



Kotlin 爬虫

抓取多平台播客聆听数据实战



范圣佑
JetBrains 技术布道师
COSCon'23 (10/28-29)



《Kotlin 炉边漫谈》背后的故事

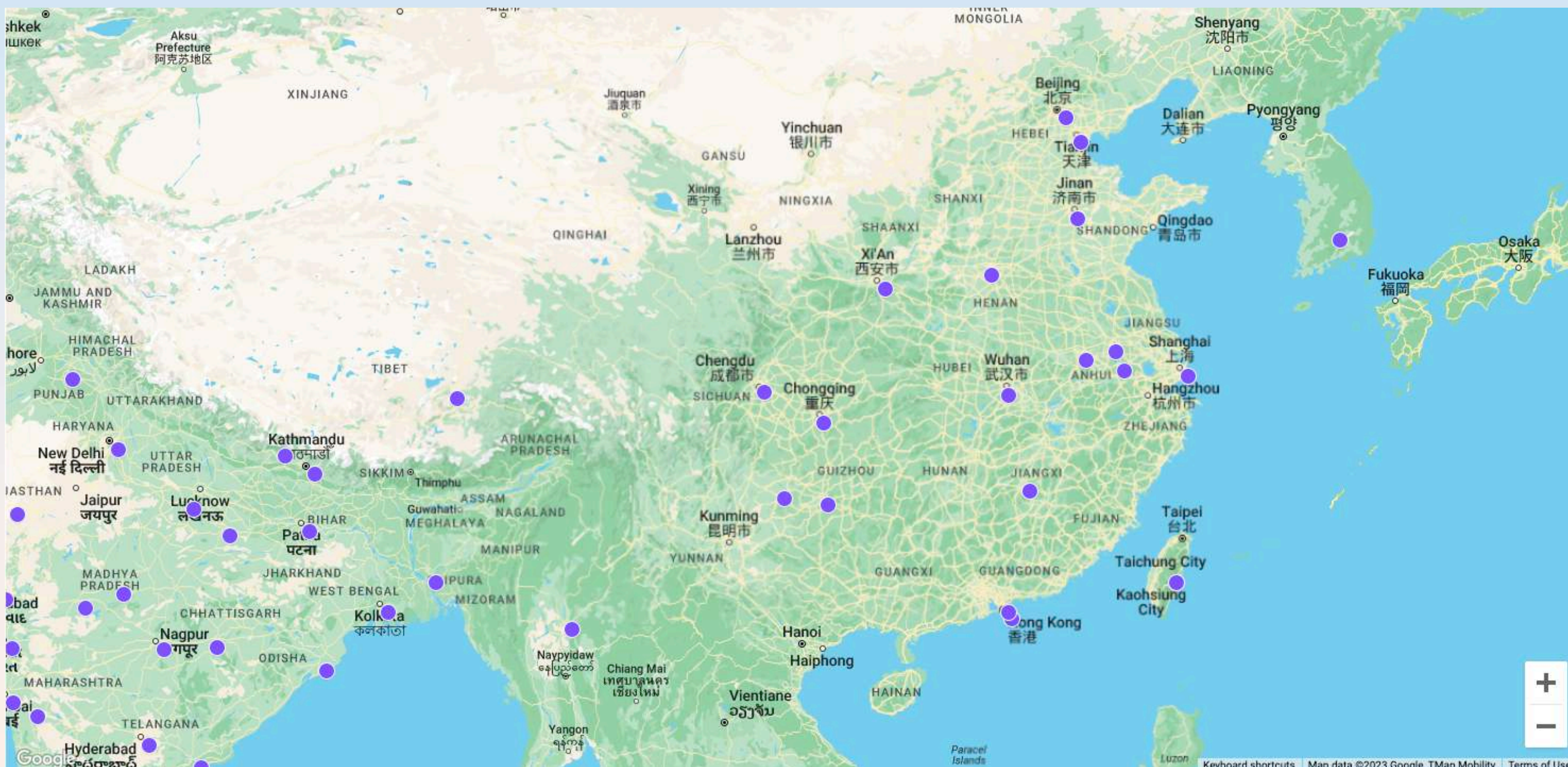


2021 年起举办 Kotlin 中文开发者大会



<p>Advent of Code Idiomatic Kotlin: Solving Advent of Code Puzzles 17:19</p> <p>今年的 Advent of Code 活动开始了，来用 Kotlin 挑战每日一道算法 1380 2022-12-3</p>	<p>Kotlin 中文开发者大会 01:33</p> <p>成为 Kotlin 的新生力量！【Kotlin 中文开发者大会预告之 KUG 伙伴】 1106 2022-11-25</p>	<p>Kotlin 中文开发者大会 01:47</p> <p>Kotlin 多平台项目实践抢先看！【Kotlin 中文开发者大会预告之乔】 1024 2022-11-24</p>	<p>Kotlin 中文开发者大会 01:01</p> <p>她在用一种很新的方式刷 LeetCode! 【Kotlin 中文开发者大会预告之阿】 2241 2022-11-23</p>	<p>Kotlin 中文开发者大会 01:15</p> <p>码住这场必看的 Compose 主题分享！【Kotlin 中文开发者大会预告之阿】 1418 2022-11-22</p>	<p>Kotlin 中文开发者大会 01:09</p> <p>Kotlin 中文开发者大会报名进行时，剧透走一波！【AB 篇】 814 2022-11-21</p>
<p>Kotlin 社区的大小事 Kotlin User Group 组织者 20:13</p> <p>【直播回放】那些运营 Kotlin 社区的大小事 2022 Kotlin 中文开发者大会 209 1-19</p>	<p>使用 Kotlin 编写 IntelliJ Plugin 张志豪 30:16</p> <p>【直播回放】使用 Kotlin 编写 IntelliJ Plugin 2022 Kotlin 中文开发者大会 1129 1-19</p>	<p>使用 Vaadin 搭配 Kotlin 快速开发 Web 应用 郭晋宜 (Maggie) 55:24</p> <p>【直播回放】使用 Vaadin 搭配 Kotlin 快速开发 Web 应用 2022 Kotlin 中文开发者大会 1056 1-19</p>	<p>从零开始欣赏 Coroutine 的精湛设计 Gary Lo 01:04:52</p> <p>【直播回放】从零开始欣赏 Coroutine 的精湛设计 2022 Kotlin 中文开发者大会 2437 1-19</p>	<p>与时俱进： 使用 Kotlin 尝鲜 Spring 6 贾彦伟 49:00</p> <p>【直播回放】与时俱进：使用 Kotlin 尝鲜 Spring 6 2022 Kotlin 中文开发者大会 1.1万 1-19</p>	<p>在你的城市 寻找 Kotlin 伙伴 Alina Dolgikh & 范圣佑 14:02</p> <p>【直播回放】在你的城市寻找 Kotlin 伙伴 2022 Kotlin 中文开发者大会 946 1-19</p>
<p>Kotlin 很简单， 一起来学习吧！ 楠楚伶 & Noelle & 林洁彬 & 范圣佑 19:14</p> <p>【直播回放】Kotlin 很简单，一起来学习吧！ 2022 Kotlin 中文开发者大会 765 1-19</p>	<p>享受用 Kotlin 刷 LeetCode 的乐趣 李盈莹 (Kate) 21:07</p> <p>【直播回放】享受用 Kotlin 刷 LeetCode 的乐趣 2022 Kotlin 中文开发者大会 1685 1-19</p>	<p>通过 Ktor 框架 同步完成 Server 及 Client Side 开发 赵家笙 (Recca) 34:13</p> <p>【直播回放】通过 Ktor 框架同步完成 Server 及 Client Side 开发 2022 Kotlin 中文开发者大会 1061 1-19</p>	<p>Kotlin Symbol Processor 应用与技巧 2BAB 31:46</p> <p>【直播回放】Kotlin Symbol Processor 应用与技巧 2022 Kotlin 中文开发者大会 1184 1-19</p>	<p>使用 DSL + KSP 打造跨平台的 Kotlin SQLite 框架 乔禹昂 35:22</p> <p>【直播回放】使用 DSL + KSP 打造跨平台的 Kotlin SQLite 框架 2022 Kotlin 中文开发者大会 1533 1-19</p>	<p>从 Jetpack Compose 到 Compose Multiplatform 王鹏 41:43</p> <p>【直播回放】从 Jetpack Compose 到 Compose Multiplatform 2022 Kotlin 中文开发者大会 1921 1-19</p>
<p>十分钟带您了解 Kotlin 的 2022 Pamela Hill & 范圣佑 14:58</p> <p>【直播回放】十分钟带您了解 Kotlin 的 2022 2022 Kotlin 中文开发者大会 4420 1-19</p>					

