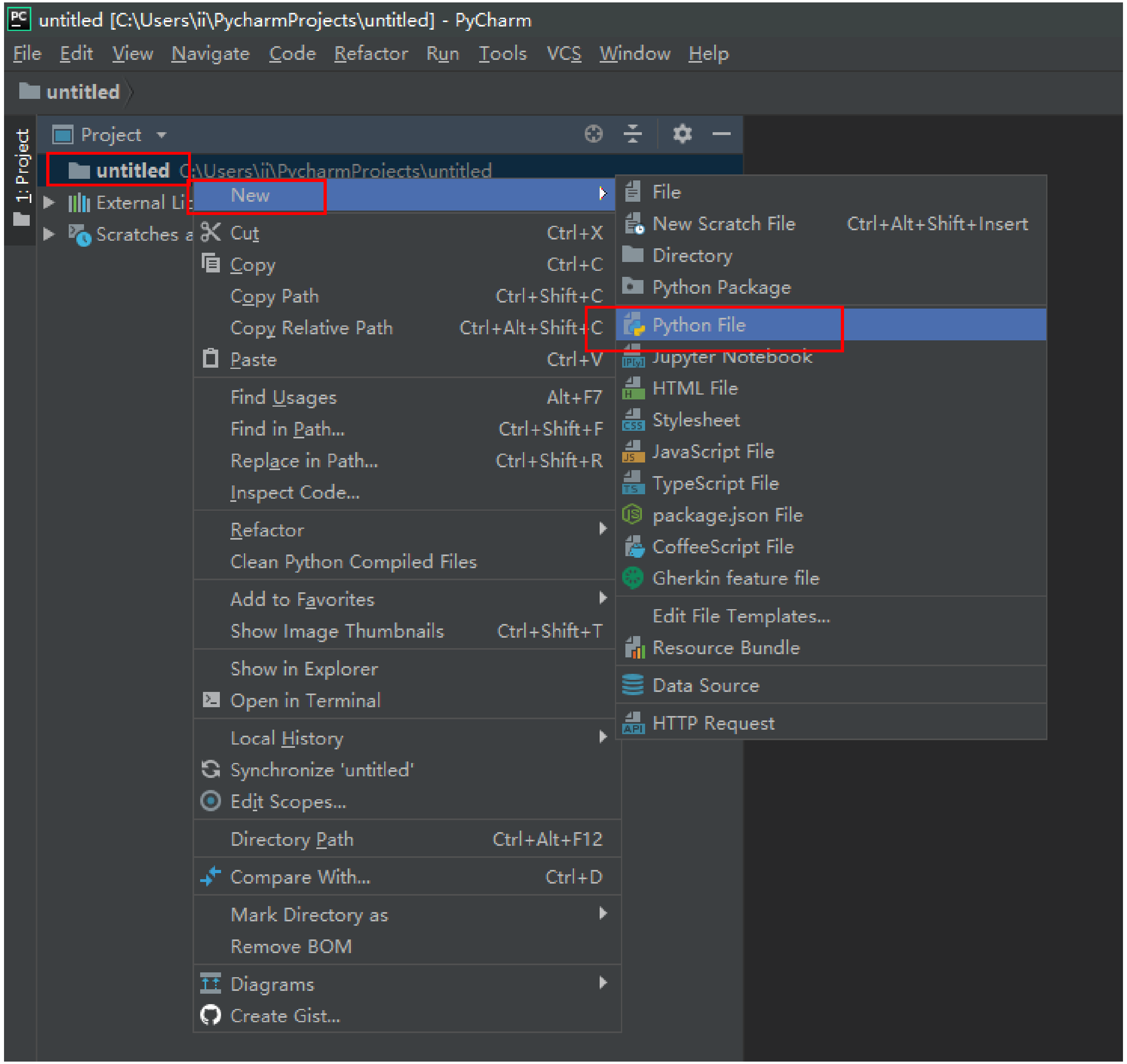
Base interpreter: Python 3.7 C:\ProgramData\Anaconda3\python.exe

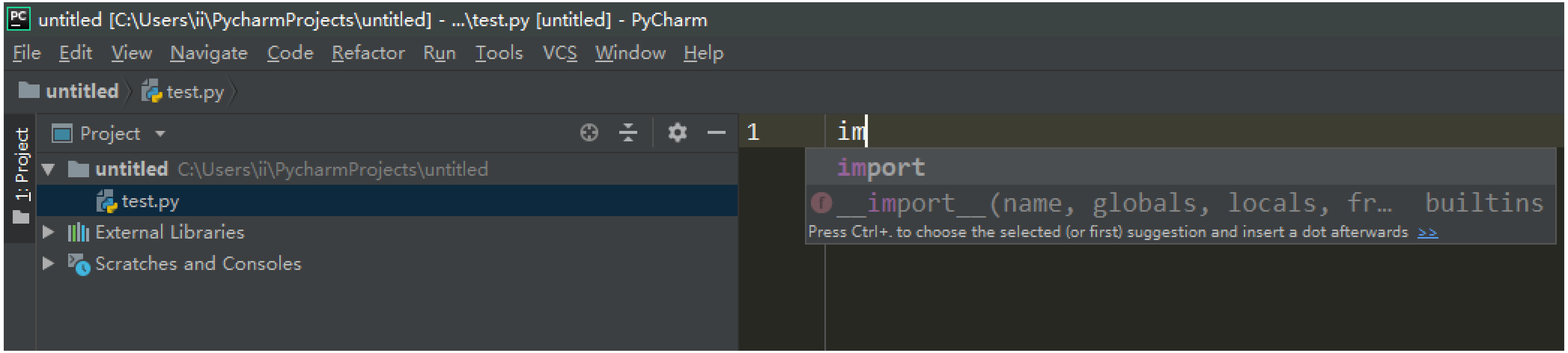
□ Inherit global site-packages

□ Make available to all projects

● Existing interpreter

Interpreter: Python 3.7 C:\ProgramData\Anaconda3\python.exe


untitled [C:\Users\ii\PycharmProjects\untitled] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
untitled
1: Project
Project
untitled C:\Users\ii\PycharmProjects\untitled
External Li
Scratches a
New
Cut Ctrl+X
Copy Ctrl+C
Copy Path Ctrl+Shift+C
Copy Relative Path Ctrl+Alt+Shift+C
Paste Ctrl+V
Find Usages Alt+F7
Find in Path... Ctrl+Shift+F
Replace in Path... Ctrl+Shift+R
Inspect Code...
Refactor
Clean Python Compiled Files
Add to Favorites
Show Image Thumbnails Ctrl+Shift+T
Show in Explorer
Open in Terminal
Local History
Synchronize 'untitled'
Edit Scopes...
Directory Path Ctrl+Alt+F12
Compare With... Ctrl+D
Mark Directory as
Remove BOM
Diagrams
Create Gist...
File
New Scratch File Ctrl+Alt+Shift+Insert
Directory
Python Package
Python File
Jupyter Notebook
HTML File
Stylesheet
JavaScript File
TypeScript File
package.json File
CoffeeScript File
Gherkin feature file
Edit File Templates...
Resource Bundle
Data Source
HTTP Request



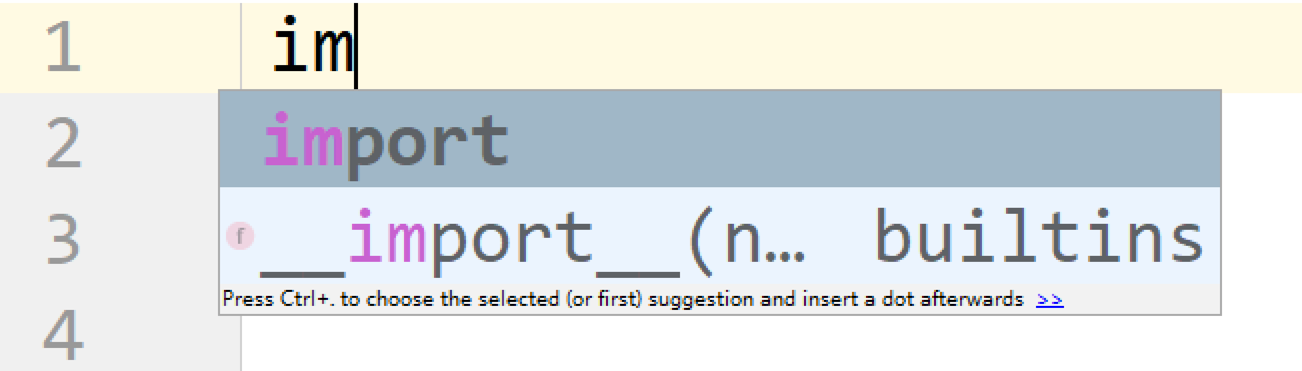
untitled [C:\Users\ii\PycharmProjects\untitled] - ...\test.py [untitled] - PyCharm
File Edit View Navigate Code Refactor Run Tools VCS Window Help
untitled
test.py
1: Project
Project
untitled C:\Users\ii\PycharmProjects\untitled
test.py
External Libraries
Scratches and Consoles

```
im
```

```
import
__import__(name, globals, locals, fr…  builtins
```

Press Ctrl+. to choose the selected (or first) suggestion and insert a dot afterwards >>

