# Python 数据分析与应用

## 实验二：网络爬虫

丁烨

dingye@dgut.edu.cn

网络空间安全学院

2023-10-18

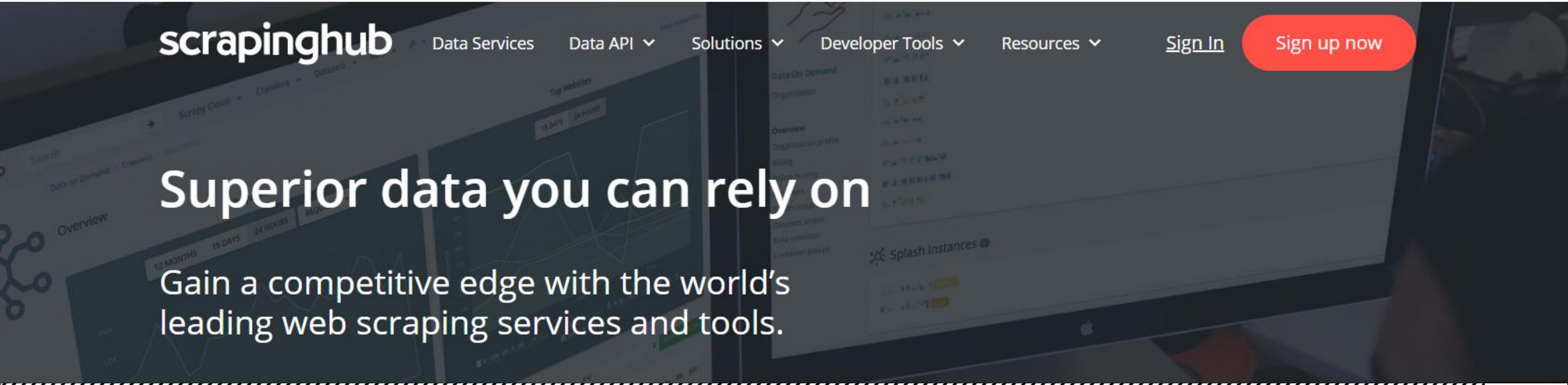
東莞理工學院

DONGGUAN UNIVERSITY OF TECHNOLOGY

# 爬虫框架

## Scrapy

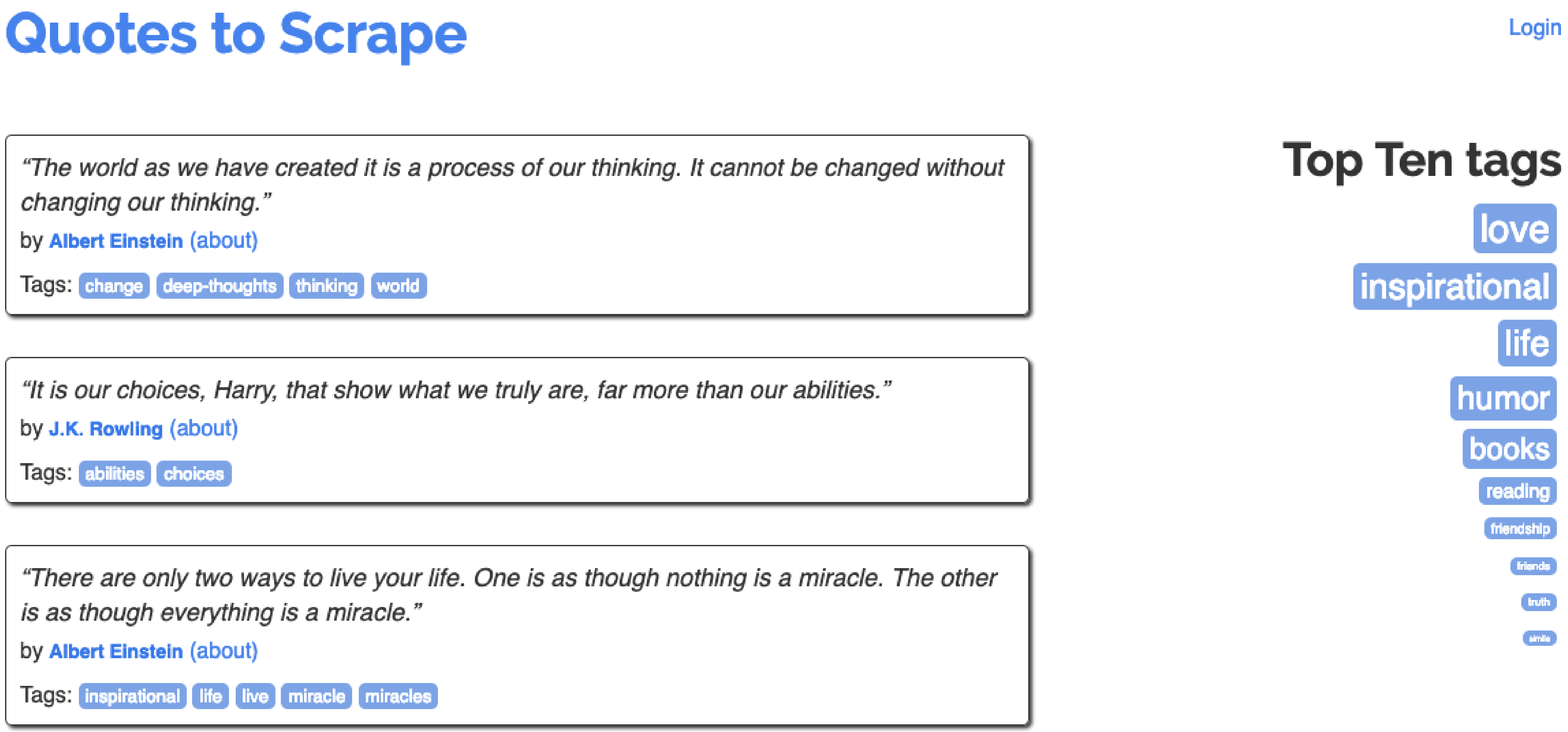
- Quotes to Scrape
- http://quotes.toscrape.com/
- Scrapy“官方”练习网站
- 由 Scrapinghub 提供：http://www.scrapinghub.com/



Superior data you can rely on

Gain a competitive edge with the world's leading web scraping services and tools.

## Scrapy



# Quotes to Scrape

Login

“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”

by **Albert Einstein** (about)

Tags: change deep-thoughts thinking world

“It is our choices, Harry, that show what we truly are, far more than our abilities.”

by **J.K. Rowling** (about)

Tags: abilities choices

“There are only two ways to live your life. One is as though nothing is a miracle. The other is as though everything is a miracle.”

by **Albert Einstein** (about)

Tags: inspirational life live miracle miracles

## Top Ten tags

love

inspirational

life

humor

books

reading

friendship

friends

truth

simile

# 爬虫框架

## Scrapy

- Scrapy
- https://scrapy.org/
- 一个基于 Python 开发的爬虫框架

An open source and collaborative framework
for extracting the data you need from websites.
In a fast, simple, yet extensible way.