培育各城市的 Kotlin User Group



China Mainland

Beijing KUG

Chengdu KUG

Chongqing KUG

Guangzhou KUG

Guizhou KUG

Hefei KUG

JiangXi KUG

Jinan KUG

Nanjing KUG

Shanghai KUG

Shenzhen KUG

Suzhou KUG

Tianjin KUG

Wuhan KUG

Xi'An KUG

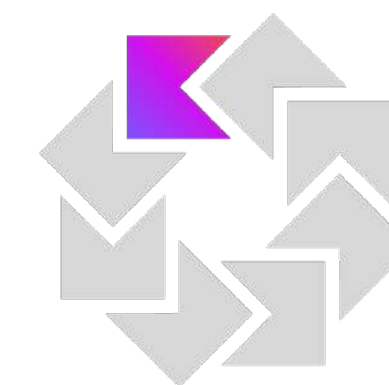
Zhengzhou KUG

Hong Kong

Hong Kong KUG

Taiwan

Taiwan KUG



目前有 18 个城市
有说中文的 Kotlin
社区！

跟 KUG 的小伙伴一时兴起

《Kotlin 炉边漫谈》正式开播！



炉边漫
谈

Kotlin
talk

共同主持



禹昂



Maggie

现阶段成果

3 位共同主持人

1.5 年

12 期节目

2023 第八届中国

Kotlin 爐邊漫談 Podcast
Kraftsman:Coding 職人塾
9 videos 90 views Updated 7 days ago

▶ Play all Shuffle

- 1 #Kotlin 爐邊漫談 Podcast #1 - Kotlin 爐邊漫談開播啦
Kraftsman:Coding 職人塾 • 355 views • 1 year ago
- 2 #Kotlin 爐邊漫談 Podcast #2 - 聽說 Kotlin 能算命？
Kraftsman:Coding 職人塾 • 141 views • 10 months ago
- 3 #Kotlin 爐邊漫談 Podcast #3 - 技術社群主辦人都在忙什麼？
Kraftsman:Coding 職人塾 • 146 views • 9 months ago
- 4 #Kotlin 爐邊漫談 Podcast #4 - Kotlin 的異國工作經驗 feat. 二分電台
Kraftsman:Coding 職人塾 • 98 views • 7 months ago
- 5 #Kotlin 爐邊漫談 Podcast #5 - 手機開發編年史
Kraftsman:Coding 職人塾 • 128 views • 5 months ago
- 6 #Kotlin 爐邊漫談 Podcast #6 - 年終特輯
Kraftsman:Coding 職人塾 • 114 views • 4 months ago
- 7 #Kotlin 爐邊漫談 Podcast #7 - 聊聊技術領域的多元 · 公平 · 共融以及演講經驗談
Kraftsman:Coding 職人塾 • 148 views • 2 months ago
- 8 #Kotlin 爐邊漫談 Podcast #8 - 來自阿里巴巴及美团的 Kotlin Multiplatform Mobile 應用案例
Kraftsman:Coding 職人塾 • 112 views • 1 month ago
- 9 #Kotlin 爐邊漫談 Podcast #9 - 隨性地在新加坡路邊漫談
Kraftsman:Coding 職人塾 • 56 views • 7 days ago

尽力满足各种聆听习惯

—

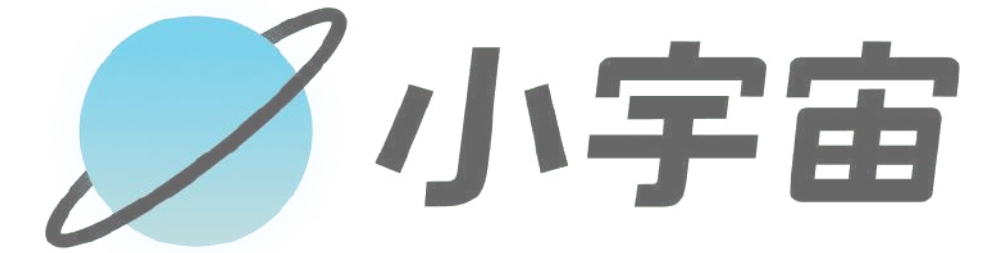
2 种语言

10 个平台



喜马拉雅





该怎么知道各平台的聆听数据呢？



喜马拉雅



部份平台把数据写在网页上

喜马拉雅 发现 频道 电台 专辑/声音/主播

第10期：在 KotlinConf'23 现场录播客

2023-07-20 16:34:02 37:19 161

所属专辑：Kotlin 炉边漫谈

bilibili 首页 番剧 直播 游戏中心 会员购 漫画 赛事 王者荣耀活动

Kotlin 炉边漫谈 Podcast #10 在 KotlinConf'23 现场录播客

283 2023-07-20 17:00:46 未经作者授权，禁止转载

荔枝 发现 主播入口 家族合作 Investor Relations

布道师圣佑 #10 在 KotlinConf'23 现场录播客

2023-07-20 37:19 35

YouTube

#Kotlin 爐邊漫談 Podcast #10 - 在 KotlinConf'23 現場錄 Podcast

Kraftsman:Coding 職人塾 1.05K subscribers

54 views 4 days ago

xiaoyuzhoufm.com

通勤路上 听播客，上小宇宙!

