云计算与大数据应用开发

实验三：网络爬虫

丁烨
dingye@dgut.edu.cn
计算机科学与技术学院

2024-04-15

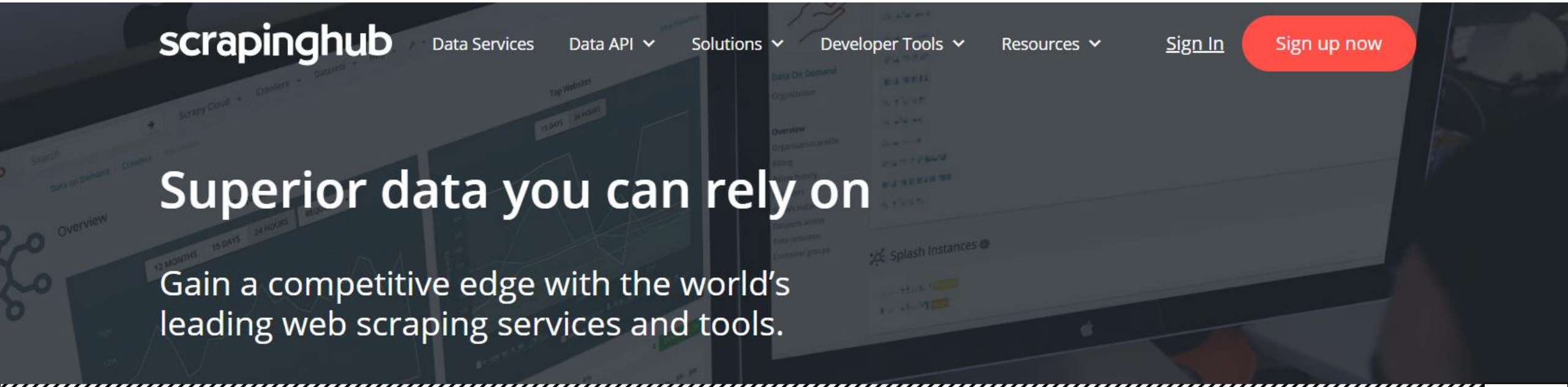
東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

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/



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
smile

爬虫框架

Scrapy

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

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

- 如果在安装过程中提醒~/.local/bin未加入环境变量
- 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

❖ 测试 Scrapy

❖ scrapy

..e/test/python ..space/test/js

```
~/Workspace/test/python » scrapy                                                valency@Zeppelin
Scrapy 2.7.1 - no active project

Usage:
  scrapy <command> [options] [args]

Available commands:
  bench         Run quick benchmark test
  commands
  fetch         Fetch a URL using the Scrapy downloader
  genspider     Generate new spider using pre-defined templates
  runspider     Run a self-contained spider (without creating a project)
  settings      Get settings values
  shell         Interactive scraping console
  startproject  Create new project
  version       Print Scrapy version
  view          Open URL in browser, as seen by Scrapy
```

❖ 创建 Scrapy 项目

❖ scrapy startproject <name>

..e/test/python ..space/test/js

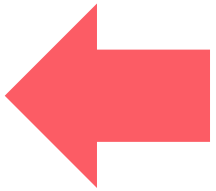
```
~/Workspace/test/python » scrapy startproject tutorial                                valency@Zeppelin
New Scrapy project 'tutorial', using template directory '/home/valency/.local/lib/python3.10/site-packages/scrapy/templates/project', creat
ed in:
    /mnt/e/Workspace/test/python/tutorial

You can start your first spider with:
    cd tutorial
    scrapy genspider example example.com
~/Workspace/test/python »                                                             valency@Zeppelin
```

❖ Scrapy 将会创建如下的目录结构：

```
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
```

❖ 在 tutorial/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)
```



- 运行 Scrapy 爬虫
- scrapy crawl quotes

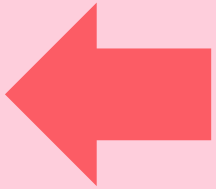
..thon/tutorial ..space/test/js