# 爬虫框架

## Scrapy

- 安装 Scrapy

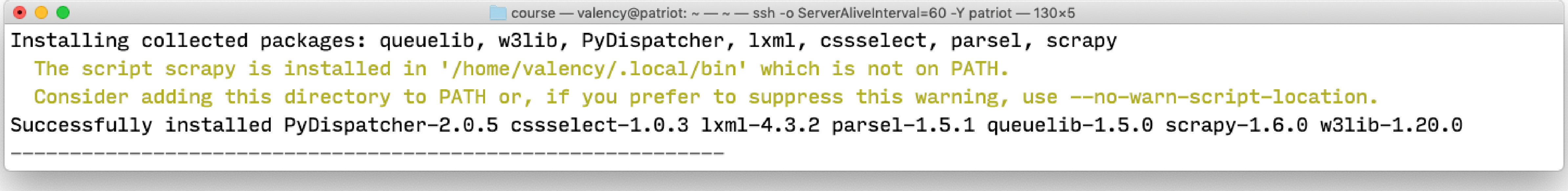
- 安装依赖：
- sudo apt install libxml2-dev libxslt1-dev zlib1g-dev libffi-dev libssl-dev

- 使用 pip 安装 Scrapy:
- pip3 install -U scrapy

# 爬虫框架

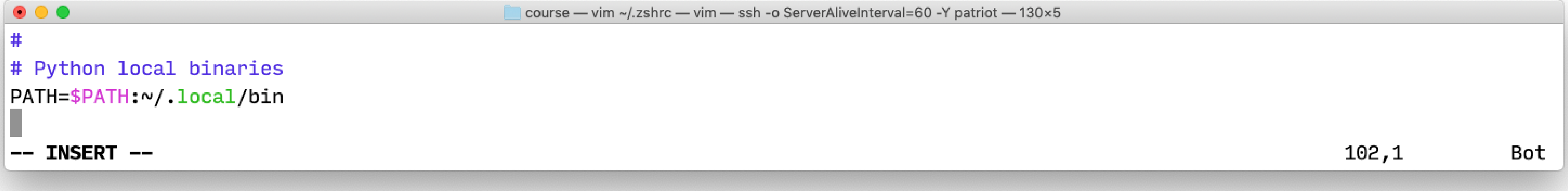
## Scrapy

- 如果在安装过程中出现如下警告：


course — valency@patriot: ~ — ~ — ssh -o ServerAliveInterval=60 -Y patriot — 130×5

```
Installing collected packages: queuelib, w3lib, PyDispatcher, lxml, cssselect, parsel, scrapy
  The script scrapy is installed in '/home/valency/.local/bin' which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed PyDispatcher-2.0.5 cssselect-1.0.3 lxml-4.3.2 parsel-1.5.1 queuelib-1.5.0 scrapy-1.6.0 w3lib-1.20.0
------------------------------------------------------------
```

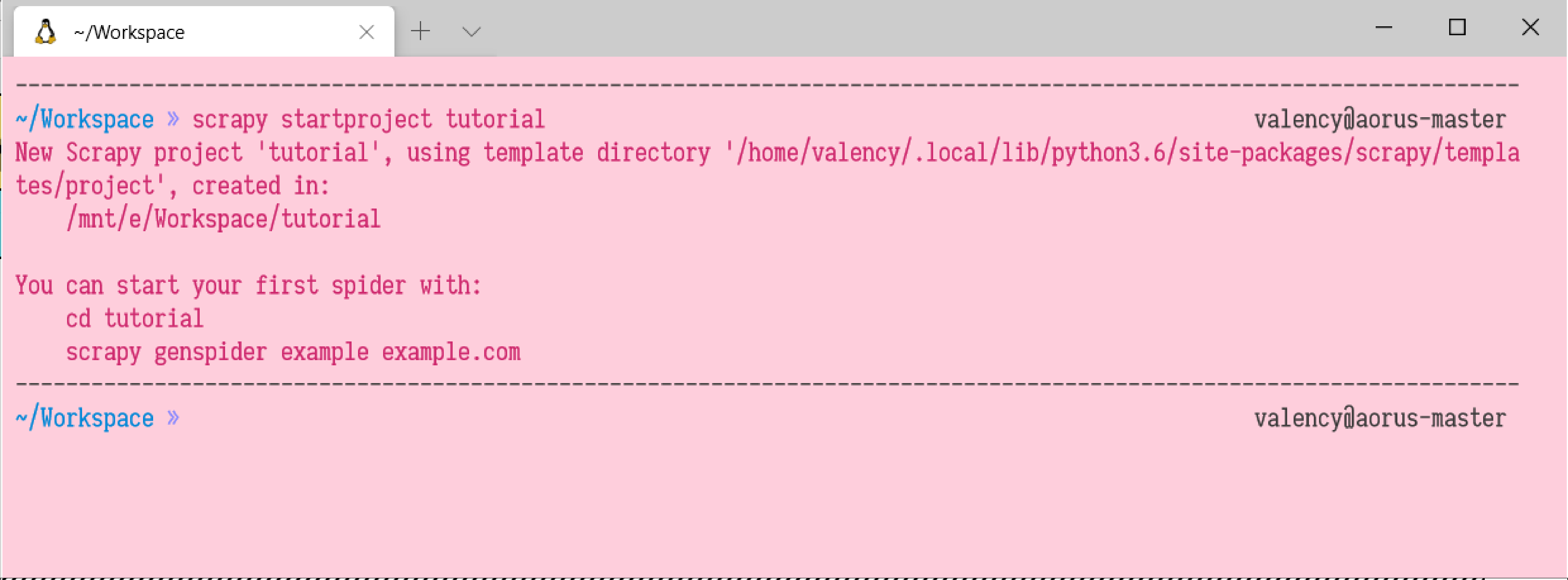
- 则需要添加环境变量：
- vim ~/.bashrc ➔ 添加：PATH=$PATH:~/.local/bin
- 然后退出重新连接，或：source ~/.bashrc


course — vim ~/.zshrc — vim — ssh -o ServerAliveInterval=60 -Y patriot — 130×5

```
#
# Python local binaries
PATH=$PATH:~/.local/bin

-- INSERT --                                                                                    102,1         Bot
```

## ❖ 创建 Scrapy 项目

❖ scrapy startproject <name>



```
--------------------------------------------------------------------------------------------
~/Workspace » scrapy startproject tutorial                            valency@aorus-master
New Scrapy project 'tutorial', using template directory '/home/valency/.local/lib/python3.6/site-packages/scrapy/templa
tes/project', created in:
    /mnt/e/Workspace/tutorial

You can start your first spider with:
    cd tutorial
    scrapy genspider example example.com
--------------------------------------------------------------------------------------------
~/Workspace »                                                        valency@aorus-master
```

# 爬虫框架

## Scrapy

```
tutorial/
    scrapy.cfg            # 配置文件，用于存储项目的配置信息
    tutorial/             # 项目的 Python 模块，将会从这里开始引用代码
        __init__.py
        items.py          # 实体文件，用于定义项目的目标实体
        middlewares.py    # 中间件文件，用于定义 Spider 中间件
        pipelines.py      # 管道文件，用于定义项目使用的各种管道
        settings.py       # 设置文件，用于存储项目的设置信息
        spiders/          # 存储爬虫代码的目录
            __init__.py
```