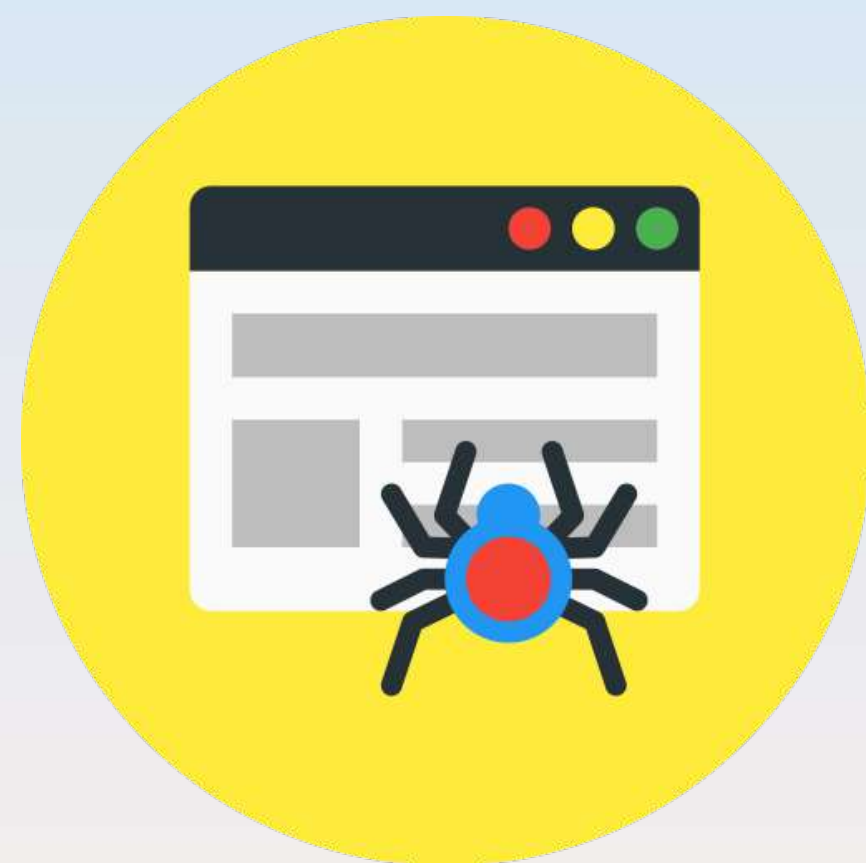
第10期：在 KotlinConf'23 现场录播客

Kotlin 炉边漫谈

37分钟·5天 24

《Kotlin 炉边漫谈》是一档谈论 Kotlin 相关信息的中文 Podcast，由上海 Kotlin User Group 主办人禹昂、Kotlin 开发者 Maggie 和 JetBrains 技术布道师圣佑共同主持，除了介绍两岸三地 Kotlin User Group 技术社群活动资讯外，还会邀请各地 Kotlin 开发者一起来聊聊 Kotlin 应用场景。