为什么要用 PyCharm

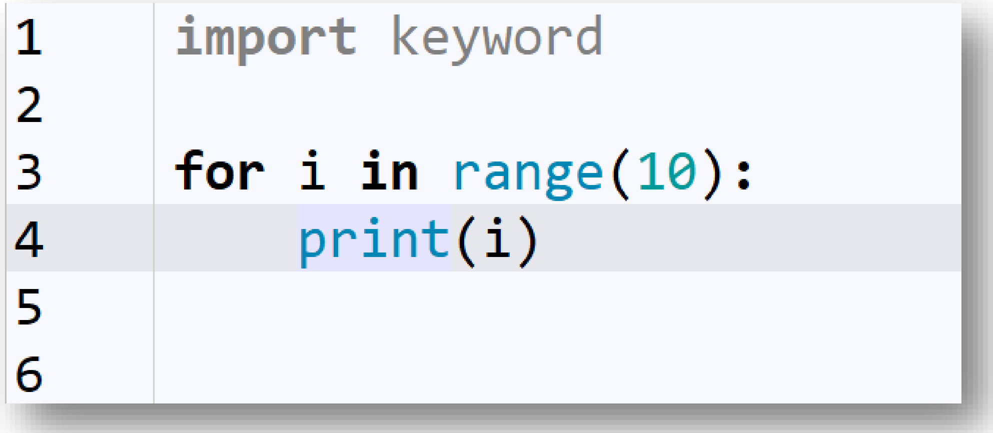


```
im
```

Press Ctrl+. to choose the selected (or first) suggestion and insert a dot afterwards >>

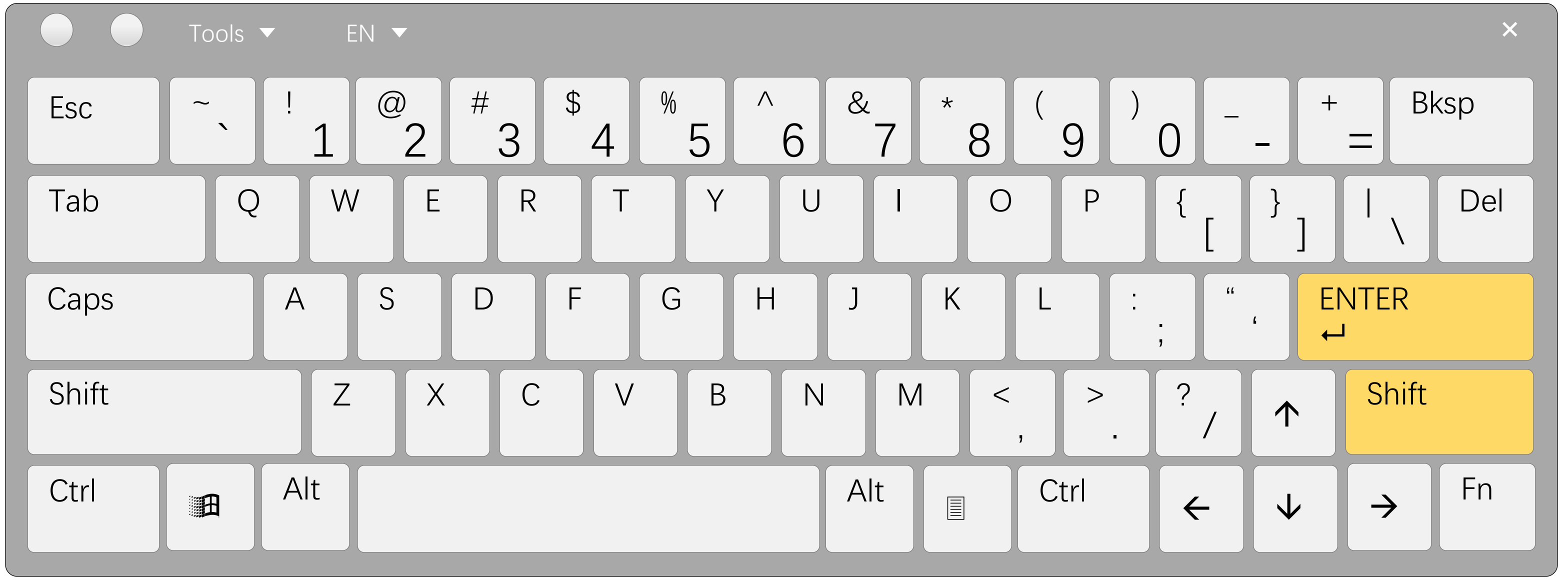
提供(自检,提醒,智能补全)功能…

专注代码逻辑实现,…

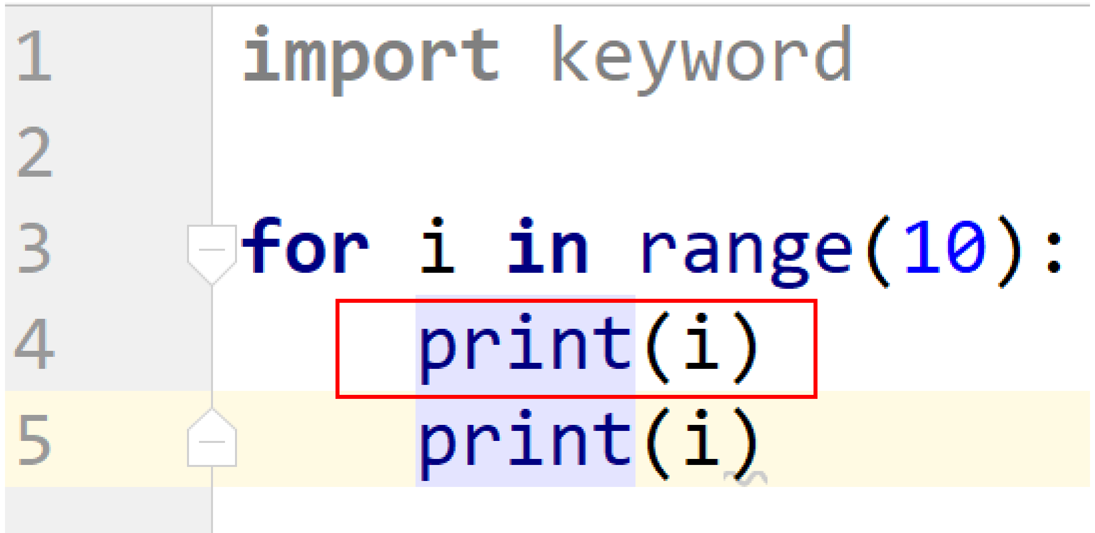


```python
import keyword

for i in range(10):
    print(i)


```

光标任意位置,切换一行,…

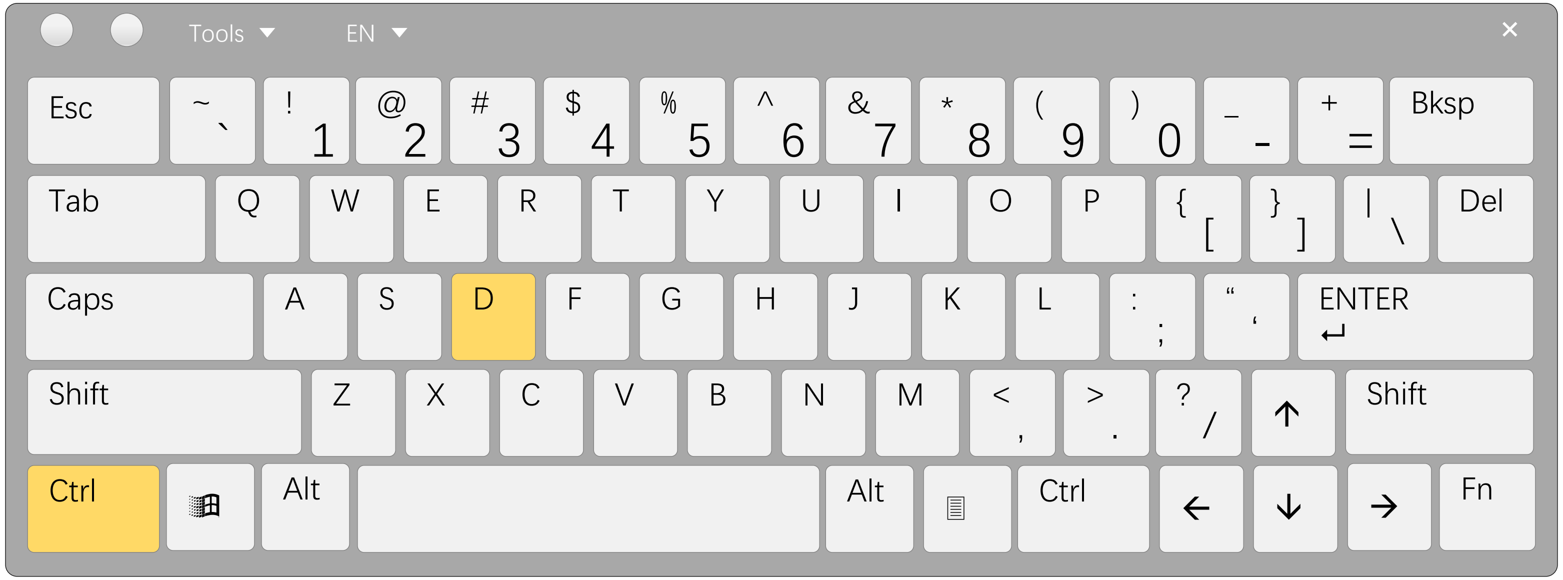


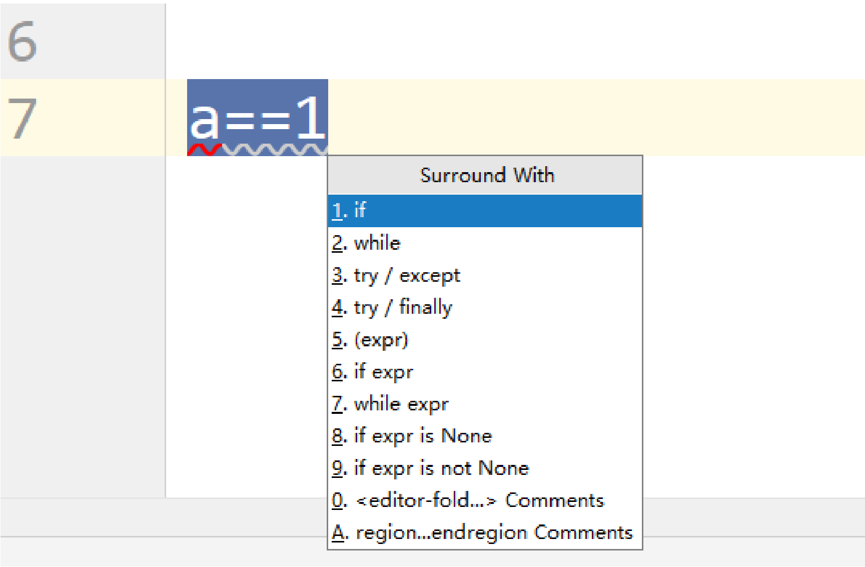
Tools
EN
×
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab Q W E R T Y U I O P { [}] | \ Del
Caps A S D F G H J K L : ; " ' ENTER ↵
Shift Z X C V B N M < , > . ? / ↑ Shift
Ctrl Alt Alt Ctrl ← ↓ → Fn



```python
import keyword

for i in range(10):
    print(i)
    print(i)
```