# 爬虫框架

## Scrapy

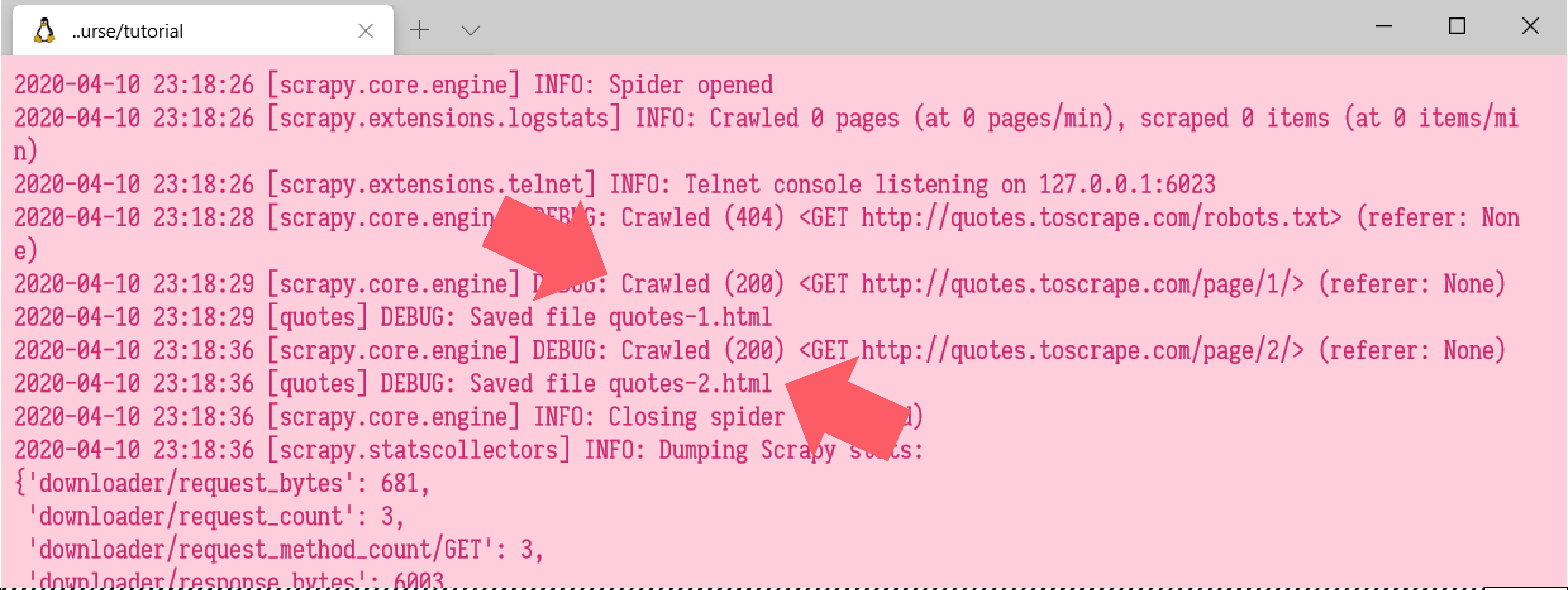
❖ 在 spiders 目录下创建 quotes_spider.py 文件:

```
import scrapy
class QuotesSpider(scrapy.Spider):
    name = "quotes"

    def start_requests(self):
        urls = ['http://quotes.toscrape.com/page/1/', 'http://quotes.toscrape.com/page/2/']
        for url in urls:
            yield scrapy.Request(url=url, callback=self.parse)

    def parse(self, response):
        page = response.url.split("/")[-2]
        filename = 'quotes-%s.html' % page
        with open(filename, 'wb') as f:
            f.write(response.body)
        self.log('Saved file %s' % filename)
```

# 爬虫框架

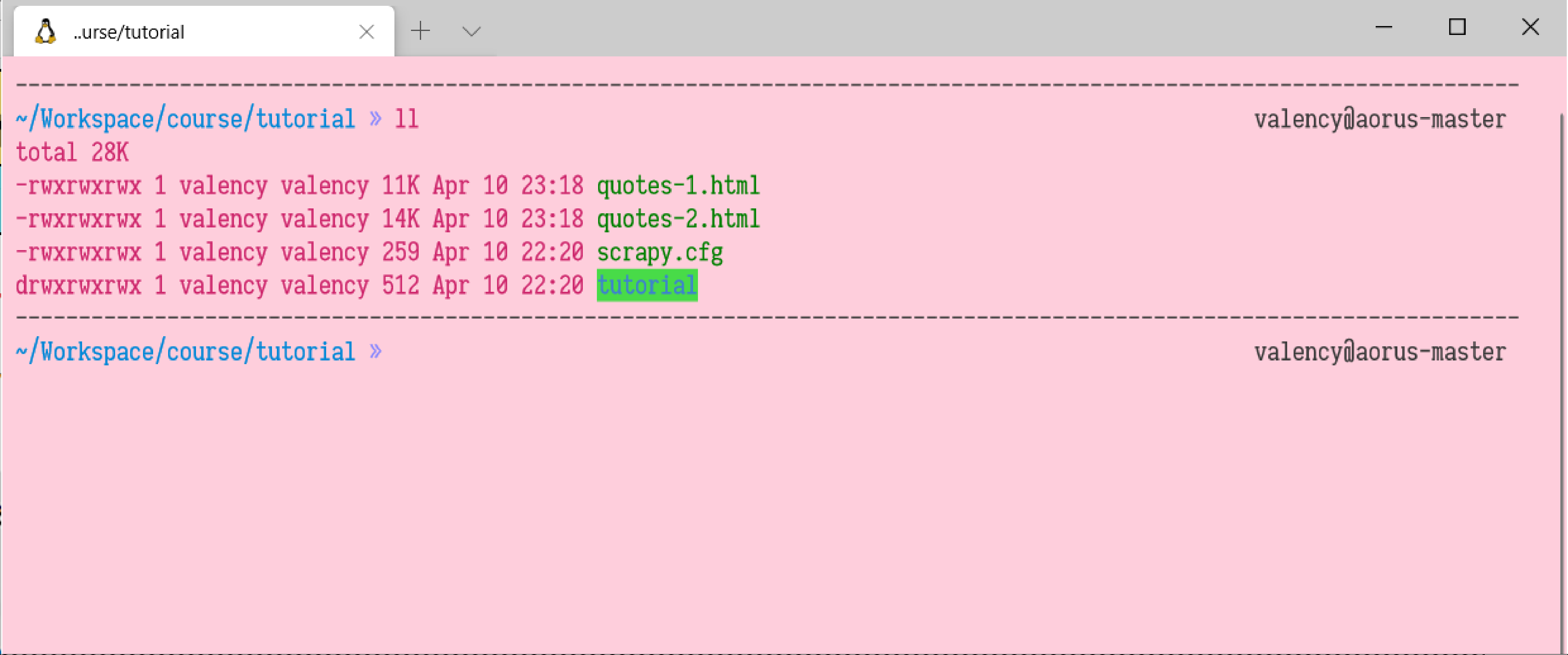
## Scrapy

- 运行 Scrapy 爬虫
- scrapy crawl quotes



```
2020-04-10 23:18:26 [scrapy.core.engine] INFO: Spider opened
2020-04-10 23:18:26 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/mi
n)
2020-04-10 23:18:26 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2020-04-10 23:18:28 [scrapy.core.engine] DEBUG: Crawled (404) <GET http://quotes.toscrape.com/robots.txt> (referer: Non
e)
2020-04-10 23:18:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/1/> (referer: None)
2020-04-10 23:18:29 [quotes] DEBUG: Saved file quotes-1.html
2020-04-10 23:18:36 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/2/> (referer: None)
2020-04-10 23:18:36 [quotes] DEBUG: Saved file quotes-2.html
2020-04-10 23:18:36 [scrapy.core.engine] INFO: Closing spider (finished)
2020-04-10 23:18:36 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 681,
 'downloader/request_count': 3,
 'downloader/request_method_count/GET': 3,
 'downloader/response_bytes': 6003,
```