本期嘉宾：二分电台主理人 - 2BAB



说到网络爬虫，大家想到的编程语言是？



在京东搜索「爬虫」关键字

爬虫 ×

全部 店铺 口碑

人民邮电出版社京东自营官方... 自营 113万人关注 8年老店 进店

Python 3 网络爬虫开发实战 ¥69.00

Python 网络爬虫权威指南 ¥68.70

Python 3 反爬虫原理与绕过实战 ¥77.40

明日科技京东自营官方旗舰店 自营 2.9万人关注 5年老店 进店

Python 网络爬虫 从入门到实践 ¥47.70

Python 项目开发 实战入门 ¥47.70

Python 从入门到项目实践 ¥48.60

机械工业出版社京东自营官方... 自营 40.3万人关注 8年老店 进店

爬虫逆向进阶实战 ¥100.40

Python 网络爬虫 入门到实战 ¥68.60

Python 爬虫、数据分析与可视化 从入门到精通 ¥65.10

搜书文化图书专营店 26人关注 进店

爬虫 ×

全部 店铺 口碑

综合推荐 销量 价格 筛选

京东物流 配送全球 形态 品牌 总净含量

自营 实战Python网络爬虫

¥87.40 定价¥99.00 8.83折

7天价保

2万+条评价 99%好评

清华大学出版社京东自营官方旗舰店 进店

仅剩1件

自营 Python网络爬虫权威指南 第2版

¥68.70 定价¥79.00 8.70折

7天价保

5000+条评价 98%好评

人民邮电出版社京东自营官方旗舰店 进店

仅剩2件

自营 python网络爬虫与数据可视化应用实战

¥101.70 定价¥109.00 9.34折

7天价保 每满100减50

5万+条评价 98%好评

中国水利水电出版社京东自营官方... 进店

仅剩1件

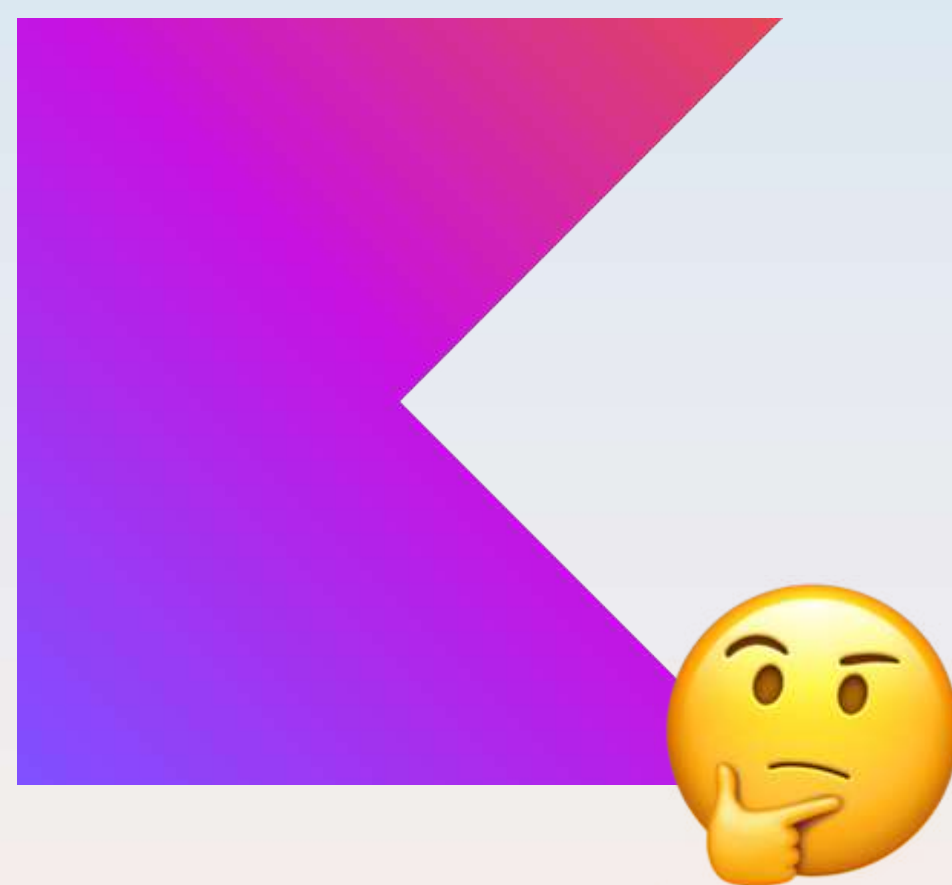
自营 Python网络爬虫从入门到精通

¥88.10 定价¥99.80 8.83折

7天价保

5000+条评价 97%好评

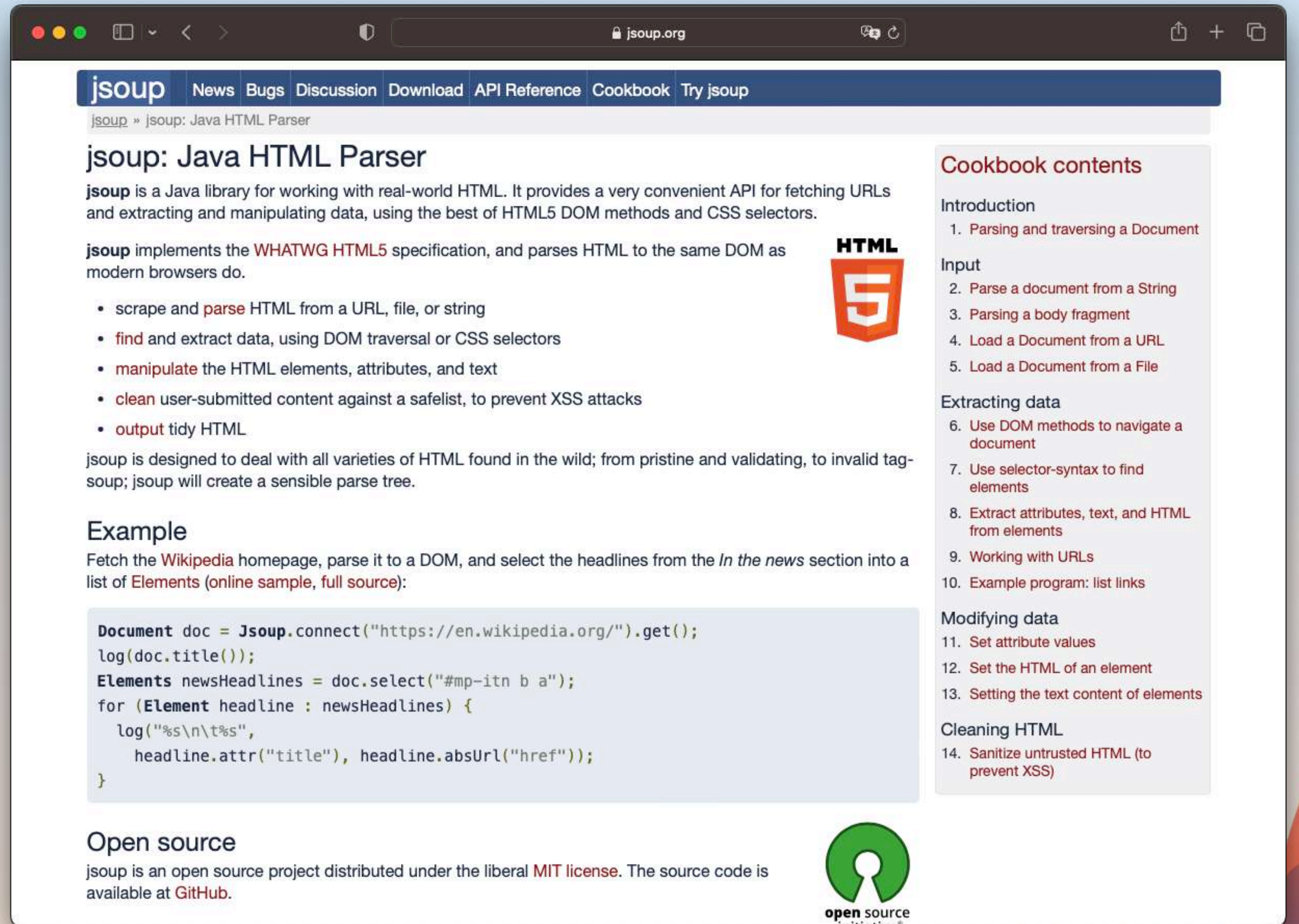
清华大学出版社京东自营官方旗舰店 进店



Python 做得到的事，Kotlin 也行么？

当然行！而且也是 soup！

<https://jsoup.org/>

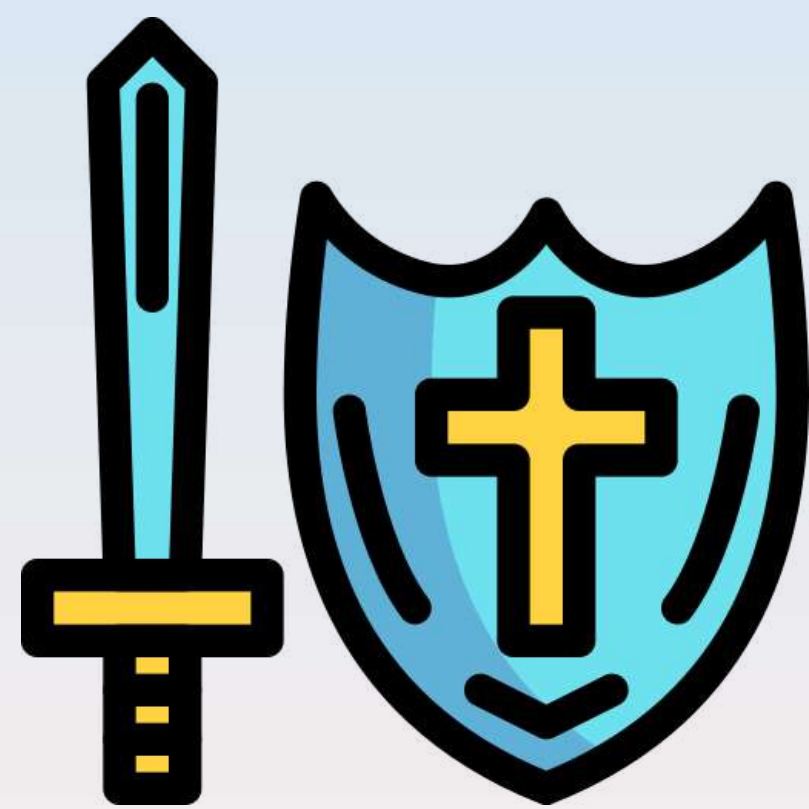


The screenshot shows the jsoup.org website. At the top, there is a navigation bar with links for News, Bugs, Discussion, Download, API Reference, Cookbook, and Try jsoup. The main heading is "jsoup: Java HTML Parser". Below this, a paragraph describes jsoup as a Java library for working with real-world HTML. A list of features includes scraping and parsing HTML, finding and extracting data, manipulating HTML elements, cleaning user-submitted content, and outputting tidy HTML. An example code block shows how to fetch the Wikipedia homepage, parse it to a DOM, and select headlines. A sidebar on the right contains a "Cookbook contents" section with a table of contents listing 14 items, including parsing, extracting data, modifying data, and cleaning HTML. The Open Source Initiative logo is visible at the bottom right of the page.

网页信息抓取

—

- 使用 HTTP Client 下载目标网页
- 使用 DOM Parser 套件抓取目标位置的字符串



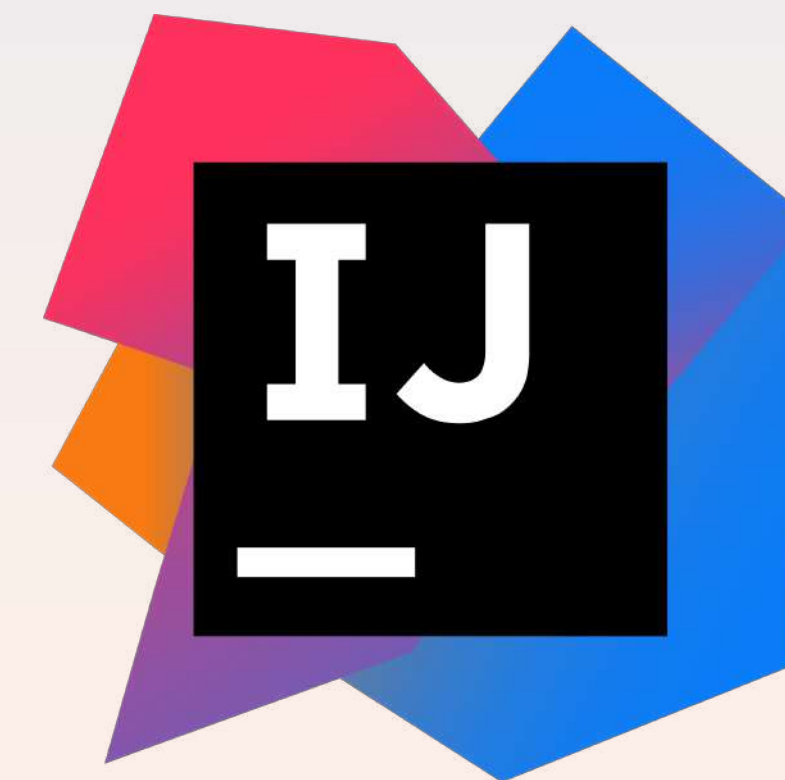
实战 Kotlin 爬虫抓取播客聆听数据



开发环境

—

- Java LTS 17+
- Gradle (or Maven)
- IntelliJ IDEA



添加 jsoup 库

—

```
dependencies {  
    // ...  
    implementation("org.jsoup:jsoup:$version")  
    // ...  
}
```