```
2023-02-01 17:27:46 [scra    ore.engine] DEBUG: Crawled (404) <GET http://quotes.toscrape.com/robots.txt> (referer: None)
2023-02-01 17:27:46 [sc      ngine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/1/> (referer: None)
2023-02-01 17:27:46 [quotes]      Saved file quotes-1.html
2023-02-01 17:27:47 [scrapy.core.engine] DEBUG: Crawled (200) <GET http://quotes.toscrape.com/page/2/> (referer: None)
2023-02-01 17:27:47 [quotes] DEBUG: Saved file quotes-2.html
2023-02-01 17:27:47 [scrapy.core.engine] INFO: Closing spider (finished)
2023-02-01 17:27:47 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/exception_count': 1,
 'downloader/exception_type_count/twisted.internet.error.TCPTimedOutError': 1,
 'downloader/request_bytes': 926,
 'downloader/request_count': 4,
 'downloader/request_method_count/GET': 4,
 'downloader/response_bytes': 5359,
 'downloader/response_count': 3,
 'downloader/response_status_count/200': 2,
 'downloader/response_status_count/404': 1,
 'elapsed_time_seconds': 132.807208,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2023, 2, 1, 9, 27, 47, 729376)
```

❖ 检查运行结果

..thon/tutorial ..space/test/js

```
~/Workspace/test/python/tutorial » ll                                    valency@Zeppelin
total 28K
-rwxrwxrwx 1 valency valency   0 Feb  1 17:24 __init__.py
-rwxrwxrwx 1 valency valency 11K Feb  1 17:27 quotes-1.html
-rwxrwxrwx 1 valency valency 14K Feb  1 17:27 quotes-2.html
-rwxrwxrwx 1 valency valency 259 Feb  1 15:49 scrapy.cfg
drwxrwxrwx 1 valency valency 512 Feb  1 17:25 tutorial
~/Workspace/test/python/tutorial »                                       valency@Zeppelin
```



- 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" itemprop="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>
```

❖ 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(),
            }

...
```

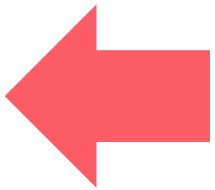
❖ 运行 Scrapy 爬虫

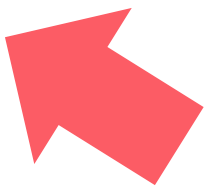
❖ scrapy crawl quotes -o quotes.json


```
[
  {
    "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 frie
    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 you
    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 an
    "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": [
```

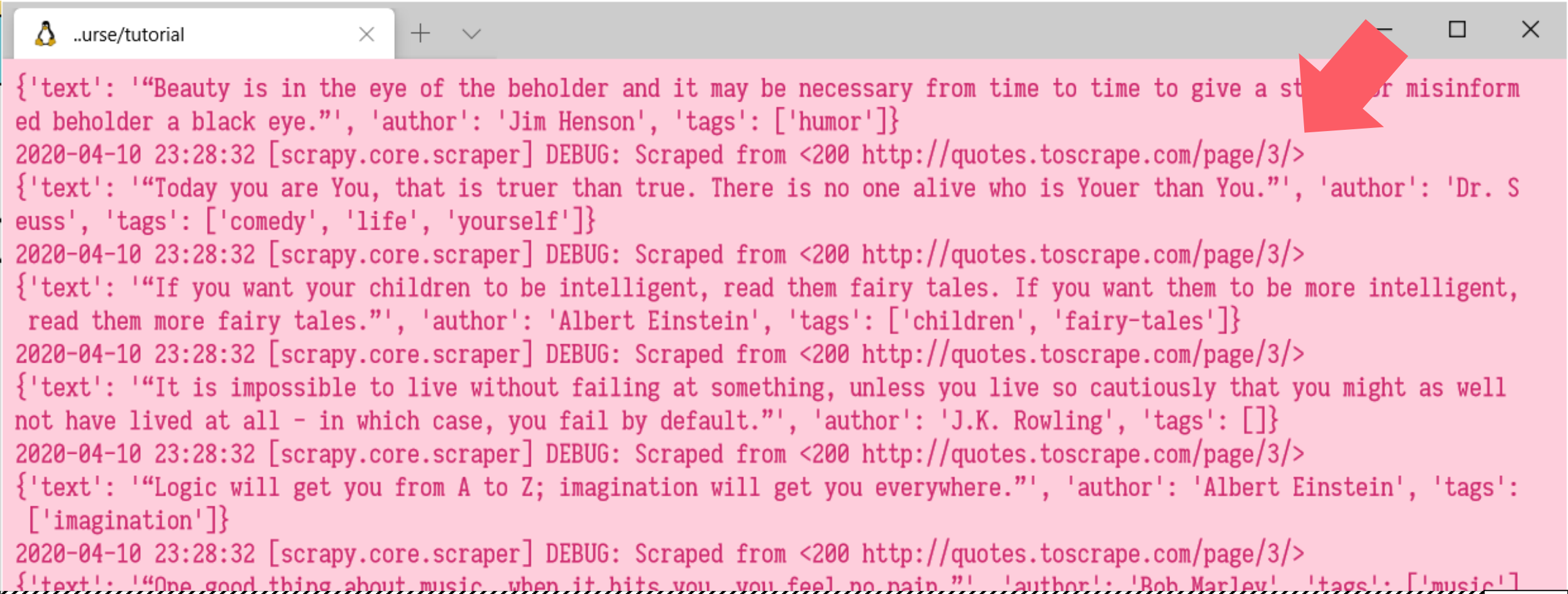
❖ 加入分页识别机制：