## ❖ 检查运行结果

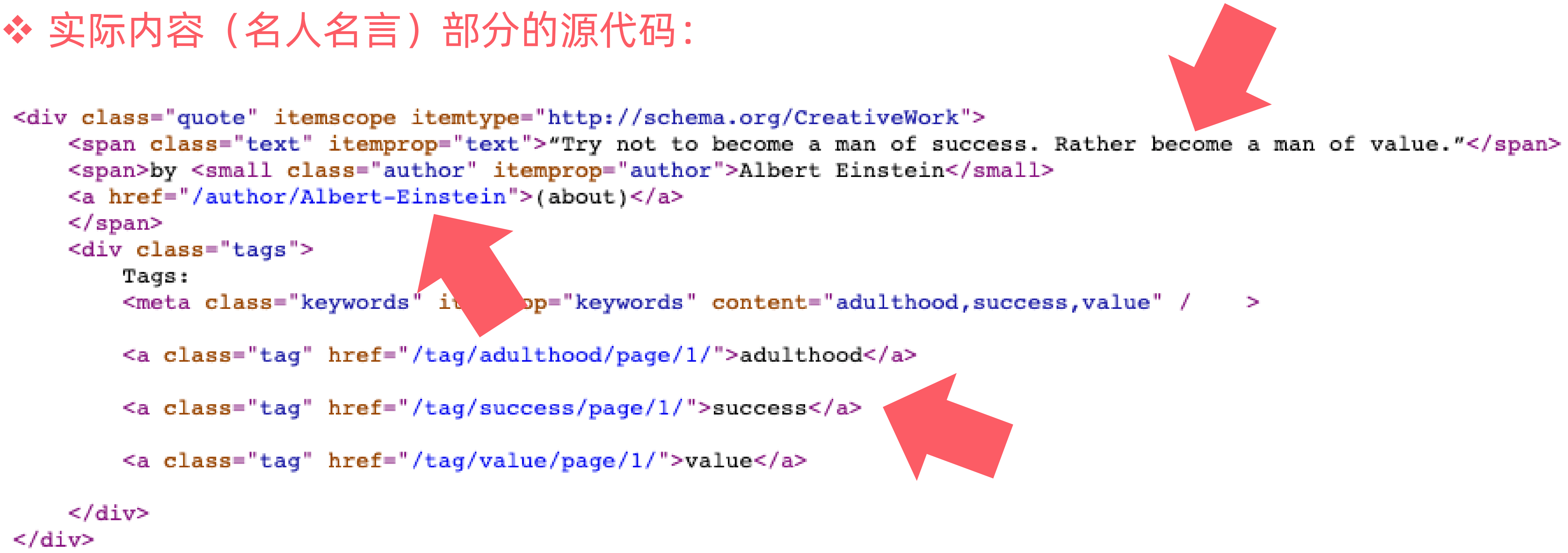


```
----------------------------------------------------------------------------------------------------------------------
~/Workspace/course/tutorial » ll                                                              valency@aorus-master
total 28K
-rwxrwxrwx 1 valency valency 11K Apr 10 23:18 quotes-1.html
-rwxrwxrwx 1 valency valency 14K Apr 10 23:18 quotes-2.html
-rwxrwxrwx 1 valency valency 259 Apr 10 22:20 scrapy.cfg
drwxrwxrwx 1 valency valency 512 Apr 10 22:20 tutorial
----------------------------------------------------------------------------------------------------------------------
~/Workspace/course/tutorial »                                                                 valency@aorus-master
```

# 爬虫框架

## Scrapy

- 通过前面两个步骤，我们已经成功爬取到了网页的源码
- 要想提取数据，需要先观察页面源码，定位目标数据，分析和了解目标数据的展示结构
- 实际内容（名人名言）部分的源代码：

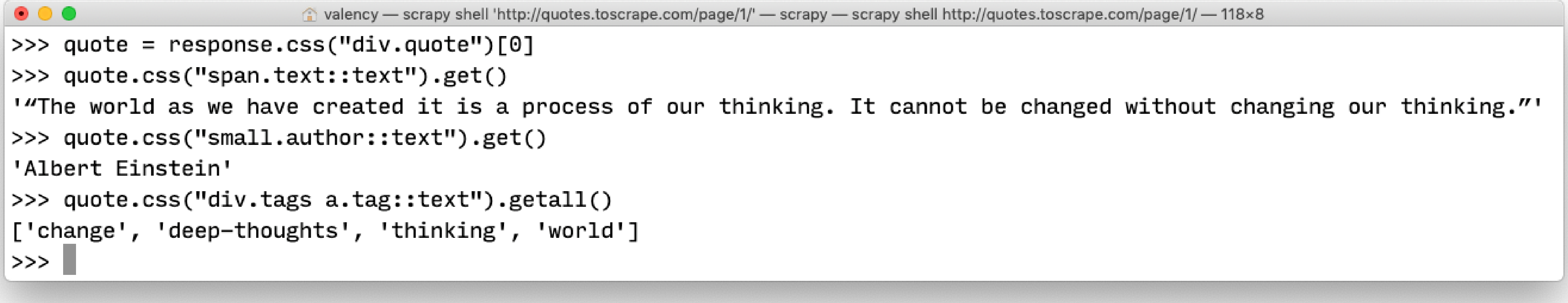


```
<div class="quote" itemscope itemtype="http://schema.org/CreativeWork">
    <span class="text" itemprop="text">“Try not to become a man of success. Rather become a man of value.”</span>
    <span>by <small class="author" itemprop="author">Albert Einstein</small>
    <a href="/author/Albert-Einstein">(about)</a>
    </span>
    <div class="tags">
        Tags:
        <meta class="keywords" it      op="keywords" content="adulthood,success,value" /    >

        <a class="tag" href="/tag/adulthood/page/1/">adulthood</a>

        <a class="tag" href="/tag/success/page/1/">success</a>

        <a class="tag" href="/tag/value/page/1/">value</a>

    </div>
</div>
```

# 爬虫框架

## Scrapy

- 使用 Scrapy Shell 探索需要解析的数据：
- scrapy shell 'http://quotes.toscrape.com/page/1/'
- quote = response.css("div.quote")[0]
- quote.css("span.text::text").get()
- quote.css("small.author::text").get()
- quote.css("div.tags a.tag::text").getall()



valency — scrapy shell 'http://quotes.toscrape.com/page/1/' — scrapy — scrapy shell http://quotes.toscrape.com/page/1/ — 118×8

```
>>> quote = response.css("div.quote")[0]
>>> quote.css("span.text::text").get()
'“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”'
>>> quote.css("small.author::text").get()
'Albert Einstein'
>>> quote.css("div.tags a.tag::text").getall()
['change', 'deep-thoughts', 'thinking', 'world']
>>>
```