jsoup 基本语法

—

```
// 抓取目标网页
```

```
val url = "..."
```

```
val doc = Jsoup.connect(url).get()
```

```
// 通过 CSS Selector 定位至目标位置
```

```
val elements = doc.select("div.info div.category .count")
```

```
// 再从目标位置开始抓取其中的字符串
```

```
elements.text()
```

```
elements.html()
```

```
elements.outerHtml()
```

```
elements.attr("...")
```


从喜马拉雅抓数据

第7期：聊聊科技行业的性别多元化

2023-03-08 09:47:18 60:28 546

所属专辑：Kotlin 炉边漫谈



喜欢

下载

```
<!-- 页面内容 -->
```

```
<div class="info kn_">
```

```
  <div class="category kn_">
```

```
    <span class="count kn_">
```

```
      // ...
```

```
    </span>
```

```
    <span class="count kn_">
```

```
      <i class="xuicon xuicon-erji1 kn_"></i>
```

```
      546
```

```
    </span>
```

```
  </div>
```

```
</div>
```

```
// 抓取网页并定位
```

```
val url = "https://www.ximalaya.com/sound/$soundId"
```

```
val doc = Jsoup.connect(url).get()
```

```
val element = doc.select("div.info div.category .count")[1]
```

```
// 抓取字符串并转成整数
```

```
element.text().trim().toInt()
```


从蜻蜓 FM 抓数据



```
<!-- 页面内容 -->
<div class="info">
  <div>
    <span class="propTxt">播放: </span>
    <span>11次</span>
  </div>
</div>
```

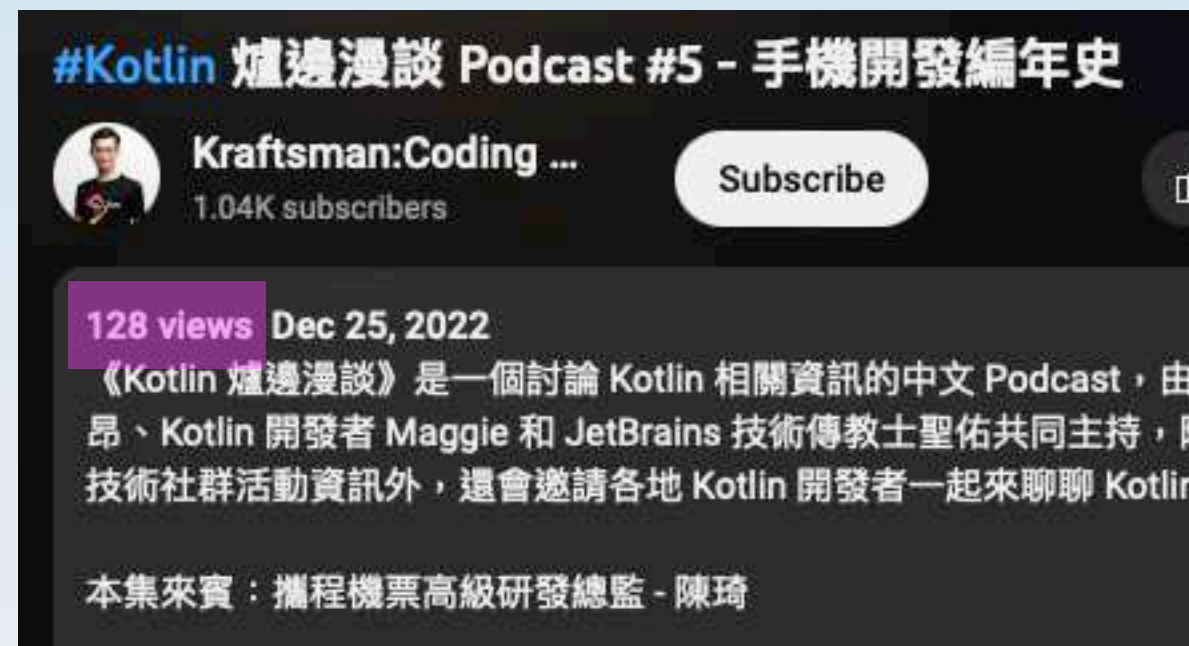
```
// 抓取网页并定位
val url = "https://www.qingting.fm/.../programs/$programId"
val doc = Jsoup.connect(url).get()
val element = doc.select("div.info div span")
    .first { node -> node.html().contains("次") }
```

```
// 抓取字符串并转成整数
element.text().replace("次", "").toInt()
```




等一下！抓回来的画面一片白阿？

从 YouTube 抓数据



```
<!-- 页面内容 -->
```

```
<!DOCTYPE html>
```

```
<html>
```

```
<head>
```

```
  <div>
```

```
    <meta itemprop="interactionCount" content="128">
```

```
  </div>
```

```
</head>
```

```
</html>
```

```
// 抓取网页并定位
```

```
val url = "https://www.youtube.com/watch?v=$videoId"
```

```
val doc = Jsoup.connect(url).get()
```

```
val element = doc
```

```
    .select("meta[itemprop=interactionCount]").first()
```

```
// 抓取字符串并转成整数
```

```
element?.attr("content")?.toIntOrNull() ?: 0
```

