Headless Browser

演讲者：yaoj


先说是什么


无UI的浏览器
开局一个CONSOLE，控制全靠命令

提供DOM环境模拟，允许执行DOM API操作

提供完整JS运行环境

↓↓↓↓

前端测试容器、预渲染

无头浏览器都有啥

全平台

PhantomJS - Command line/all platforms.

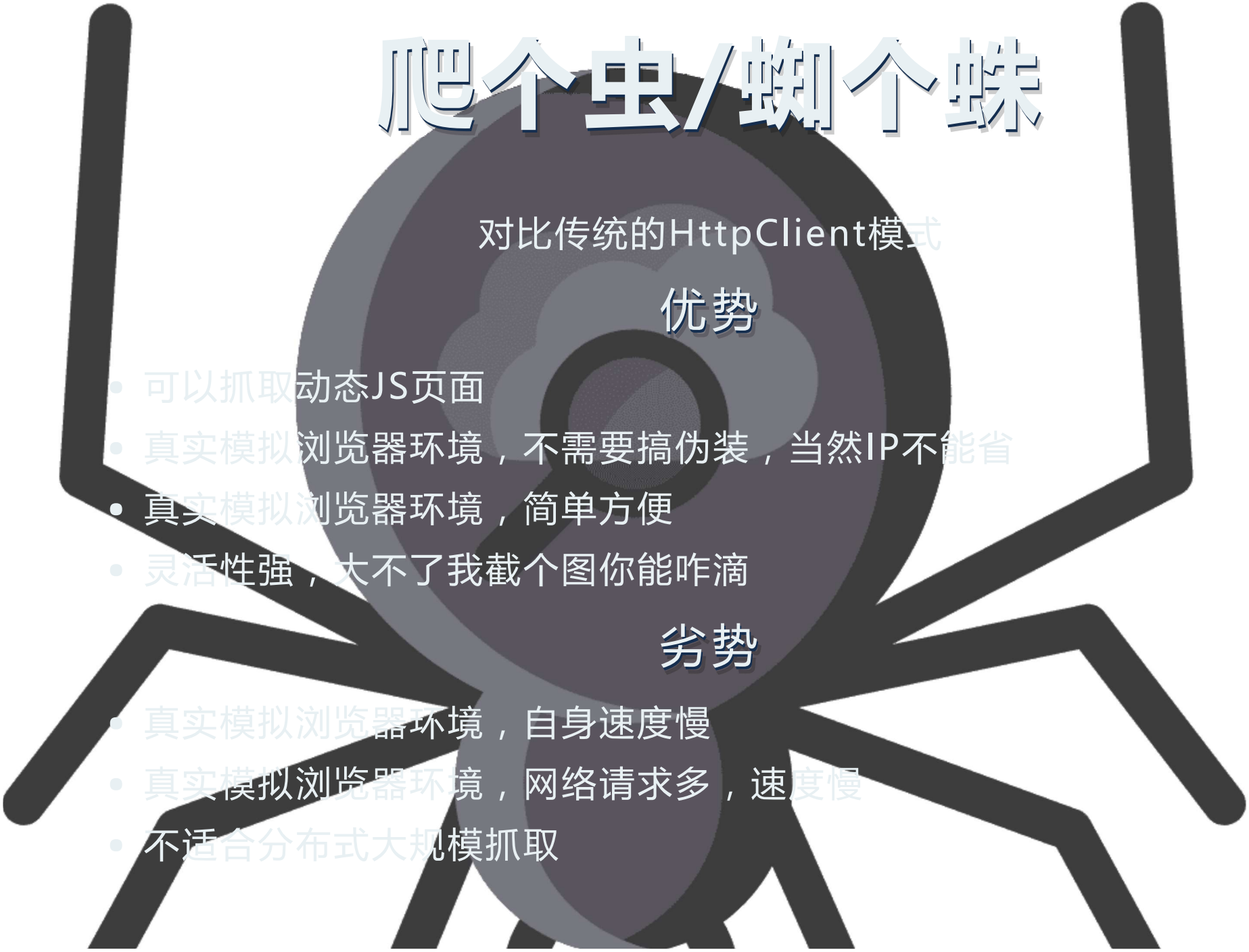
Python

- Spynner - Python only.
- Ghost - Python only.
- Twill - Python/command line.
- Ghost - Python only.
- Twill - Python/command line.

Others

- HtmlUnit - Java.
- Awesomium - C++/.Net/all platforms.
- SimpleBrowser - .Net 4/C#.
- ZombieJS - Node.js.
- EnvJS - JavaScript via Java/Rhino.
- Watir-webdriver with headless gem - Ruby via WebDriver.