# 爬虫框架

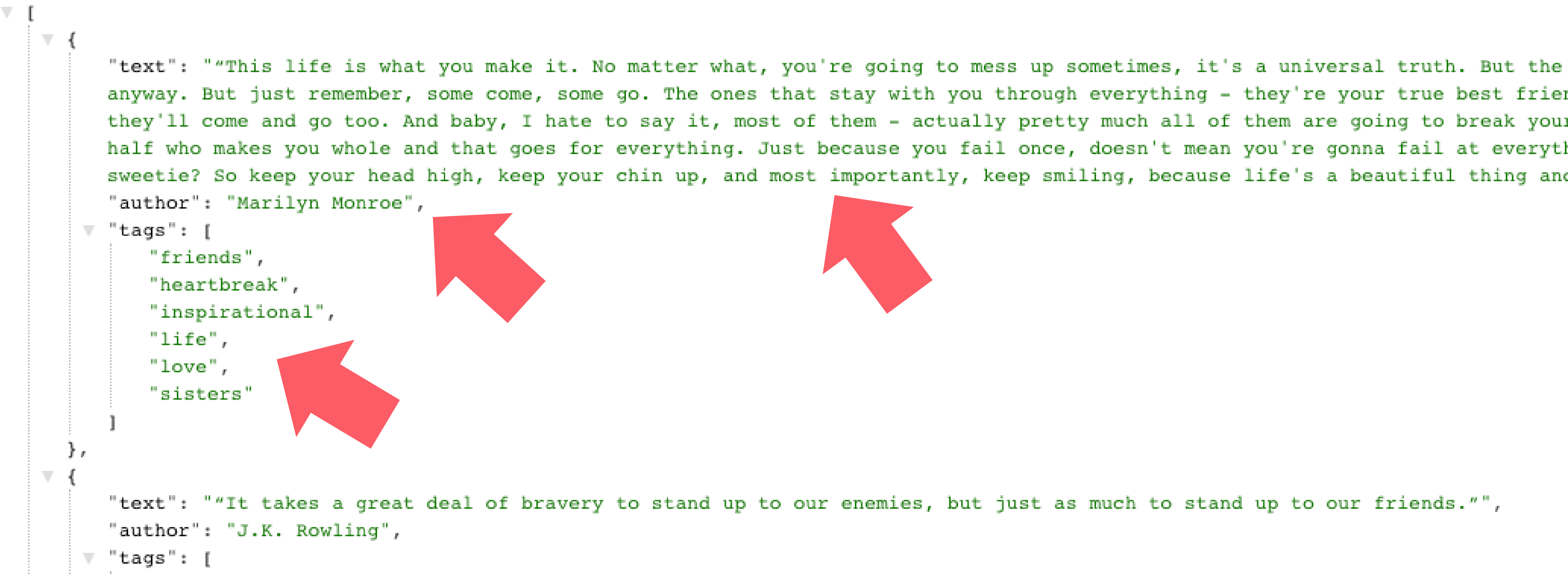
## Scrapy

❖ 修改 quotes_spider.py 文件:

```
...

    def parse(self, response):
        for quote in response.css('div.quote'):
            yield {
                'text': quote.css('span.text::text').get(),
                'author': quote.css('small.author::text').get(),
                'tags': quote.css('div.tags a.tag::text').getall(),
            }

...
```

## ❖ 运行 Scrapy 爬虫

❖ `scrapy crawl quotes -o quotes.json`



```
[
    {
        "text": "“This life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the
        anyway. But just remember, some come, some go. The ones that stay with you through everything - they're your true best frien
        they'll come and go too. And baby, I hate to say it, most of them - actually pretty much all of them are going to break your
        half who makes you whole and that goes for everything. Just because you fail once, doesn't mean you're gonna fail at everyth
        sweetie? So keep your head high, keep your chin up, and most importantly, keep smiling, because life's a beautiful thing and
        "author": "Marilyn Monroe",
        "tags": [
            "friends",
            "heartbreak",
            "inspirational",
            "life",
            "love",
            "sisters"
        ]
    },
    {
        "text": "“It takes a great deal of bravery to stand up to our enemies, but just as much to stand up to our friends.”",
        "author": "J.K. Rowling",
        "tags": [
```

❖ 加入分页识别机制：

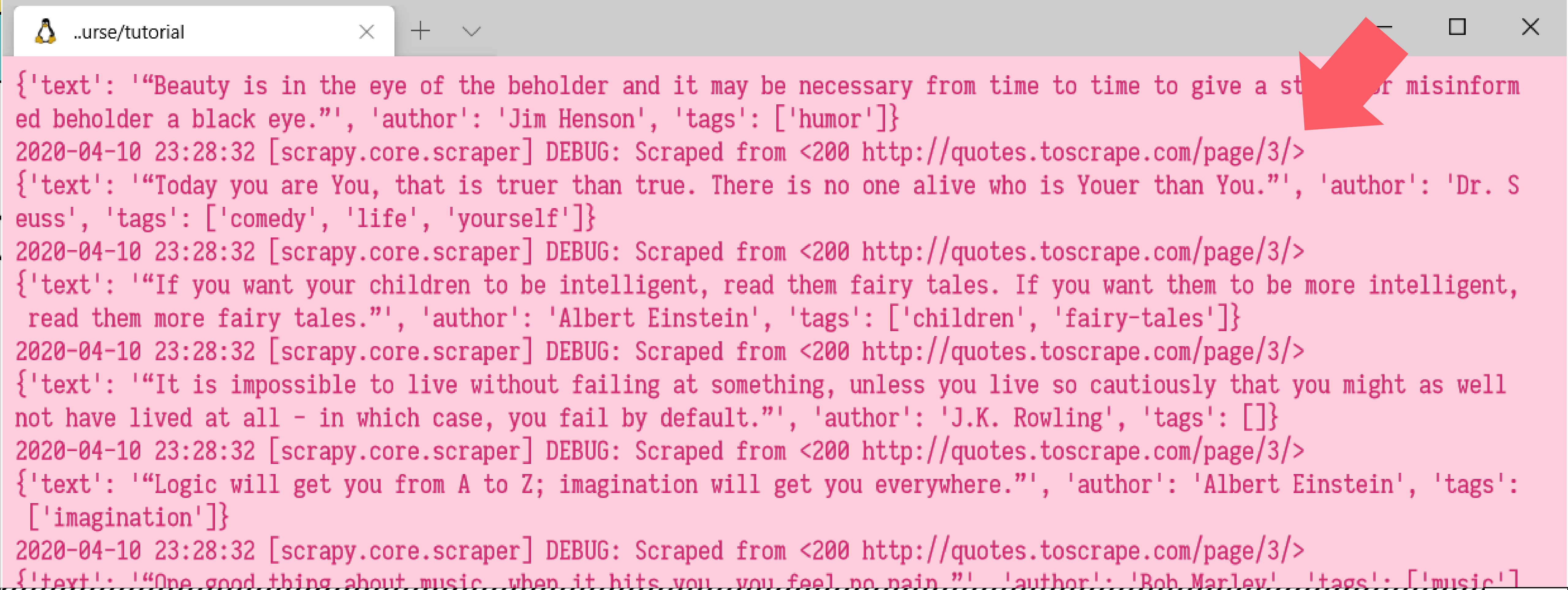
```
...

    def start_requests(self):
        urls = ['http://quotes.toscrape.com/page/1/']
        ...

    def parse(self, response):
        ...
        next_page = response.css('li.next a::attr(href)').get()
        if next_page is not None:
            yield response.follow(next_page, self.parse)

...
```

# 爬虫框架

## Scrapy

### ❖ 运行 Scrapy 爬虫

❖ scrapy crawl quotes -o quotes.json



```
{'text': '“Beauty is in the eye of the beholder and it may be necessary from time to time to give a st          r misinform
ed beholder a black eye.”', 'author': 'Jim Henson', 'tags': ['humor']}
2020-04-10 23:28:32 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/3/>
{'text': '“Today you are You, that is truer than true. There is no one alive who is Youer than You.”', 'author': 'Dr. S
euss', 'tags': ['comedy', 'life', 'yourself']}
2020-04-10 23:28:32 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/3/>
{'text': '“If you want your children to be intelligent, read them fairy tales. If you want them to be more intelligent,
 read them more fairy tales.”', 'author': 'Albert Einstein', 'tags': ['children', 'fairy-tales']}
2020-04-10 23:28:32 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/3/>
{'text': '“It is impossible to live without failing at something, unless you live so cautiously that you might as well
not have lived at all - in which case, you fail by default.”', 'author': 'J.K. Rowling', 'tags': []}
2020-04-10 23:28:32 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/3/>
{'text': '“Logic will get you from A to Z; imagination will get you everywhere.”', 'author': 'Albert Einstein', 'tags':
 ['imagination']}
2020-04-10 23:28:32 [scrapy.core.scraper] DEBUG: Scraped from <200 http://quotes.toscrape.com/page/3/>
{'text': '“One good thing about music, when it hits you, you feel no pain.”', 'author': 'Bob Marley', 'tags': ['music']
```