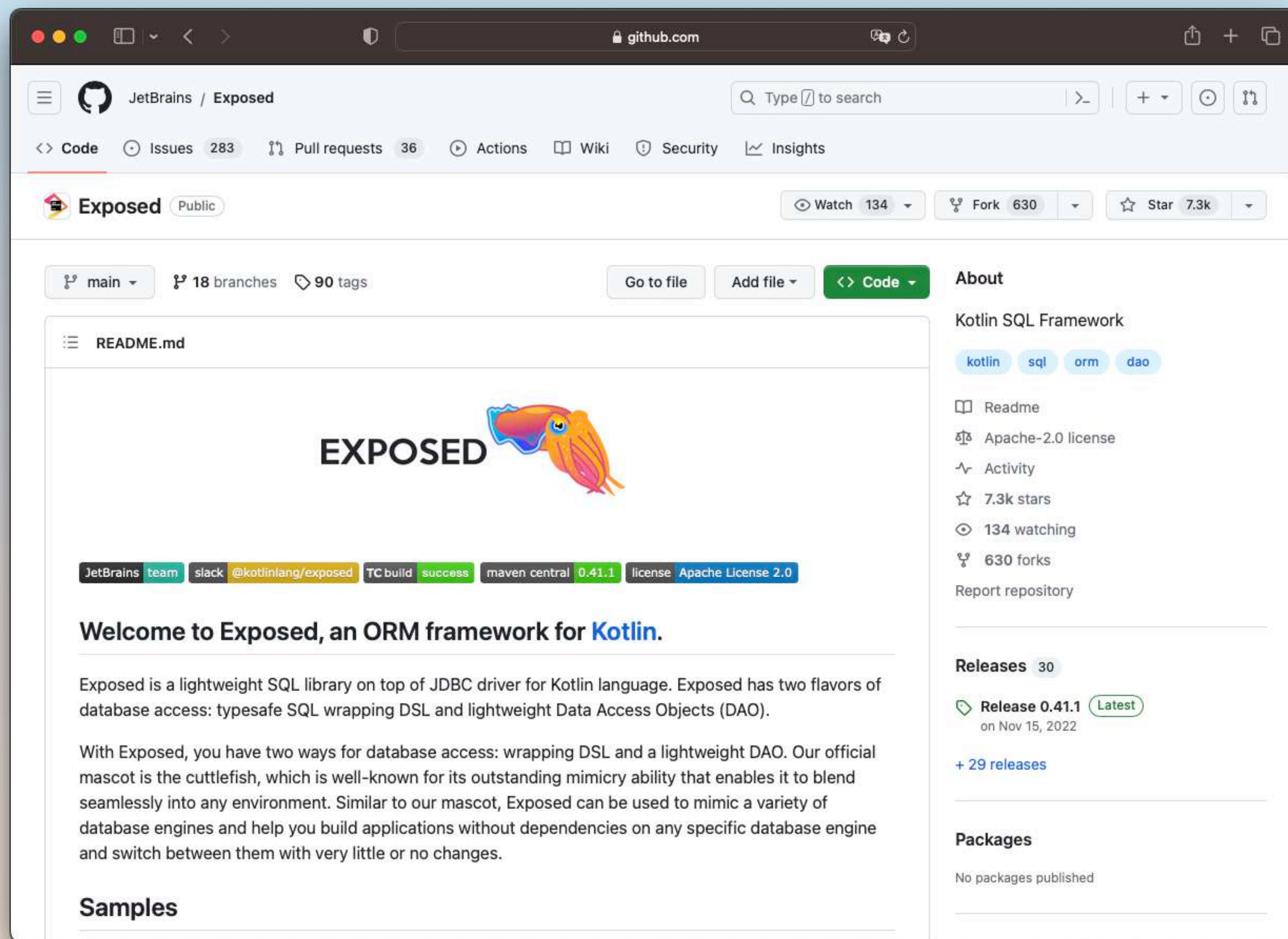



需要将抓取的数据存储至数据库

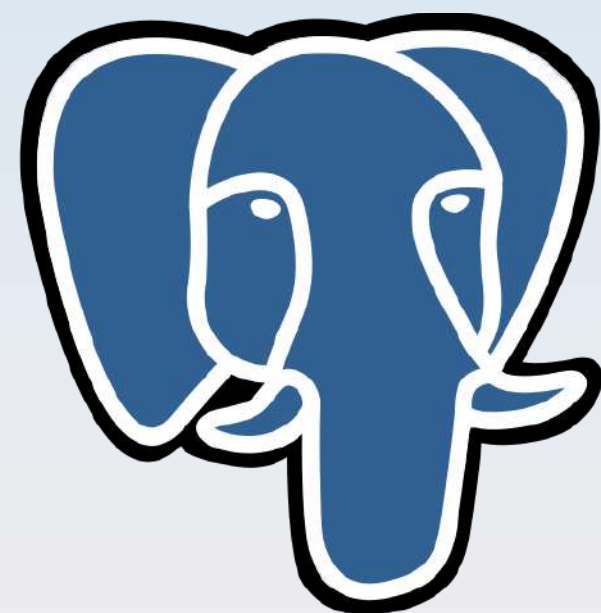


集成 Exposed ORM 框架

<https://github.com/JetBrains/Exposed>



支持主流数据库



PostgreSQL



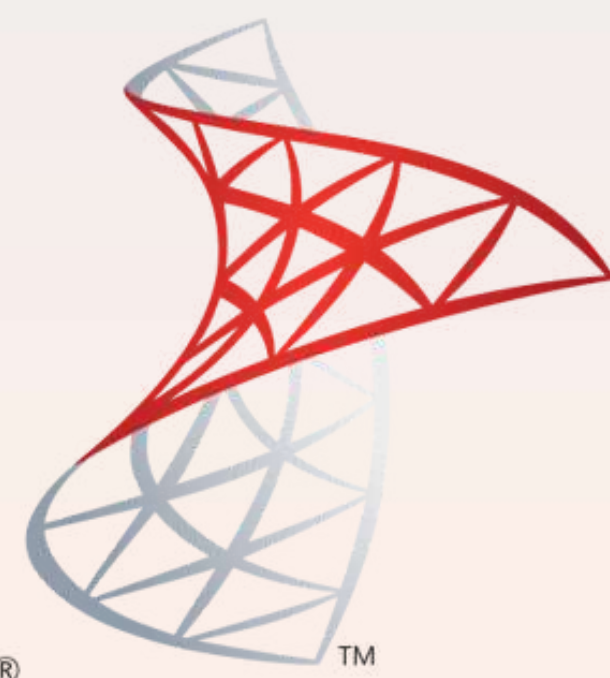
SQLite



H2 database



Oracle



SQL Server



添加 Exposed 及 Connector 库

```
dependencies {  
  
    // ...  
  
    implementation("org.jetbrains.exposed:exposed-core:$version")  
    implementation("org.jetbrains.exposed:exposed-dao:$version")  
    implementation("org.jetbrains.exposed:exposed-java-time:$version")  
    implementation("org.jetbrains.exposed:exposed-jdbc:$version")  
    implementation("mysql:mysql-connector-java:$version")  
  
    // ...  
  
}
```

两种方式操作数据

—



typesafe SQL wrapping DSL



lightweight data access objects

设定数据结构

```
// 声明 Table
object Statistics : IntIdTable(name = "statistics") {
    val episode = integer("episode")
    val platform = integer("platform")
    val listenNumber = integer("listen_number")
    val scrapedAt = datetime("scraped_at")
}

// 声明 Entity
class Statistic(id: EntityID<Int>) : IntEntity(id) {
    companion object : IntEntityClass<Statistic>(Statistics)

    var episode by Statistics.episode
    var platform by Statistics.platform
    var listenNumber by Statistics.listenNumber
    var scrapedAt by Statistics.scrapedAt
}
```

将抓取的数据写入

```
// 数据库连线
Database.connect(
    url = "jdbc:mysql://.../...",
    driver = "com.mysql.cj.jdbc.Driver",
    user = "...",
    password = "...",
)

// 写入数据
transaction {
    Statistic.new {
        episode = ...
        platform = ...
        listenNumber = scraper.scrape(...)
        scrapedAt = LocalDateTime.now()
    }
}
```




需要处理不同数据格式？



处理不同数据格式

—

- Text
- HTML
- JSON
- XML
- CSV



页面上的数据需等 JavaScript 载入后才能抓？



添加 Selenium 及 WebDriverManager 库

```
dependencies {  
    // ...  
    implementation("io.github.bonigarcia:webdrivermanager:$version")  
    implementation("org.seleniumhq.selenium:selenium-java:$version")  
    // ...  
}
```


通过浏览器 抓取网页数据

```
// 取 WebDriverManager 取得 Selenium Driver  
WebDriverManager.firefoxdriver().setup()  
  
// 设置浏览器参数  
val options = FirefoxOptions()  
                .addArguments("--headless")  
val driver = FirefoxDriver(options)  
  
// 以浏览器读取目标网页  
driver.get("...")  
  
// 抓取页面上指定位置的数据  
val element = driver.findElements(  
    By.cssSelector("...")  
).first()  
element.text.toIntOrNull() ?: 0  
  
// 关闭浏览器  
driver.quit()
```

