App 爬虫基础

Zhenghui Wang *

International School, Beijing University of Posts & Telecommunications

2020 年 11 月 2 日

1 动因与概述

很多时候我们无法通过 web 爬虫解决问题，比如说我想知道北邮体育馆预约明天游泳的名额是否满了，我就需要打开微信，进入北邮企业号，找到北邮体育馆程序，从中获知。这是 web 爬虫无法获得的数据，而对这一部分数据的需求让 app 爬虫技术产生了。

在 web 爬虫中，只需要 requests 和 re 两个 python 扩展库就可以完成绝大多数任务。但在 app 中获取数据是很有技术含量的，仅涉及到的各种框架就要花费很多时间熟悉。

由于我的需求仅仅是在得物 app 中爬取鞋类数据，所以对 app 爬虫技术研究得并不深入，我仅针对我的探索历程作简单介绍。

软件需求：Charles、mitmproxy、appium

硬件需求：IOS、安卓 6.0 以下版本

2 实验过程

2.1 Charles 的使用

Charles 是一个网络抓包工具，可以用于 App 的抓包分析，得到 App 运行过程中发生的所有网络请求和相应内容，这与在 web 端浏览器的开发者工具 Netwrok 部分看到的结果一致。

与 Charles 类似的软件还有 Fiddler（这是我最开始用的软件，界面略原始，但功能基本一致，我一开始用 Fiddler 抓安卓系统的包总是被拒绝，以为是软件的问题，后发现用 Charles 也一样，问题并不是出在软件上）

在使用 Charles 抓包前要确保已经正确安装了 Charles 并开启了代理服务，手机和 Charles 处于同一局域网下，Charles 代理和 CharlesCA 证书设置好，并开启 SSL 监听，具体配置请 Google。

配置完成后，Charles 应该已经可以监听移动端的网络数据传输了。图1

但是，如果你使用的是安卓系统，且安卓版本高于 6.0，那么很遗憾尽管你能监听到 App 的数据传输，但其中很多重要内容你看不见（高版本的安卓系统不再允许对用户手动安装的证书提供信任）。图2

类似上图，我想访问得物 App 中的图片，但被拒绝。网上提供了很多在不改变安卓版本前提下解决这一问题的思路，过于复杂，本宝宝实在接受不了遂放弃，有兴趣请自行 Google。

有一种解决方案是下载低于 6.0 版本的安卓模拟器，应该可以解决问题（我当时不清楚到底哪个版本的安卓系统开始不能信任证书了，于是下了一个 6.0 的结果还是不行……）