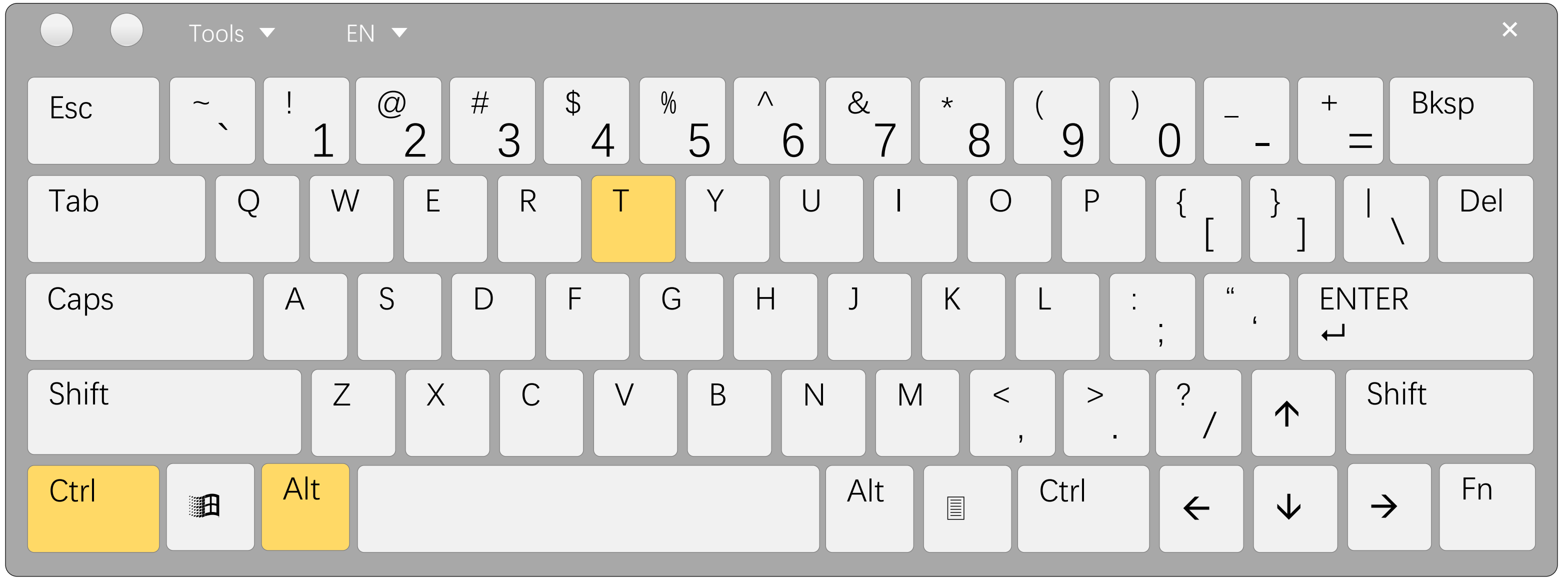
复制当前行,…

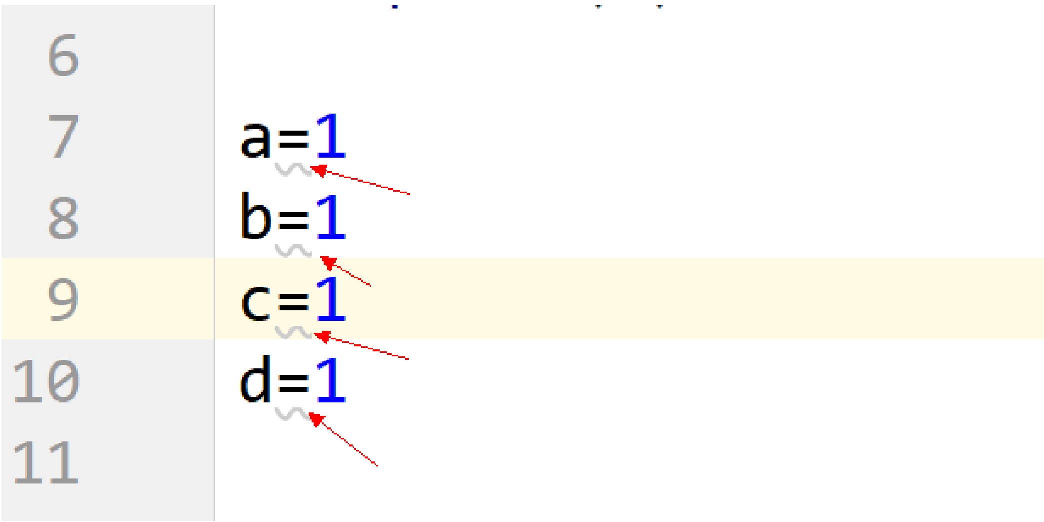

Tools
EN
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab Q W E R T Y U I O P { [}] | \ Del
Caps A S D F G H J K L : ; " ' ENTER
Shift Z X C V B N M < , > . ? / ↑ Shift
Ctrl Alt Alt Ctrl ← ↓ → Fn



```
6
7    a==1
```

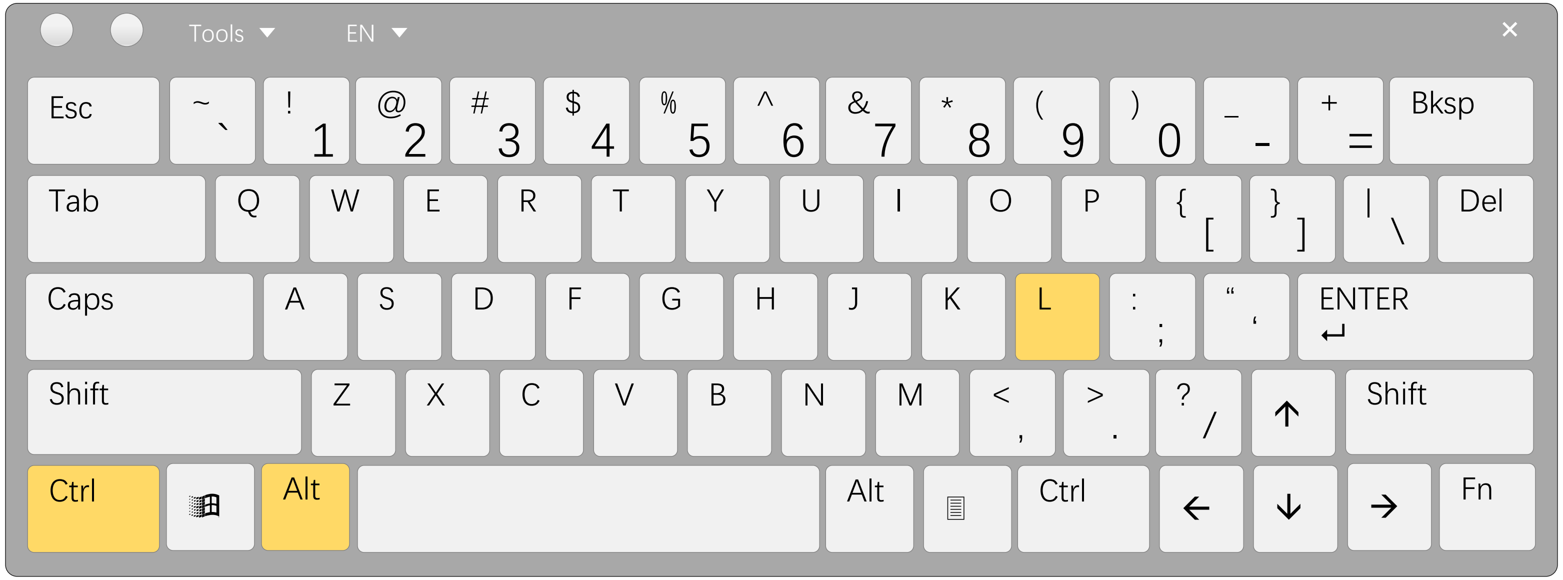
对当前行进行流程控制语句补全…

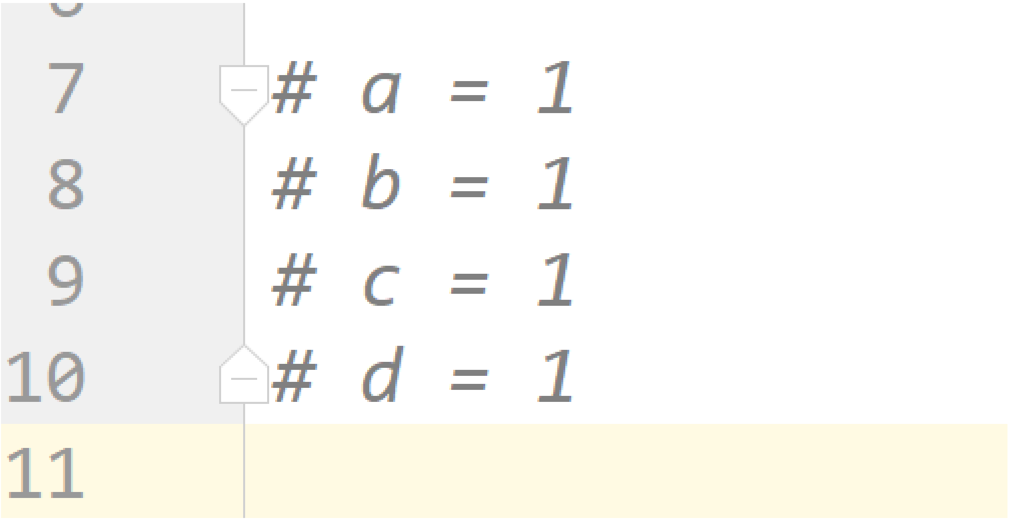

Tools
EN
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab Q W E R T Y U I O P { [}] | \ Del
Caps A S D F G H J K L : ; " ' ENTER
Shift Z X C V B N M < , > . ? / Shift
Ctrl Alt Alt Ctrl Fn



```
a=1
b=1
c=1
d=1
```