```
...

    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)

...
```





- 运行 Scrapy 爬虫
- `scrapy crawl quotes -o quotes.json`



- 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

Zombie.js

- Zombie.js
- http://zombie.js.org/
- 一个基于 JavaScript / Node.js 开发的 Headless 爬虫 / 自动测试框架

Zombie.js

Insanely fast, headless full-stack testing using Node.js

latest v6.1.4 | see CHANGELOG | build passing | js.org zombie

The Bite

If you're going to write an insanely fast, headless browser, how can you not call it Zombie? Zombie it is.

Zombie.js is a lightweight framework for testing client-side JavaScript code in a simulated environment. No browser required.

爬虫框架

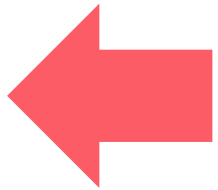
Zombie.js

- 无头软件（Headless Software）
- 是指一般具备图形用户界面（GUI）的软件的无图形用户界面版本
- 例如：Windows 操作系统、网络浏览器、视频播放软件等

- 无头浏览器（Headless Browser）
- 指的是没有图形用户界面的浏览器
- 无头浏览器通常用于自动化测试，通过远程指令控制浏览器自动浏览页面、填写表格等
- 在爬虫领域，由于部分网站源代码极其复杂（如使用了大量 JavaScript 脚本等），或网页使用了非常复杂的渲染机制，导致一般爬虫爬取数据难度较高时，使用无头浏览器可以大幅节省对网站分析的时间成本

Zombie.js

❖ Zombie.js 的基本用法：

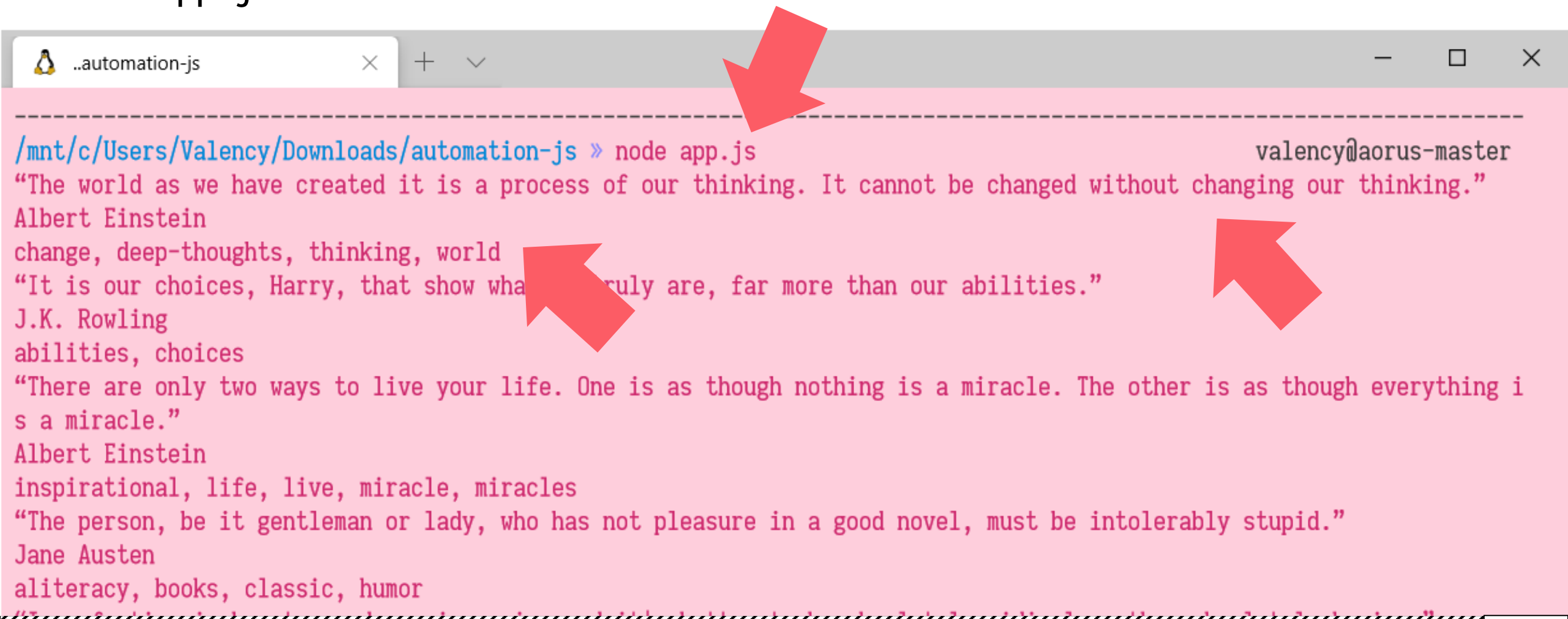