其他支持 Java 的 Headless 浏览器

- HTMLUnit
- Playwright
- Puppeteer (with Jvppeteer)
- 其他第三方服务

框架比较 📌 [Playwright vs Selenium vs Cypress: A Detailed Comparison](#)



需要定期抓取数据？



添加 JobRunr 及 相依库

```
dependencies {  
  
    // ...  
  
    implementation("org.jobrunr:jobrunr:$version")  
    implementation("com.fasterxml:jackson.core:jackson-databind:$ver")  
  
    // ...  
  
}
```


定时抓取数据

—

```
// 设置 JobRunr
JobRunr.configure()
    .useStorageProvider(...)
    .useBackgroundJobServer()
    .useDashboard(...)
    .initialize()

// 设置爬虫
val scraper = Scraper()

// 设置定时抓取数据
BackgroundJob.scheduleRecurrently(Cron.hourly()) {
    scraper.scrape()
}
```



除了 jsoup 外，还有其他选择吗？



添加 skrape{it} 库

—

```
dependencies {  
    // ...  
    implementation("it.skrape:skrapeit:$version")  
    // ...  
}
```

抓取页面数据

```
fun main() {
    val links: List<String> = skrape(HttpFetcher) {
        request {
            url = "https://kotlinlang.org/docs/reference/"
        }
        response {
            htmlDocument {
                a {
                    findAll {
                        eachHref
                    }
                }
            }
        }
    }
    println(links)
}
```




本日分享回顾

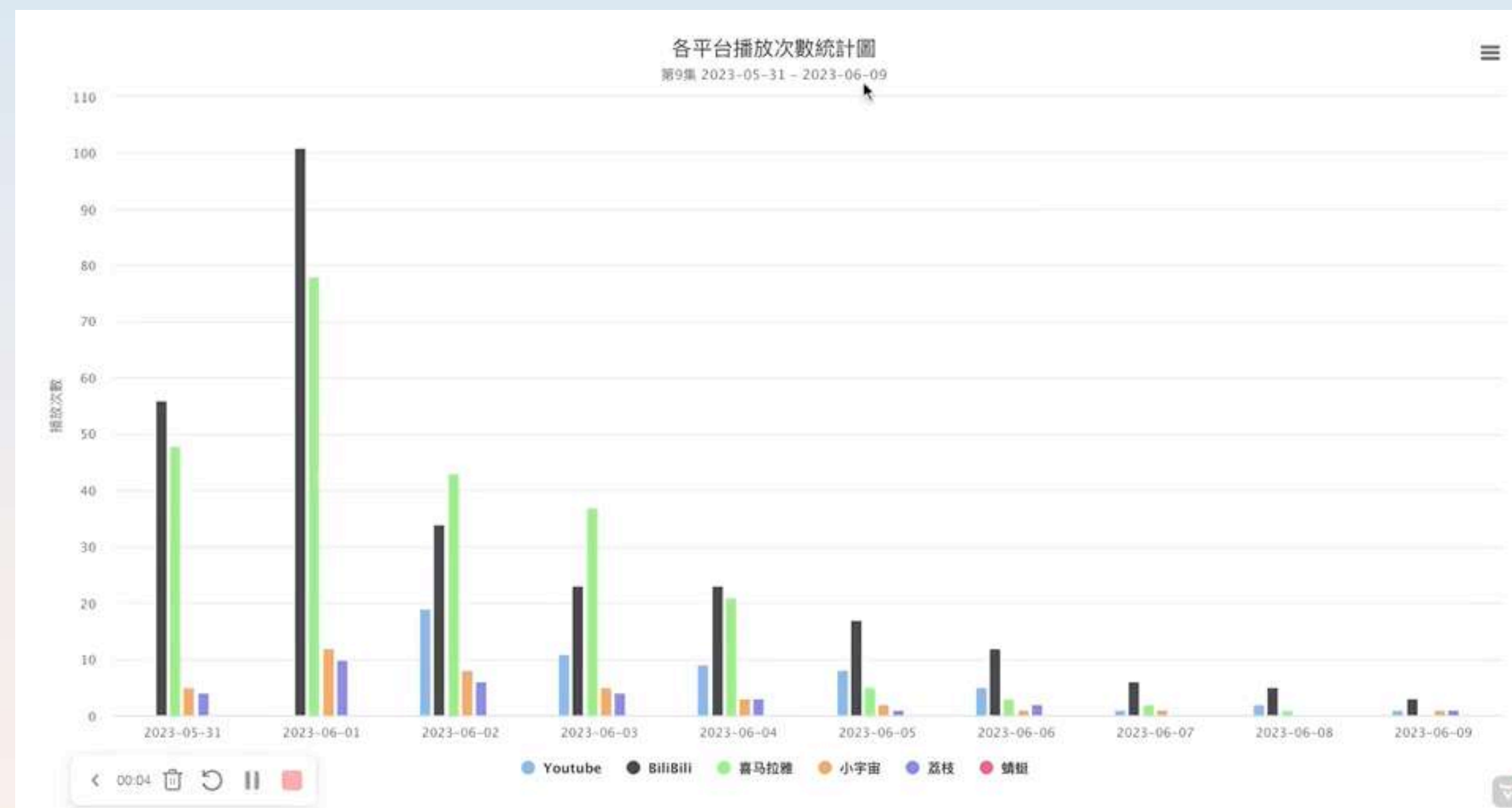
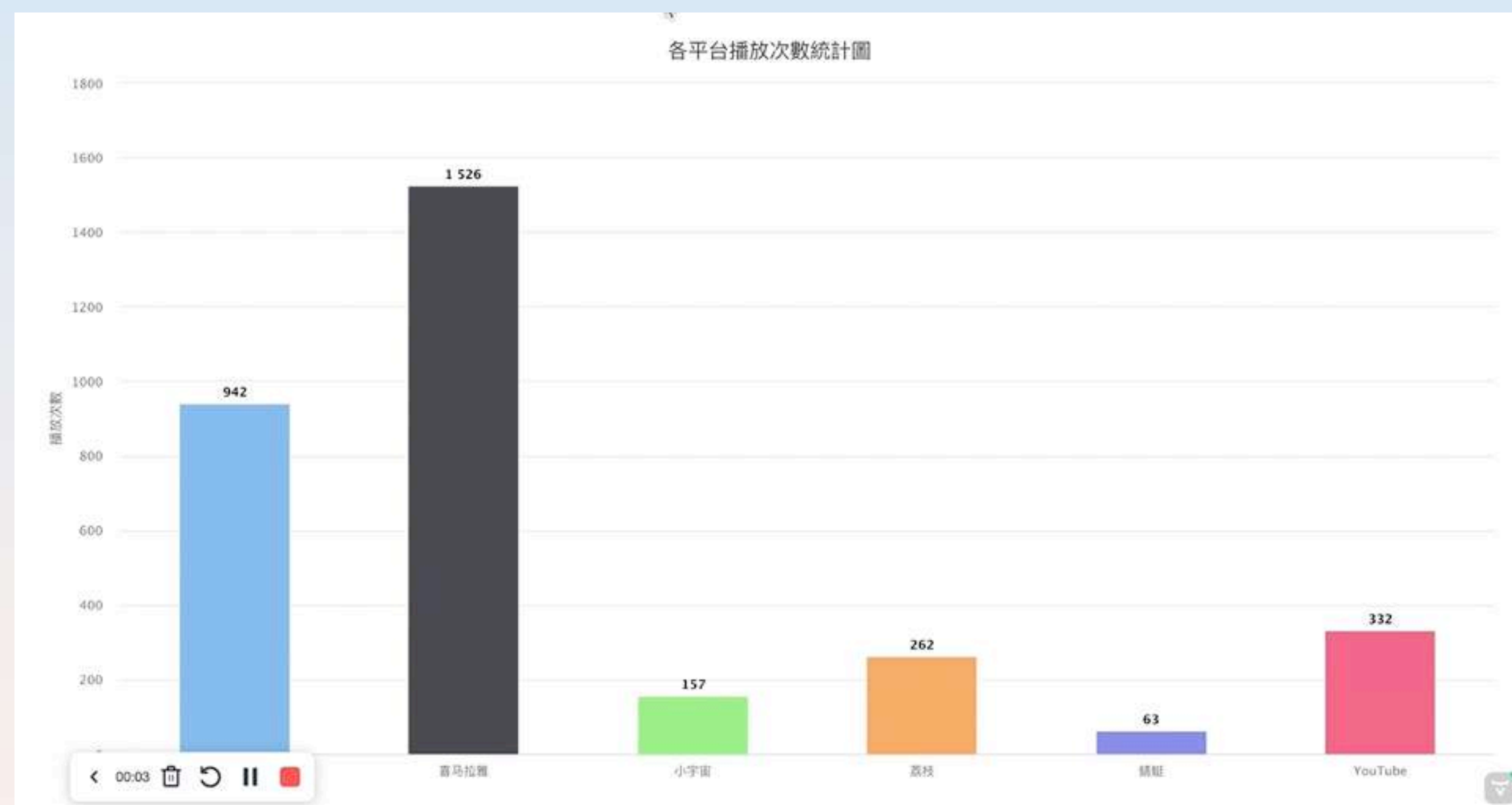