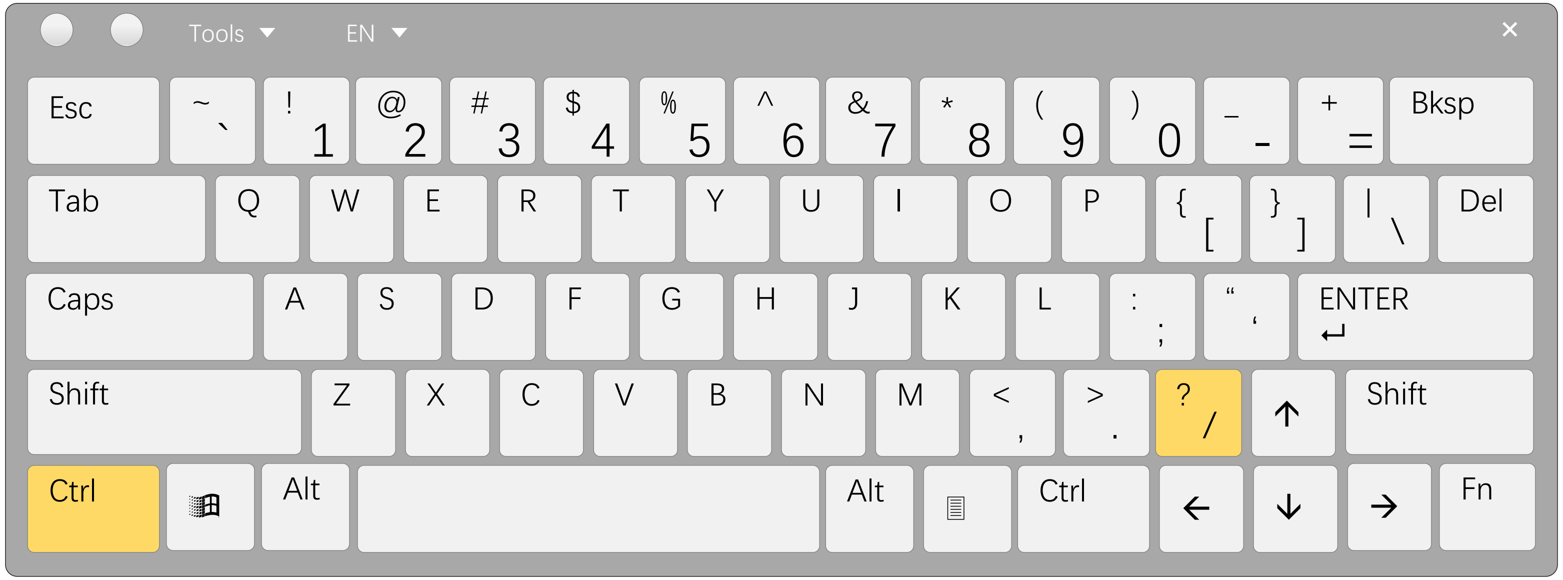
有波浪提示的时候,表示代码编写规范,…
用快捷键格式化格式即可,


Tools
EN
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab Q W E R T Y U I O P { [}] | \ Del
Caps A S D F G H J K L : ; " ' ENTER
Shift Z X C V B N M < , > . ? / Shift
Ctrl Alt Alt Ctrl Fn



```
# a = 1
# b = 1
# c = 1
# d = 1

```

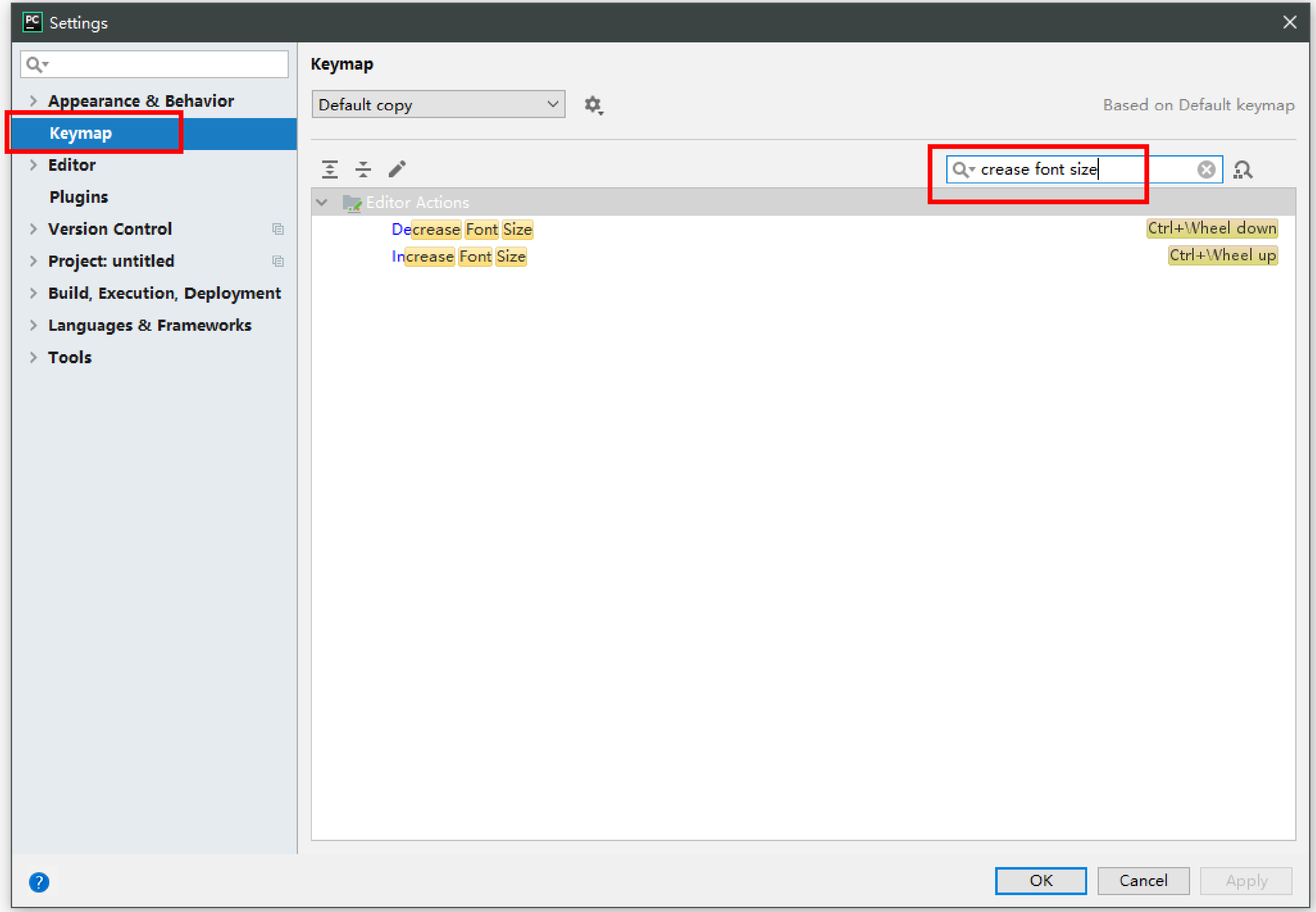
代码注释切换,


Tools ▾ EN ▾ ×
Esc ~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Bksp
Tab Q W E R T Y U I O P { [}] | \ Del
Caps A S D F G H J K L : ; " ' ENTER ↵
Shift Z X C V B N M < , > . ? / ↑ Shift
Ctrl Alt Alt Ctrl ← ↓ → Fn