```
const Browser = require('zombie');

let browser = new Browser();
browser.visit('http://quotes.toscrape.com/', null, function () {
    for (let quote of browser.queryAll('.quote')) {
        console.log(browser.text(browser.query('span.text', quote)));
        console.log(browser.text(browser.query('small.author', quote)));
        let tags = [];
        for (let tag of browser.queryAll('a.tag', quote)) {
            tags.push(browser.text(tag));
        }
        console.log(tags.join(', '));
    }
});
```

爬虫框架

Zombie.js

- 运行 Zombie.js 爬虫
- node app.js

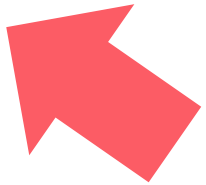


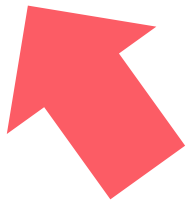
爬虫框架

Zombie.js

❖ 加入分页识别机制：

```
let urls = ['http://quotes.toscrape.com/page/1/'];
setInterval(() => {
    let url = urls.pop();
    if (url) {
        console.log(url);
        let browser = new Browser();
        browser.visit(url, null, function () {
            console.log(browser.status);
            ...
            urls.push(browser.query('a', browser.query('li.next')).href);
        });
    }
}, 1000);
```





爬虫框架

Zombie.js

- 运行 Zombie.js 爬虫
- node app.js

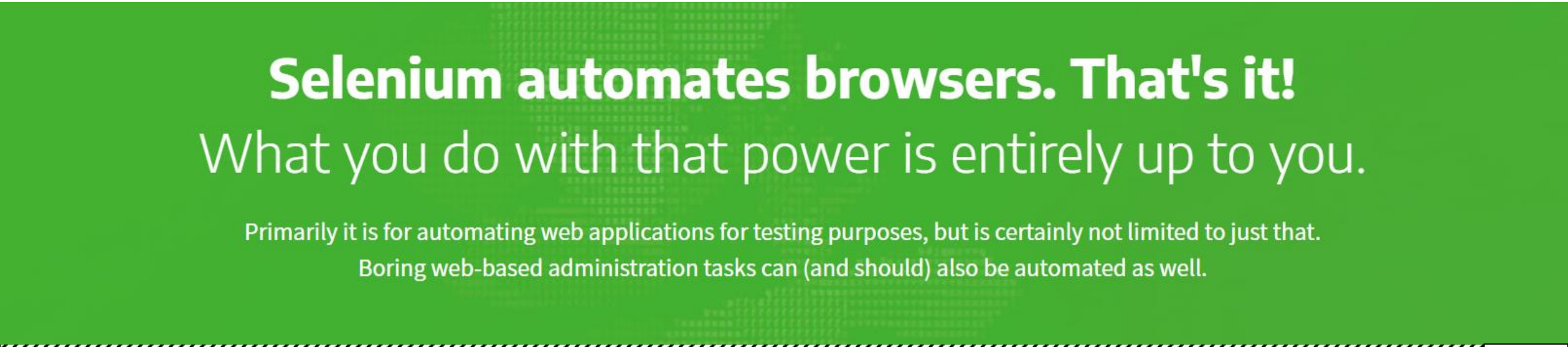
..automation-js