爬个虫/蜘个蛛



对比传统的HttpClient模式

优势

- 可以抓取动态JS页面
- 真实模拟浏览器环境，不需要搞伪装，当然IP不能省
- 真实模拟浏览器环境，简单方便
- 灵活性强，大不了我截个图你能咋滴

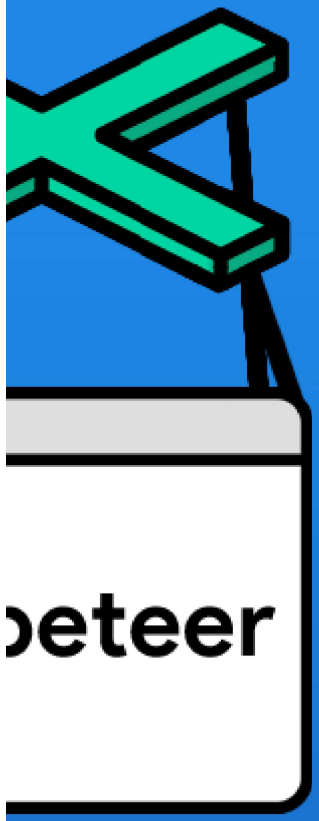
劣势

- 真实模拟浏览器环境，自身速度慢
- 真实模拟浏览器环境，网络请求多，速度慢
- 不适合分布式大规模抓取

Google Chrome/puppeteer

官方杀死同人系列

https://github.com/GoogleChrome/puppeteer



- 自带截图生成PDF API
- 为SPA应用做预渲染
- 爬虫
- 自动表单填写，UI测试，模拟键盘输入
- 及时更新的自动化的测试环境，随时在最新的Chrome环境中运行，享受最新的浏览器特性与JS功能
- 输出时间轴标记用于浏览器性能测试
- 基于DevTools protocol, 理论上devTools有的东西都能玩

截图-代码范例

```
// 获取浏览器
const browser = await puppeteer.launch();
// 打开TAB
const page = await browser.newPage();
// 输入URL
await page.goto('https://www.t66y.com');
// 等待页面完全打开
console.log(await page.content());
// 截图
await page.screenshot({path: 'screenshot.png'});
// 关闭
await browser.close();
```

运行JS-代码范例

```
// 获取浏览器
const browser = await puppeteer.launch();
// 打开TAB
const page = await browser.newPage();
 // 跳转URL
await page.goto('https://www.baidu.com')
// 等待 #kw DOM节点可用
await page.waitFor('#kw')
// 往页面内注入JS
await page.evaluate((_query) => {
    // 输入搜索关键字
    document.querySelector('#kw').value = _query
    // 点击搜索按钮
    document.querySelector('#su').click()
}, query)
// 等待网络空闲，说明搜索完成
await page.waitForNavigation({
    waitUntil: 'networkidle',
})
// 注入JS抓取本页搜索结果内容
const searchResult = await page.evaluate(getData)
// 注入JS抓取分页信息
const pageData = await page.evaluate(getPages)
return { searchResult, pageData }
```

一些技巧1

```
// 获取浏览器
const browser = await puppeteer.launch({
    headless: false //显示浏览器
    executablePath: './static/chrome-win32/chrome.exe' // 本地chrome浏览器
});
// 打开TAB
const page = await browser.newPage();
// 设置拦截过滤器
await page.setRequestInterceptionEnabled(true)
page.on('request', (interceptedRequest) => {
    if (interceptedRequest.resourceType === 'stylesheet'
        || /token/.test(interceptedRequest.url)) {
        interceptedRequest.abort()
    } else {
        interceptedRequest.continue()
    }
})
```

一些技巧2

```
// 获取浏览器
// 设置拦截过滤器
await page.setRequestInterceptionEnabled(true)
page.on('request', (interceptedRequest) => {
    if (interceptedRequest.resourceType === 'stylesheet'
        || /token/.test(interceptedRequest.url)) {
        interceptedRequest.abort()
    } else {
        interceptedRequest.continue()
    }
})
```

一些技巧3

```
// 模拟键盘
await page.keyboard.type('Hello World!')
await page.keyboard.press('ArrowLeft')
// 模拟鼠标
await mouse.click(300, 720, {
    button: 'left'
})
```

坑坑坑

- 不翻墙，下不了自带的C

```
npm set PUPPETEER_
```

- 继承Chrome的版本帝，

- 瞎生成缓存文件，

实战

抢车票，抢小米，抢卷，抢TM的

刷票刷分刷评价，刷TM的

下资源下种子合集，下TM的

抓内容，监听更改，偷TM的