更多快捷键:

https://blog.csdn.net/pipisorry/article/details/39909057

PyCharm 个性化设置

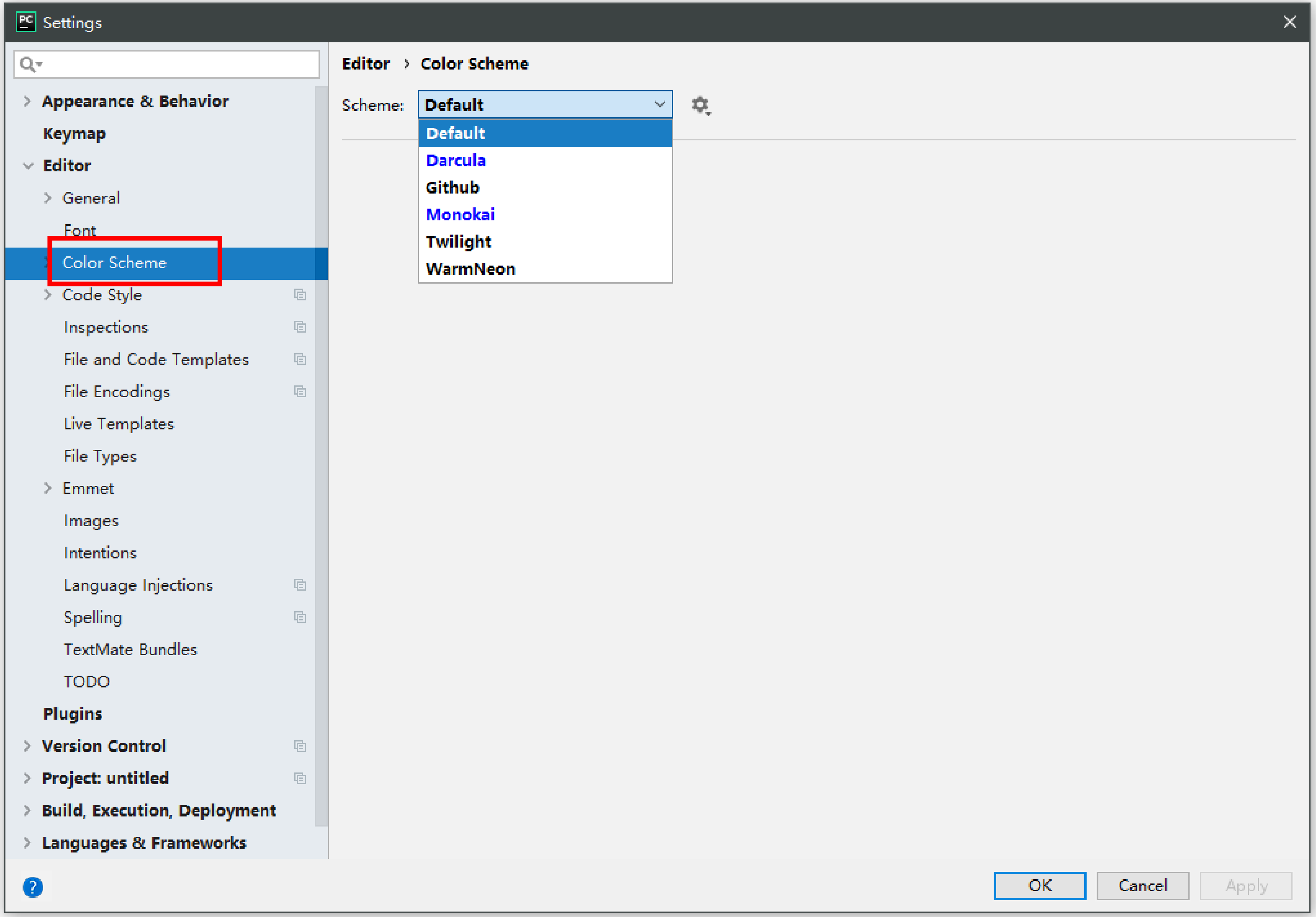


crease font size

Decrease Font Size	Ctrl+Wheel down
Increase Font Size	Ctrl+Wheel up

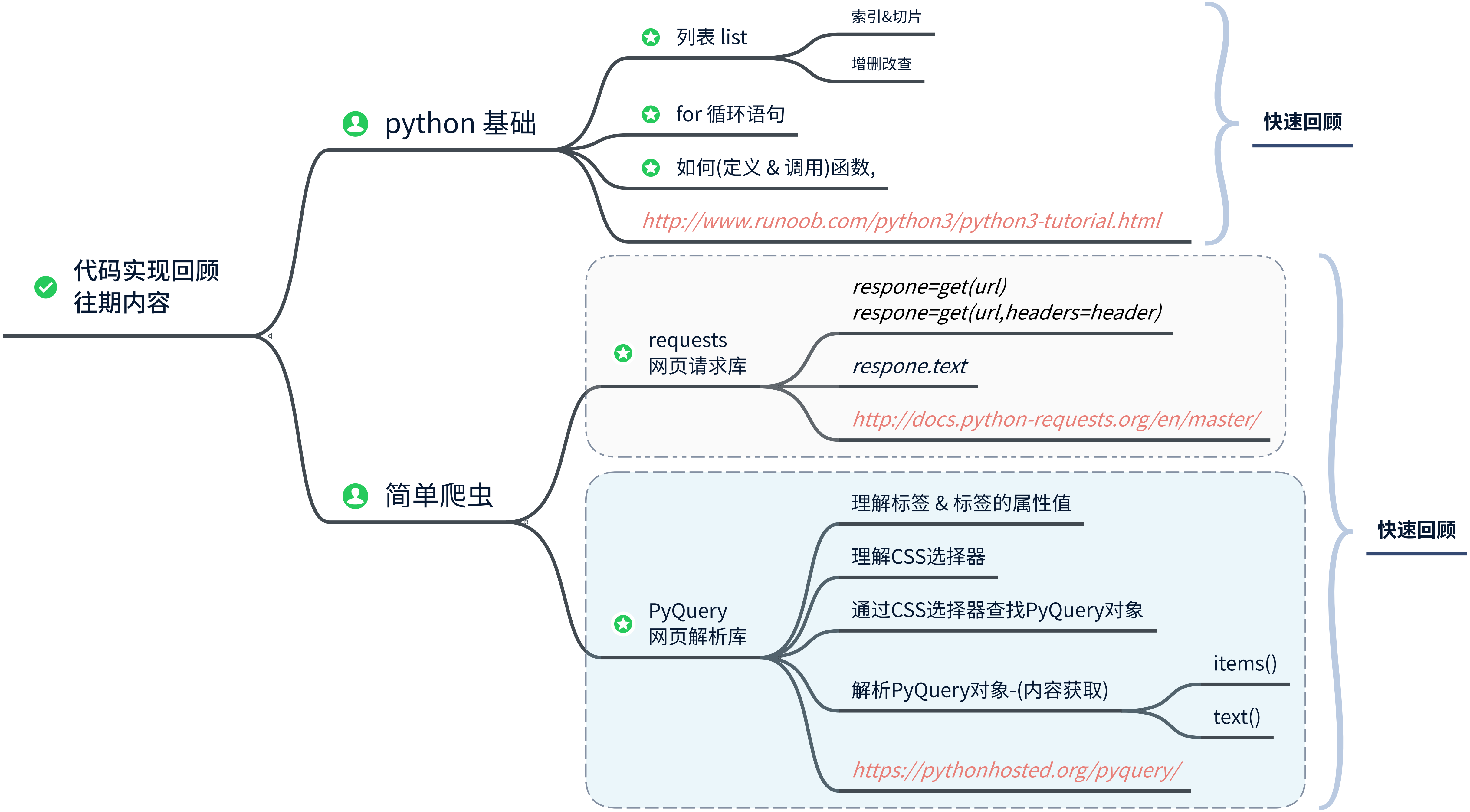
设置字体大小

CTRL+鼠标滚轮控制



Settings
Appearance & Behavior
Keymap
Editor
General
Font
Color Scheme
Code Style
Inspections
File and Code Templates
File Encodings
Live Templates
File Types
Emmet
Images
Intentions
Language Injections
Spelling
TextMate Bundles
TODO
Plugins
Version Control
Project: untitled
Build, Execution, Deployment
Languages & Frameworks
Editor › Color Scheme
Scheme: Default
Default
Darcula
Github
Monokai
Twilight
WarmNeon
OK
Cancel
Apply

设置主题



代码实现回顾往期内容

- python 基础
  - 列表 list
    - 索引&切片
    - 增删改查
  - for 循环语句
  - 如何(定义 & 调用)函数,
  - *http://www.runoob.com/python3/python3-tutorial.html*

快速回顾

- 简单爬虫
  - requests 网页请求库
    - *respone=get(url)*
      respone=get(url,headers=header)
    - *respone.text*
    - *http://docs.python-requests.org/en/master/*
  - PyQuery 网页解析库
    - 理解标签 & 标签的属性值
    - 理解CSS选择器
    - 通过CSS选择器查找PyQuery对象
    - 解析PyQuery对象-(内容获取)
      - items()
      - text()
    - *https://pythonhosted.org/pyquery/*

快速回顾

PyCharm

练习:

创建列表,索引,切片
创建函数,调用函数…

PyCharm

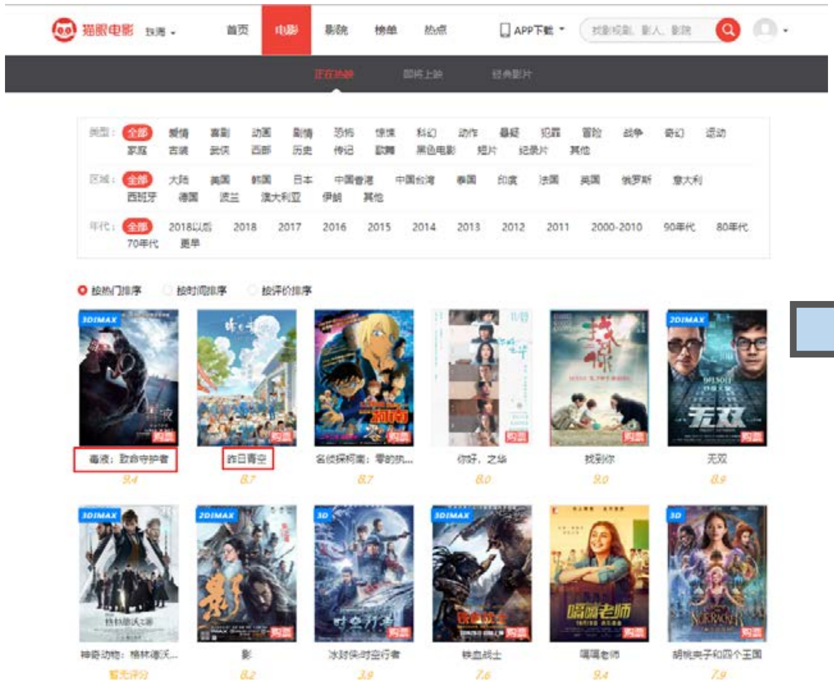
练习:简单爬虫实现

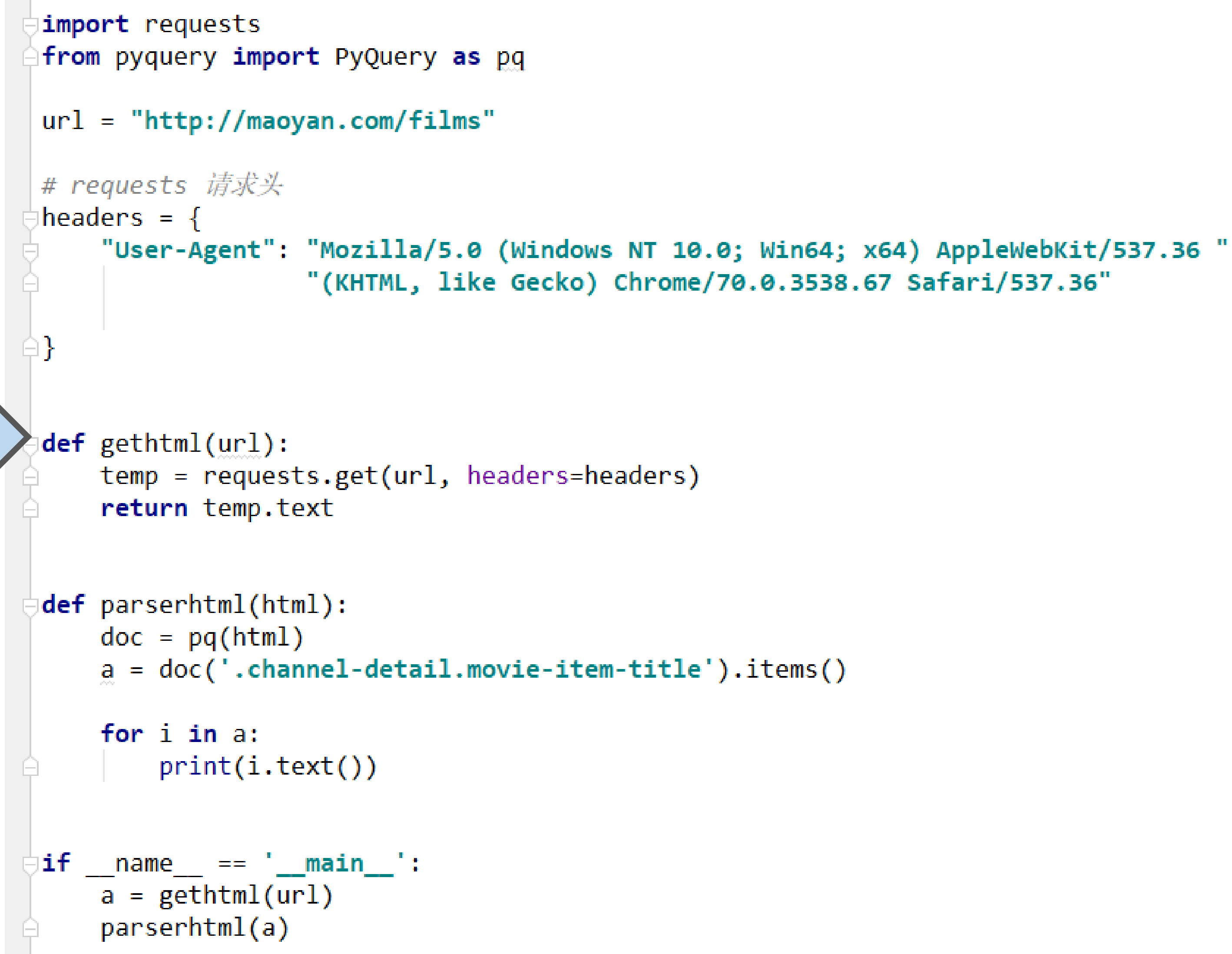
任务:猫眼网页:正在热映电影榜单
http://maoyan.com/films

<code>　PyShare D3-00.py

猫眼网页:

http://maoyan.com/films


猫眼电影
珠海
首页
电影
影院
榜单
热点
APP下载
正在热映
即将上映
经典影片
类型：全部 爱情 喜剧 动画 剧情 恐怖 惊悚 科幻 动作 悬疑 犯罪 冒险 战争 奇幻 运动 家庭 古装 武侠 西部 历史 传记 歌舞 黑色电影 短片 纪录片 其他
区域：全部 大陆 美国 韩国 日本 中国香港 中国台湾 泰国 印度 法国 英国 俄罗斯 意大利 西班牙 德国 波兰 澳大利亚 伊朗 其他
年代：全部 2018以后 2018 2017 2016 2015 2014 2013 2012 2011 2000-2010 90年代 80年代 70年代 更早
按热门排序
按时间排序
按评价排序
毒液：致命守护者
昨日青空
名侦探柯南：零的执...
你好，之华
找到你
无双
神奇动物：格林德沃...
影
冰封侠:时空行者
铁血战士
嗝嗝老师
胡桃夹子和四个王国



```python
import requests
from pyquery import PyQuery as pq

url = "http://maoyan.com/films"

# requests 请求头
headers = {
    "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 "
                  "(KHTML, like Gecko) Chrome/70.0.3538.67 Safari/537.36"

}


def gethtml(url):
    temp = requests.get(url, headers=headers)
    return temp.text


def parserhtml(html):
    doc = pq(html)
    a = doc('.channel-detail.movie-item-title').items()

    for i in a:
        print(i.text())


if __name__ == '__main__':
    a = gethtml(url)
    parserhtml(a)
```

Requests :

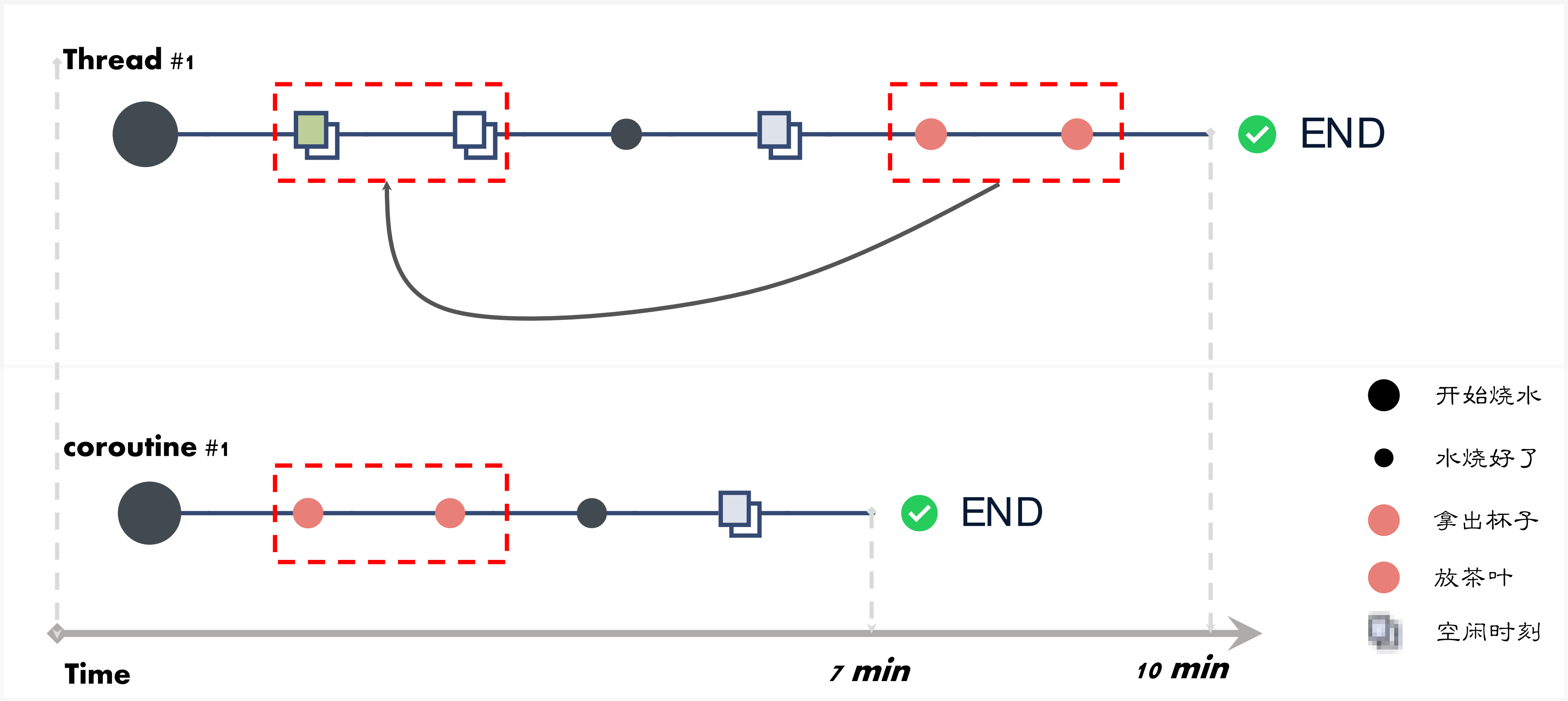
http://docs.python-requests.org/en/master/

PyQuery :