```
“A woman is like a tea bag; you never know how strong it is until it's in hot water.”
Eleanor Roosevelt
misattributed-eleanor-roosevelt
“A day without sunshine is like, you  night.”
Steve Martin
humor, obvious, simile
http://quotes.toscrape.com/page/2/
200
“This 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 get to decide how you're going to mess it up. Girls will be your friends - they'll act like it anyway. But
 just remember, some come, some go. The ones that stay with you through everything - they're your true best friends. Do
n't let go of them. Also remember, sisters make the best friends in the world. As for lovers, well, 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 heart, but yo
u can't give up because if you give up, you'll never find your soulmate.
You'll never find that half who makes you whole and that goes for everything. Just because you fail once, doesn't mean
you're gonna fail at everything. Keep trying, hold on, and always, always, always believe in yourself, because if you d
```

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

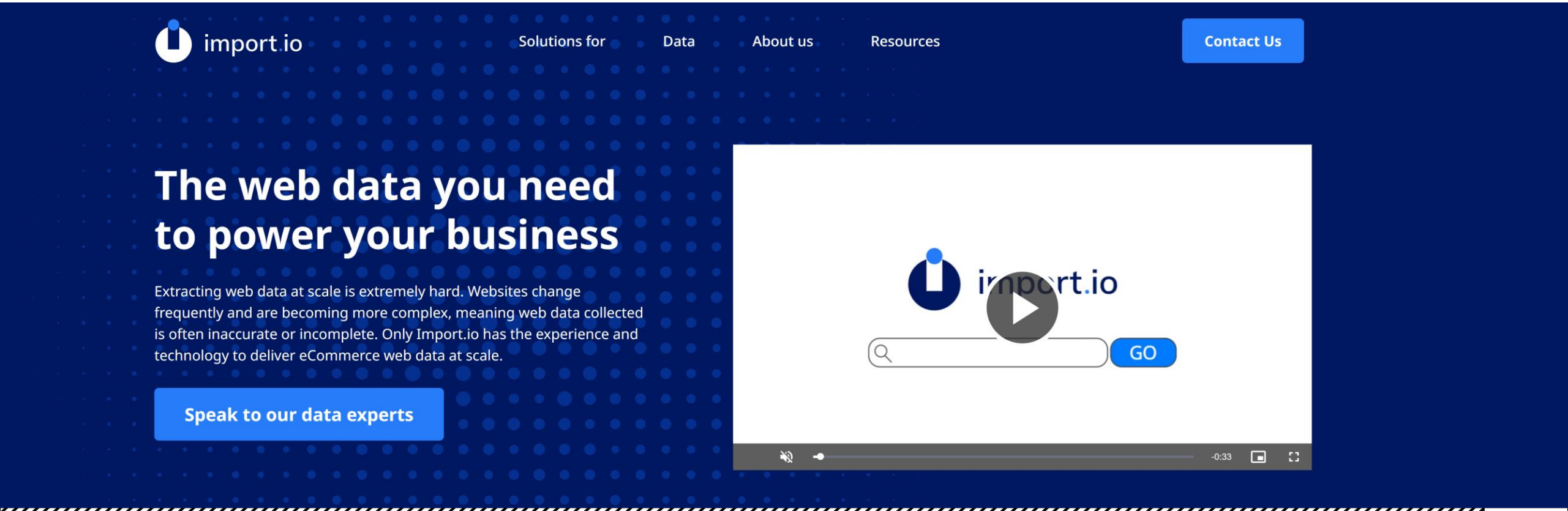


爬虫框架

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



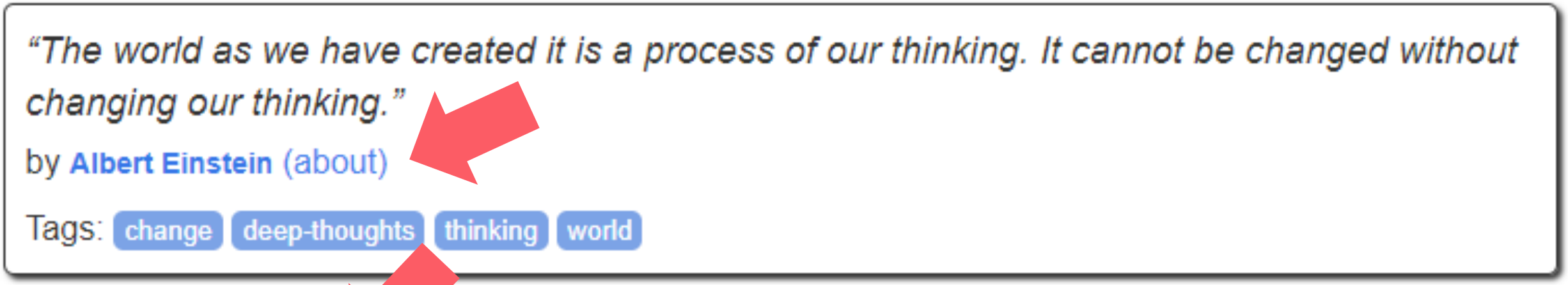
爬虫框架

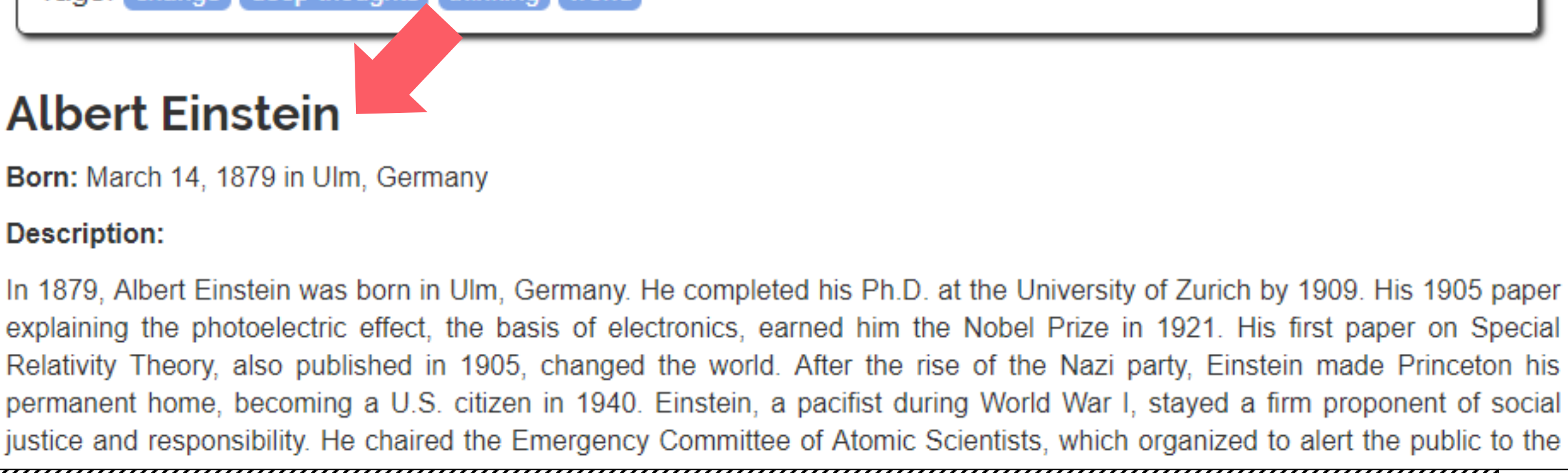
高级爬虫 / 自动测试框架 / 服务

- 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）

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





作业

quotes.json


```
{"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 ay 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": " ed 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：2024-04-22 23:59:59

课程名称：云计算与大数据应用开发　　学期：2023 年春季

实验名称	虚拟化技术			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	-
				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!