# 爬虫框架

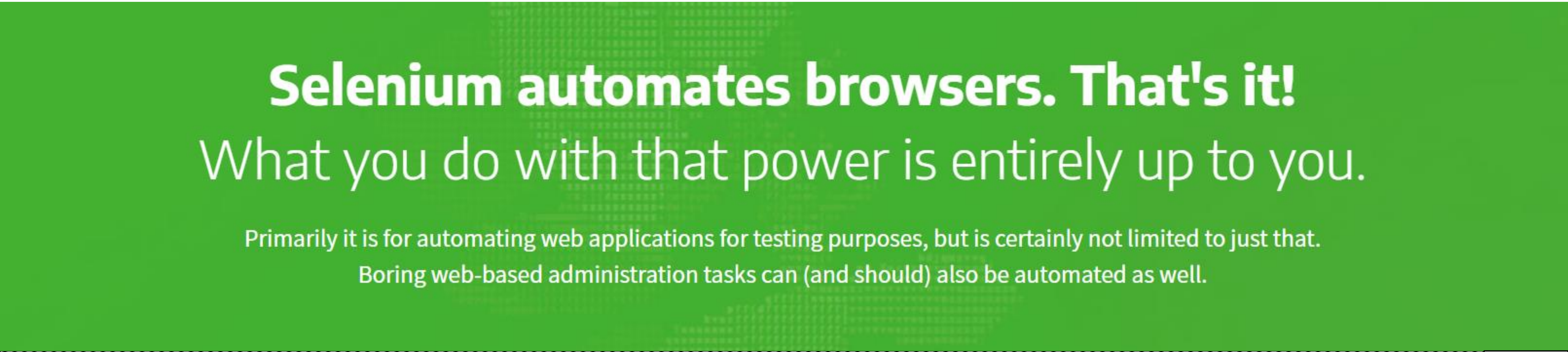
## Scrapy

- ❖ Scrapy 扩展阅读
- ❖ Scrapy 官方文档：
- ❖ https://docs.scrapy.org/
- ❖ HTML 5 速成指南：
- ❖ https://www.w3schools.com/html/
- ❖ QuotesBot：
- ❖ https://github.com/scrapy/quotesbot/
- ❖ Scrapy Cluster：
- ❖ https://github.com/istresearch/scrapy-cluster

# 爬虫框架

## 高级爬虫 / 自动测试框架 / 服务

- Selenium
- https://www.selenium.dev/
- Selenium 是一个综合性的项目，为浏览器的自动化提供了各种工具和依赖包
- Selenium IDE 是一个可录制再重放的自动化 Web 测试工具
- Selenium 为各种编程语言提供了 API，目前官方支持包括：C#、JavaScript、Java、Python、Ruby 等



**Selenium automates browsers. That's it!**

What you do with that power is entirely up to you.

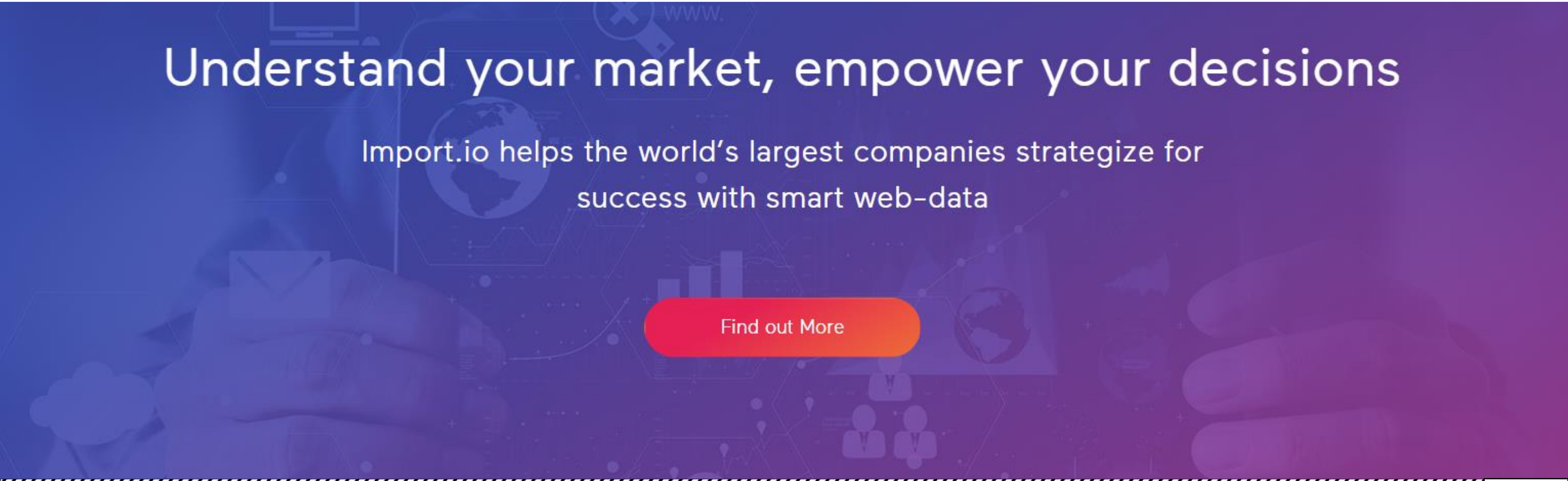
Primarily it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should) also be automated as well.

# 爬虫框架

## 高级爬虫 / 自动测试框架 / 服务

- Import.io
- https://www.import.io/
- 提供高级爬虫解决方案并交叉销售爬虫数据的互联网数据集成商（WDI）

import.io



Understand your market, empower your decisions

Import.io helps the world’s largest companies strategize for success with smart web-data