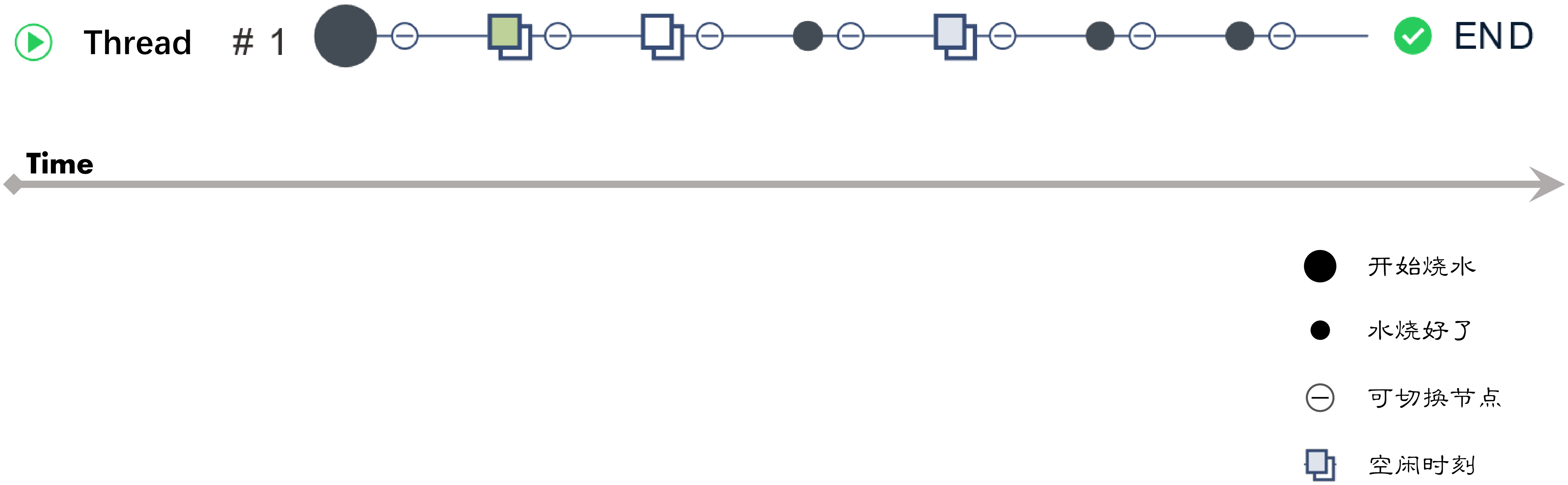
https://pythonhosted.org/pyquery/

单线程 & 协程

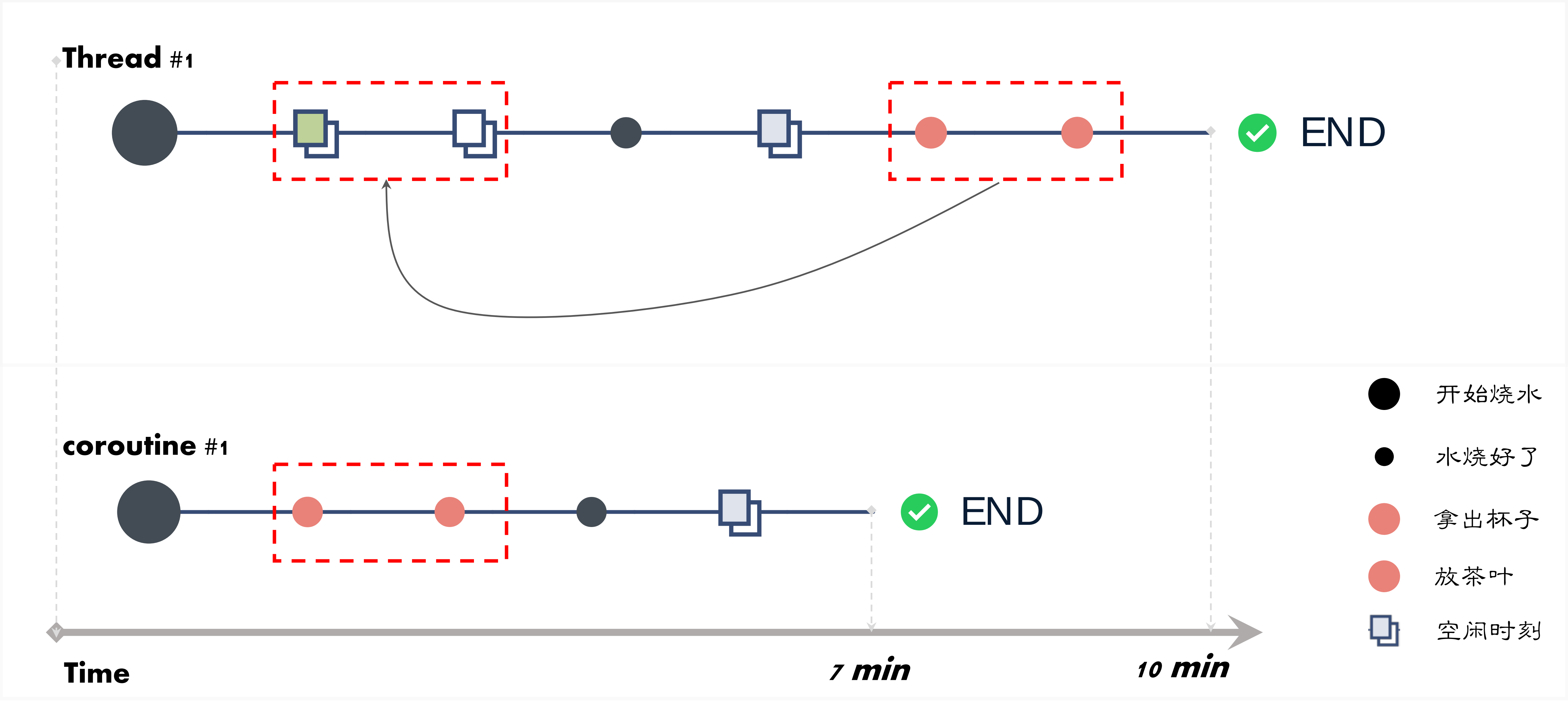
- 线程是什么
  - 执行A事时,有等待返回的操作时,处理器停止(按顺序执行)
  - 数据爬取随网页数量线性递增
- 协程是什么
  - 执行A事件时,有等待返回的操作时,先挂起A事件,执行B事件,当A事件有返回值时,执行A(异步执行)
  - 异步执行,节省时间,速度快

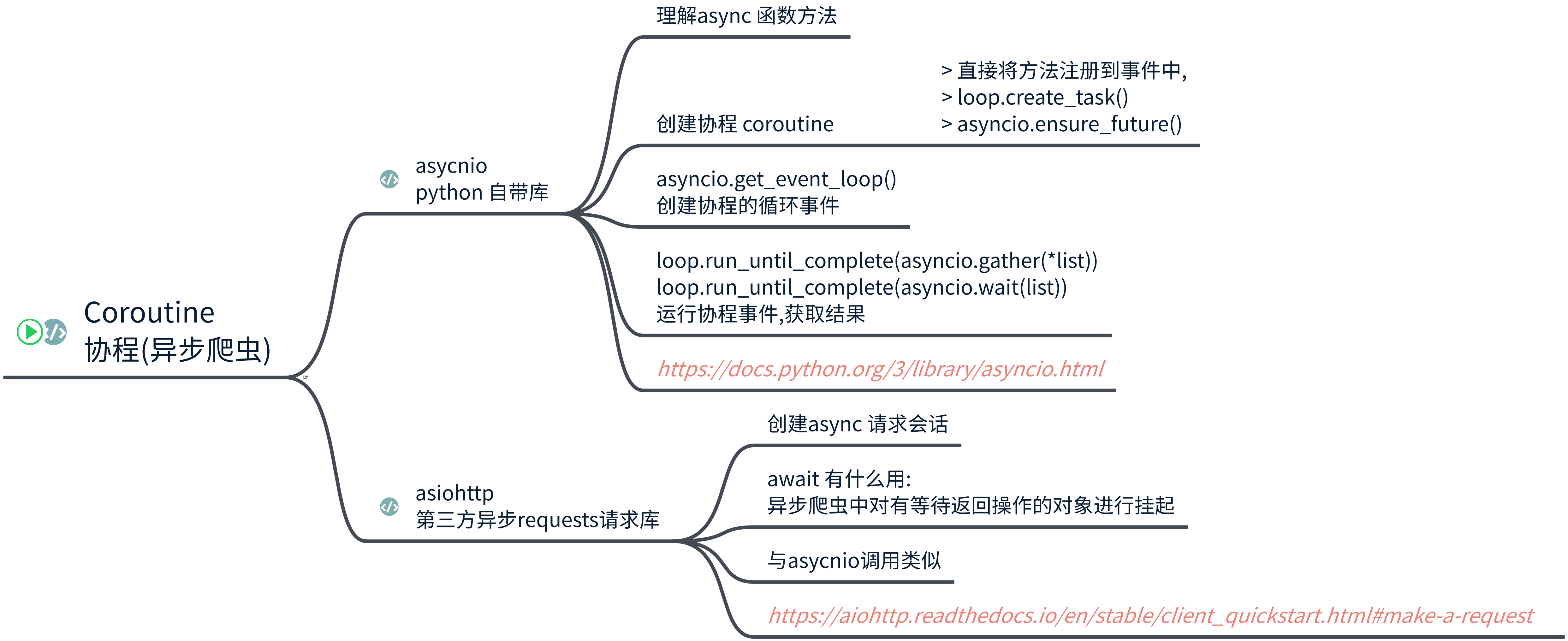


理解线程



理解协程


Thread #1
END
coroutine #1
END
开始烧水
水烧好了
拿出杯子
放茶叶
空闲时刻
Time
7 min
10 min



Coroutine
协程(异步爬虫)