Kotlin 也能写爬虫！

—

- 无需混用多种语言
- 丰富的库、框架及工具
- 成熟的开发生态系

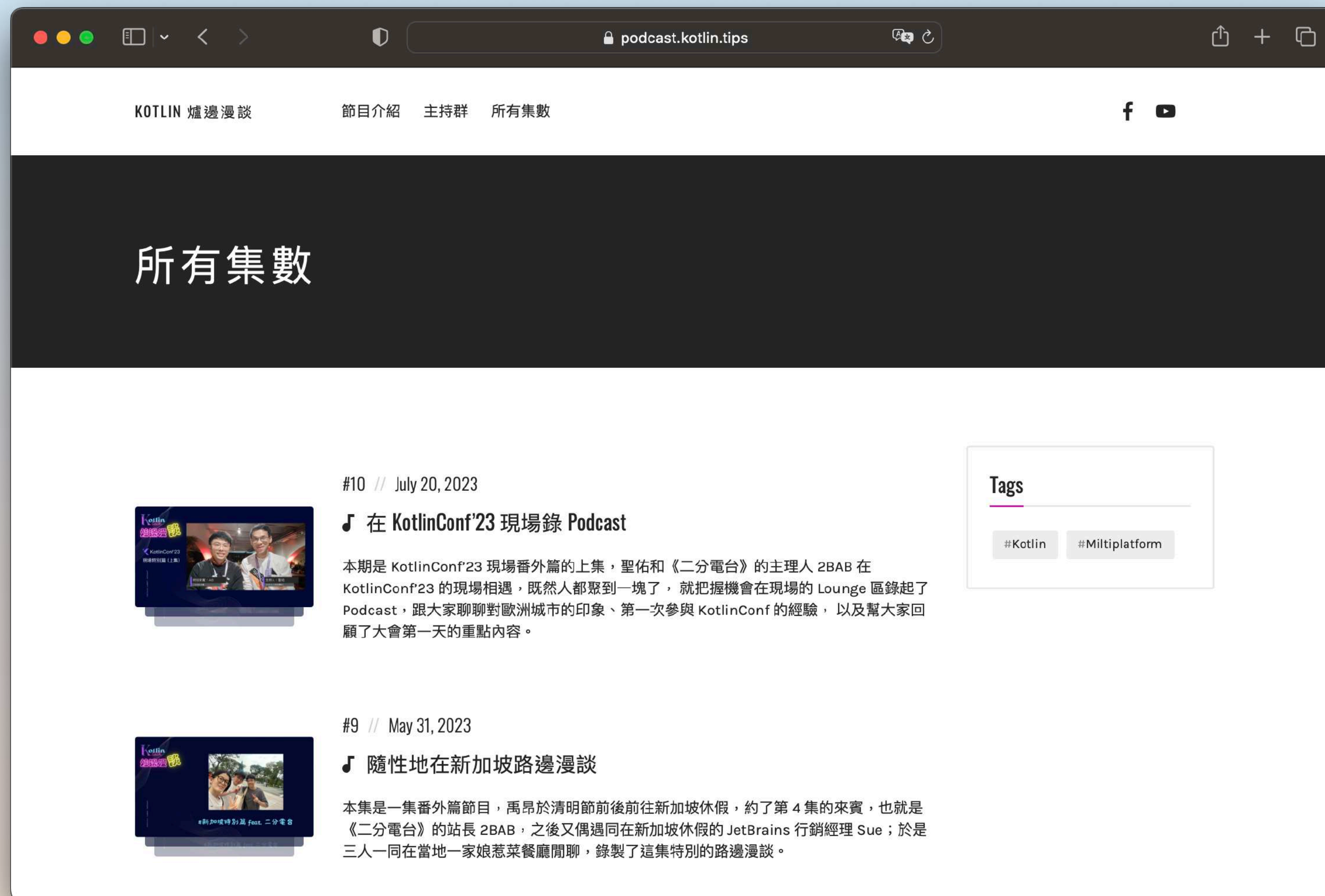
更进一步 (WIP)



欢迎收听并订阅 Kotlin 炉边漫谈播客



<https://podcast.kotlin.tips/>





想了解更多？

开源市集 (Open Source Bazaar)



THANK YOU

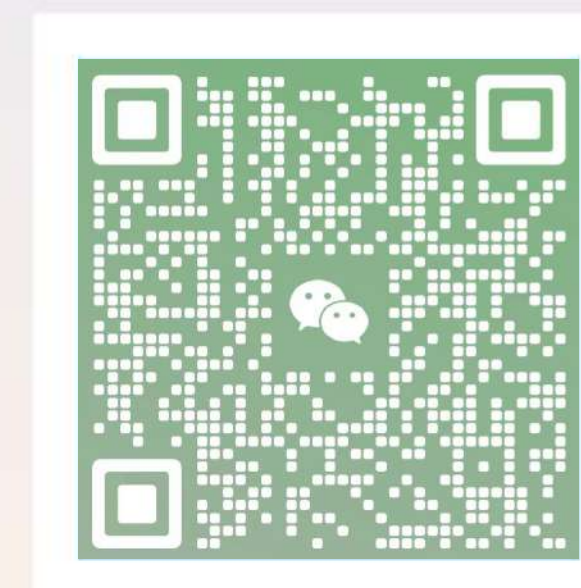
QUESTIONS?



欢迎扫码打卡
积分可兑换对应礼品哟!



扫码关注开源社公众号



扫码添加讲师联系方式

微信公众号：开源社KAIYUANSHE

视频号：开源社KAIYUANSHE

新浪微博：开源社

B站：开源社KAIYUANSHE

简书：开源社

头条：开源社

Facebook: KaiyuansheChina

Twitter: 开源社KAIYUANSHE