# 爬虫框架

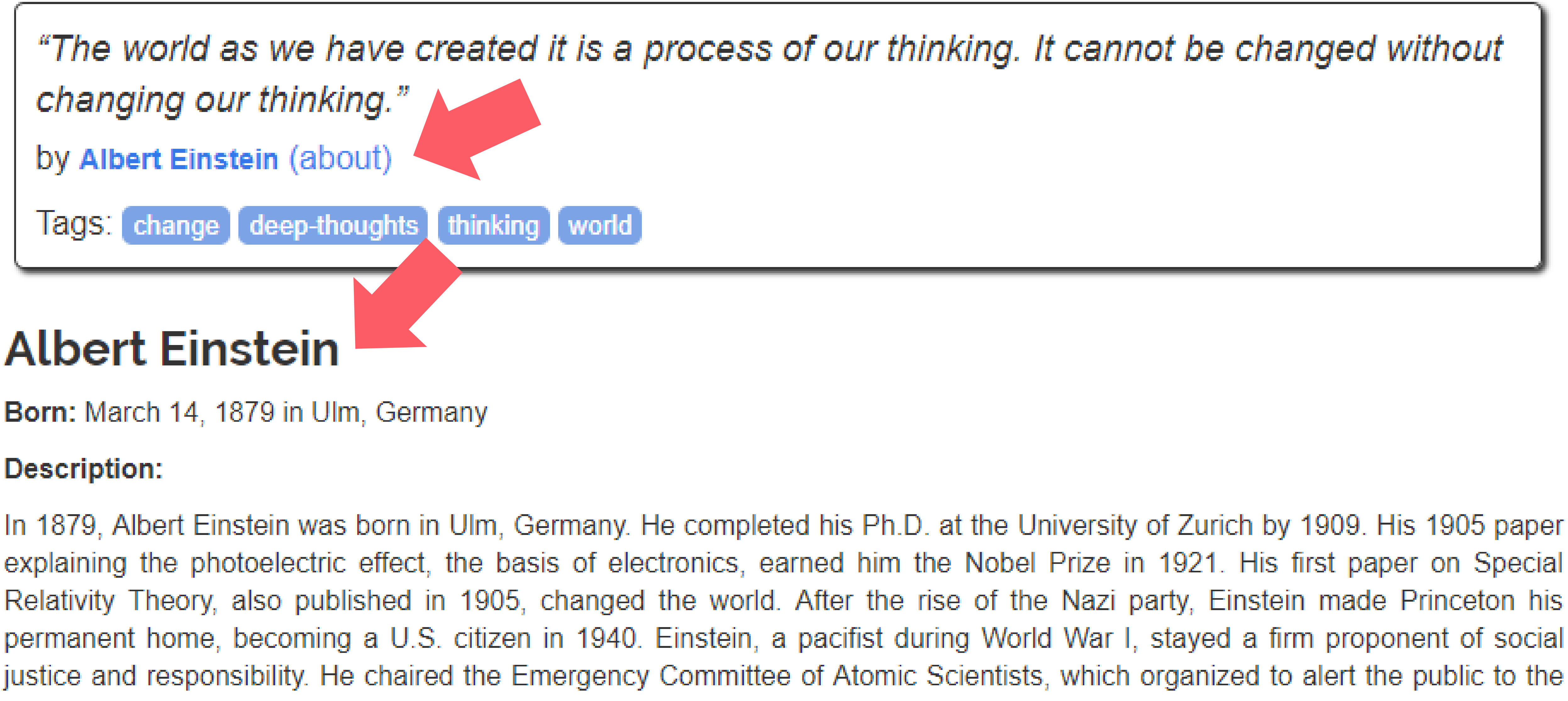
## 高级爬虫 / 自动测试框架 / 服务

- 八爪鱼
- https://www.bazhuayu.com/
- 提供高级爬虫解决方案并交叉销售爬虫数据的互联网数据集成商（WDI）

八爪鱼—全球百万用户信赖的数据采集器

# 作业

❖ 修改爬虫或使用高级爬虫工具爬取名人名言及其作者信息



*“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”*

by **Albert Einstein** (about)

Tags: change deep-thoughts thinking world

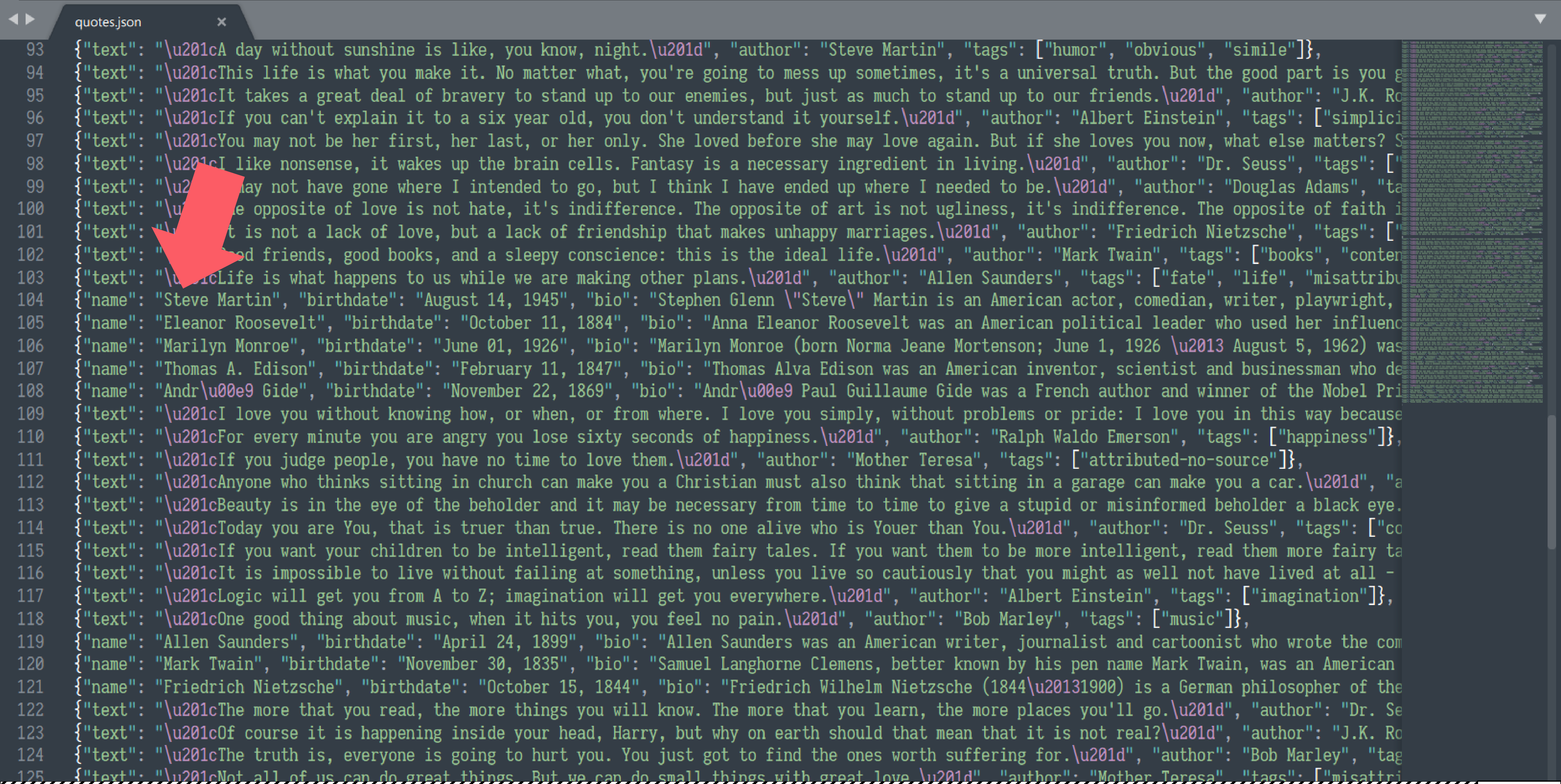
## Albert Einstein

**Born:** March 14, 1879 in Ulm, Germany

**Description:**

In 1879, Albert Einstein was born in Ulm, Germany. He completed his Ph.D. at the University of Zurich by 1909. His 1905 paper explaining the photoelectric effect, the basis of electronics, earned him the Nobel Prize in 1921. His first paper on Special Relativity Theory, also published in 1905, changed the world. After the rise of the Nazi party, Einstein made Princeton his permanent home, becoming a U.S. citizen in 1940. Einstein, a pacifist during World War I, stayed a firm proponent of social justice and responsibility. He chaired the Emergency Committee of Atomic Scientists, which organized to alert the public to the