asycnio
python 自带库

- 理解async 函数方法
- 创建协程 coroutine
  - > 直接将方法注册到事件中,
    > loop.create_task()
    > asyncio.ensure_future()
- asyncio.get_event_loop()
  创建协程的循环事件
- loop.run_until_complete(asyncio.gather(*list))
  loop.run_until_complete(asyncio.wait(list))
  运行协程事件,获取结果
- *https://docs.python.org/3/library/asyncio.html*

asiohttp
第三方异步requests请求库

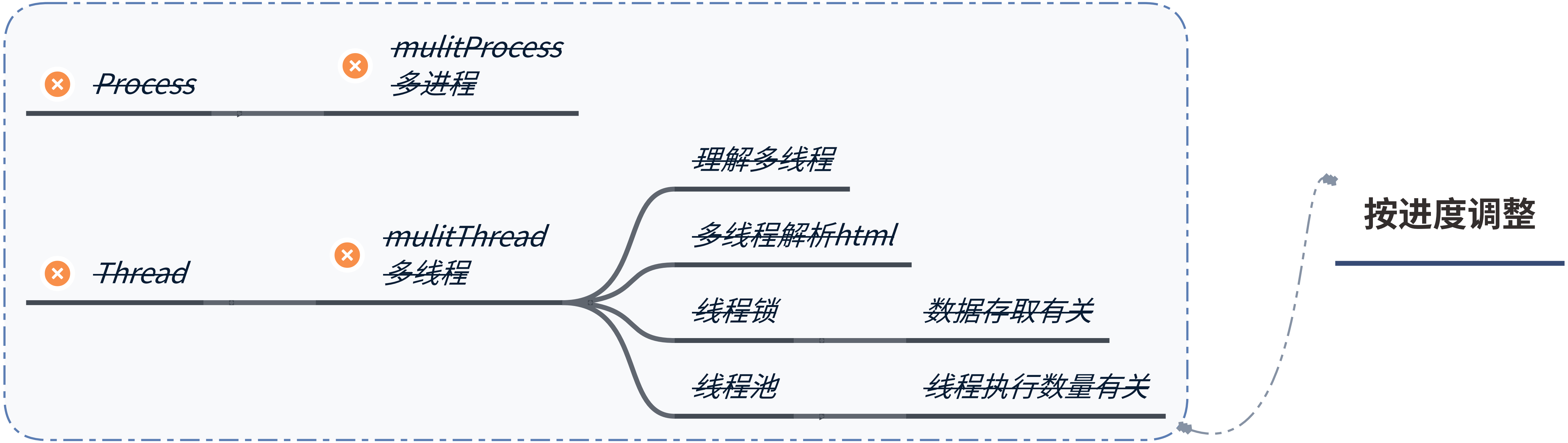
- 创建async 请求会话
- await 有什么用:
  异步爬虫中对有等待返回操作的对象进行挂起
- 与asycnio调用类似
- *https://aiohttp.readthedocs.io/en/stable/client_quickstart.html#make-a-request*

asycnio :

https://docs.python.org/3/library/asyncio.html

aiohttp :

https://aiohttp.readthedocs.io/en/stable/client_quickstart.html#make-a-request



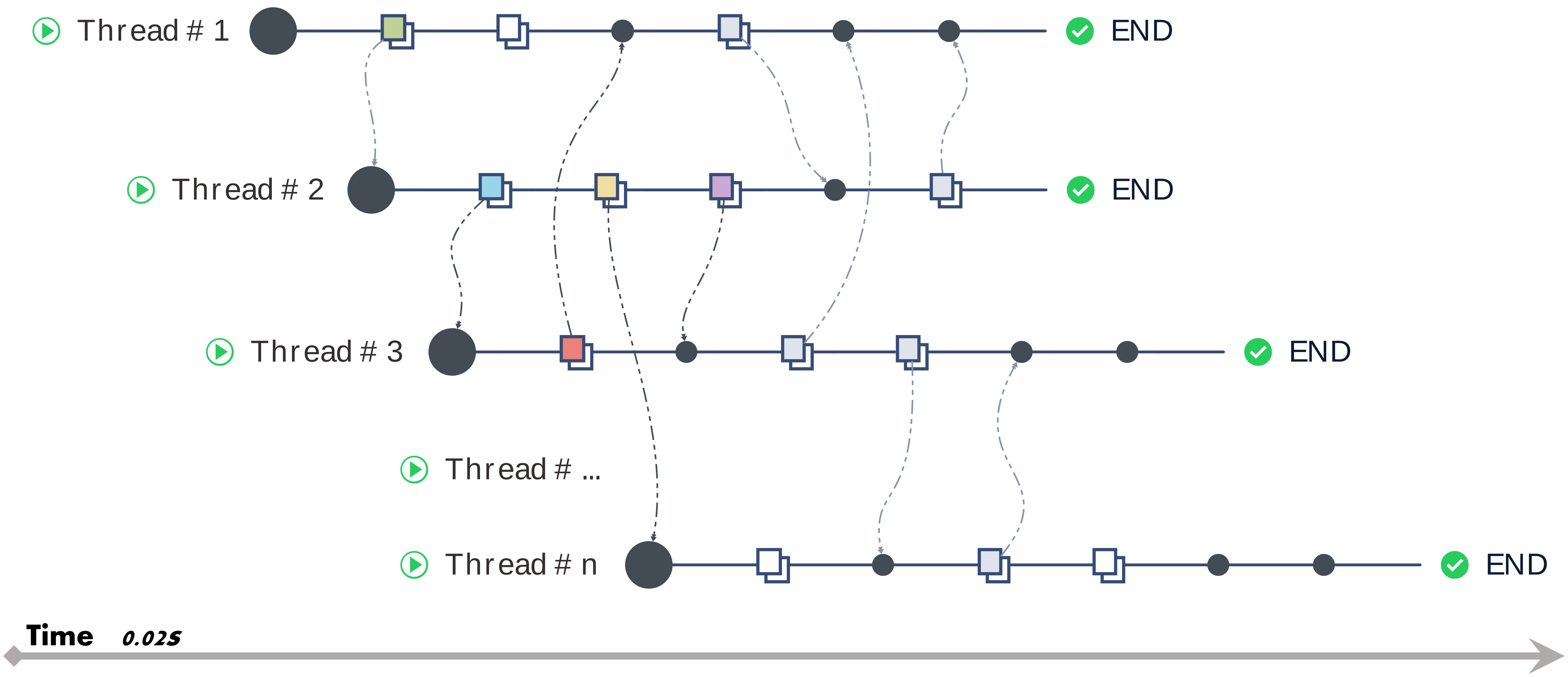
- ~~Process~~
  - ~~mulitProcess 多进程~~
- ~~Thread~~
  - ~~mulitThread 多线程~~
    - ~~理解多线程~~
    - ~~多线程解析html~~
    - ~~线程锁~~
      - ~~数据存取有关~~
    - ~~线程池~~
      - ~~线程执行数量有关~~

按进度调整

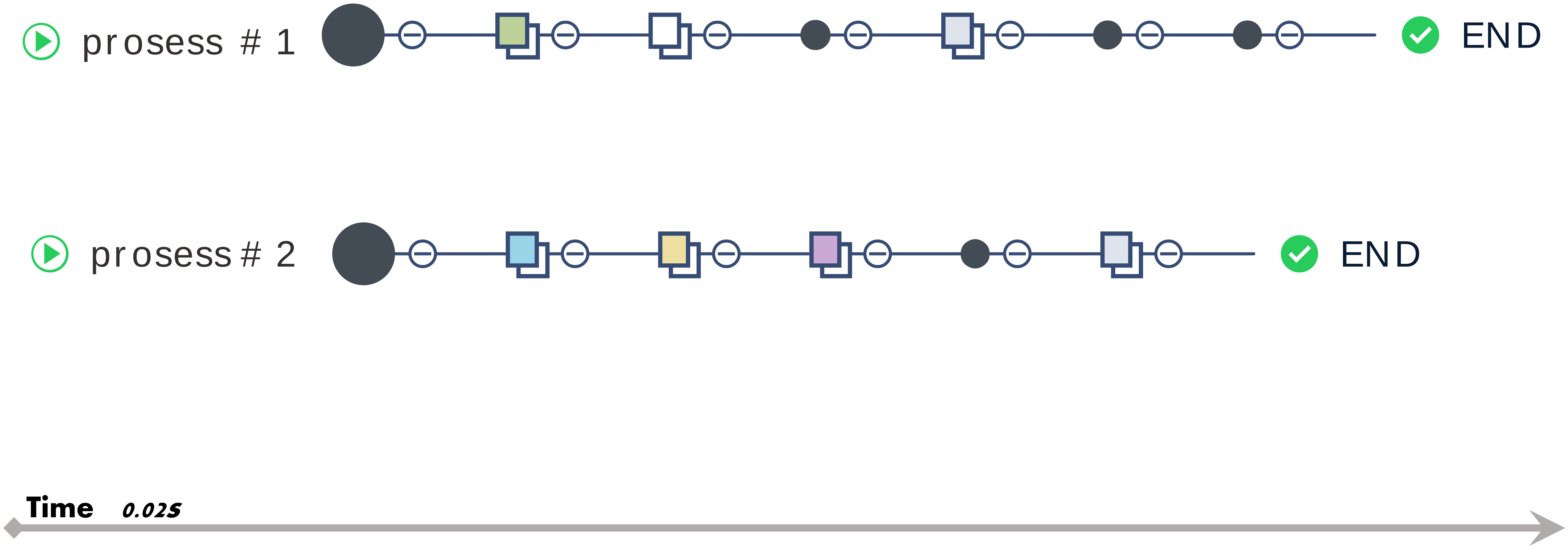
Concurrency is about dealing with lots of things at once.

Parallelism is about doing lots of things at once.

理解多线程


Thread # 1
END
Thread # 2
END
Thread # 3
END
Thread # ...
Thread # n
END
Time
0.025

理解进程



prosess # 1

END

prosess # 2

END

Time *0.025*