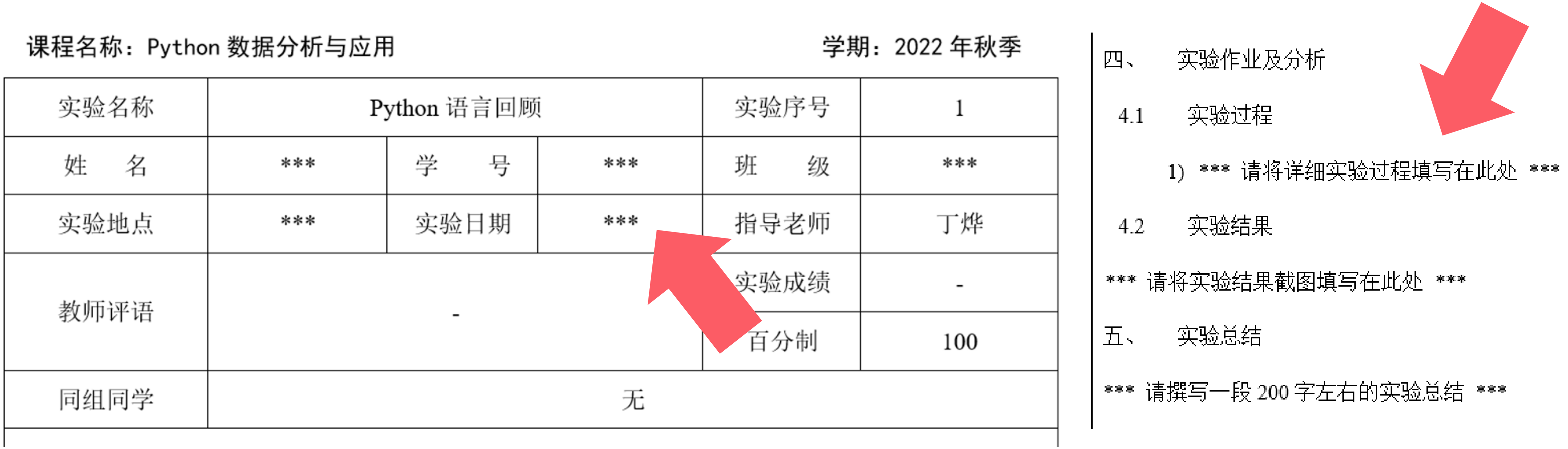
# 作业



```
{"text": "\u201cA day without sunshine is like, you know, night.\u201d", "author": "Steve Martin", "tags": ["humor", "obvious", "simile"]},
{"text": "\u201cThis life is what you make it. No matter what, you're going to mess up sometimes, it's a universal truth. But the good part is you g
{"text": "\u201cIt takes a great deal of bravery to stand up to our enemies, but just as much to stand up to our friends.\u201d", "author": "J.K. Ro
{"text": "\u201cIf you can't explain it to a six year old, you don't understand it yourself.\u201d", "author": "Albert Einstein", "tags": ["simplici
{"text": "\u201cYou may not be her first, her last, or her only. She loved before she may love again. But if she loves you now, what else matters? S
{"text": "\u201cI like nonsense, it wakes up the brain cells. Fantasy is a necessary ingredient in living.\u201d", "author": "Dr. Seuss", "tags": ["
{"text": "\u2   may not have gone where I intended to go, but I think I have ended up where I needed to be.\u201d", "author": "Douglas Adams", "ta
{"text": "\u   e opposite of love is not hate, it's indifference. The opposite of art is not ugliness, it's indifference. The opposite of faith i
{"text": "\    t is not a lack of love, but a lack of friendship that makes unhappy marriages.\u201d", "author": "Friedrich Nietzsche", "tags": ["
{"text": "     od friends, good books, and a sleepy conscience: this is the ideal life.\u201d", "author": "Mark Twain", "tags": ["books", "conten
{"text": "\u  cLife is what happens to us while we are making other plans.\u201d", "author": "Allen Saunders", "tags": ["fate", "life", "misattribu
{"name": "Steve Martin", "birthdate": "August 14, 1945", "bio": "Stephen Glenn \"Steve\" Martin is an American actor, comedian, writer, playwright,
{"name": "Eleanor Roosevelt", "birthdate": "October 11, 1884", "bio": "Anna Eleanor Roosevelt was an American political leader who used her influenc
{"name": "Marilyn Monroe", "birthdate": "June 01, 1926", "bio": "Marilyn Monroe (born Norma Jeane Mortenson; June 1, 1926 \u2013 August 5, 1962) was
{"name": "Thomas A. Edison", "birthdate": "February 11, 1847", "bio": "Thomas Alva Edison was an American inventor, scientist and businessman who de
{"name": "Andr\u00e9 Gide", "birthdate": "November 22, 1869", "bio": "Andr\u00e9 Paul Guillaume Gide was a French author and winner of the Nobel Pri
{"text": "\u201cI love you without knowing how, or when, or from where. I love you simply, without problems or pride: I love you in this way because
{"text": "\u201cFor every minute you are angry you lose sixty seconds of happiness.\u201d", "author": "Ralph Waldo Emerson", "tags": ["happiness"]},
{"text": "\u201cIf you judge people, you have no time to love them.\u201d", "author": "Mother Teresa", "tags": ["attributed-no-source"]},
{"text": "\u201cAnyone who thinks sitting in church can make you a Christian must also think that sitting in a garage can make you a car.\u201d", "a
{"text": "\u201cBeauty is in the eye of the beholder and it may be necessary from time to time to give a stupid or misinformed beholder a black eye.
{"text": "\u201cToday you are You, that is truer than true. There is no one alive who is Youer than You.\u201d", "author": "Dr. Seuss", "tags": ["co
{"text": "\u201cIf you want your children to be intelligent, read them fairy tales. If you want them to be more intelligent, read them more fairy ta
{"text": "\u201cIt is impossible to live without failing at something, unless you live so cautiously that you might as well not have lived at all -
{"text": "\u201cLogic will get you from A to Z; imagination will get you everywhere.\u201d", "author": "Albert Einstein", "tags": ["imagination"]},
{"text": "\u201cOne good thing about music, when it hits you, you feel no pain.\u201d", "author": "Bob Marley", "tags": ["music"]},
{"name": "Allen Saunders", "birthdate": "April 24, 1899", "bio": "Allen Saunders was an American writer, journalist and cartoonist who wrote the com
{"name": "Mark Twain", "birthdate": "November 30, 1835", "bio": "Samuel Langhorne Clemens, better known by his pen name Mark Twain, was an American
{"name": "Friedrich Nietzsche", "birthdate": "October 15, 1844", "bio": "Friedrich Wilhelm Nietzsche (1844\u20131900) is a German philosopher of the
{"text": "\u201cThe more that you read, the more things you will know. The more that you learn, the more places you'll go.\u201d", "author": "Dr. Se
{"text": "\u201cOf course it is happening inside your head, Harry, but why on earth should that mean that it is not real?\u201d", "author": "J.K. Ro
{"text": "\u201cThe truth is, everyone is going to hurt you. You just got to find the ones worth suffering for.\u201d", "author": "Bob Marley", "tag
{"text": "\u201cNot all of us can do great things. But we can do small things with great love.\u201d", "author": "Mother Teresa", "tags": ["misattri
```

# 作业

- ❖ 在作业系统中下载并完成本实验课对应实验报告
- ❖ https://hw.dgut.edu.cn/
- ❖ 注意：所有标识为 * 的地方都需要填写
- ❖ 截止日期：2023-10-25 23:59



课程名称：Python 数据分析与应用　　　　学期：2022 年秋季

| | | | | | |
|---|---|---|---|---|---|
| 实验名称 | Python 语言回顾 | | | 实验序号 | 1 |
| 姓　名 | *** | 学　号 | *** | 班　级 | *** |
| 实验地点 | *** | 实验日期 | *** | 指导老师 | 丁烨 |
| 教师评语 | - | | | 实验成绩 | - |
| | | | | 百分制 | 100 |
| 同组同学 | 无 | | | | |

四、　实验作业及分析

4.1　实验过程

1) *** 请将详细实验过程填写在此处 ***

4.2　实验结果

*** 请将实验结果截图填写在此处 ***

五、　实验总结

*** 请撰写一段 200 字左右的实验总结 ***

GOOD
LUCK!