*Corresponding author: wangzhenghui@bupt.edu.cn

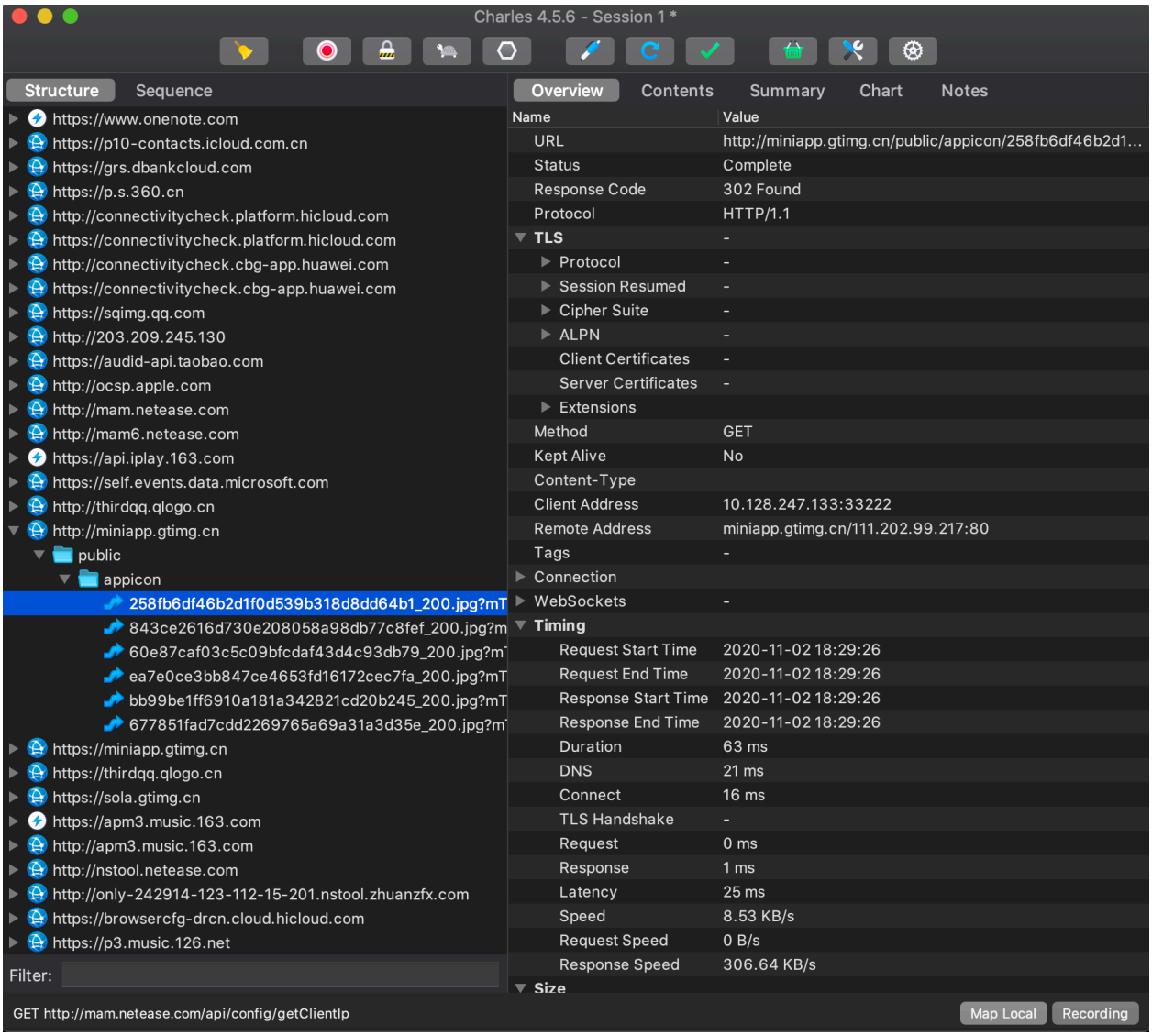

图 1: 连接成功的 Charles 界面

对安卓彻底失望后我转变思路开始用 ios 端测试，期间出现了一些小问题（移动端和 charles 接到校园网后死活连不上，最后手机建个热点立马奏效），因为 ios 端是允许信任用户安装的证书的（这一部分请自行 Google），所以配置很快就完成了。

但是，如果需要利用 appium 进行自动化测试，ios 系统下载的软件又会出问题，这一部分后面再说。

很显然的是，利用 Charles 软件我们已经可以看到 App 内部的请求和响应了。按照正常的思路，我们在 python 中利用 requests 模拟移动端发出请求，然后利用 re 捕获数据，一切就顺理成章地结束了。但显然没有这么简单，我发现不论我如何伪装，都无法骗过软件，在 python 中获取到数据。图3

当时我做到这里非常崩溃，因为真的找不到别的好的办法了。直到我发现了 mitmproxy

2.2 mitmproxy 的使用

mitmproxy 是一个支持 HTTP 和 HTTPS 的抓包程序，有类似 Fiddler 和 Charles 的功能，mitmproxy 还有两个关联组件，一个是 mitmweb，这是一个 web 程序，通过它我们可以清楚观察 mitmproxy 捕获的请求。一个是 mitmdump，通过它我们可以对接 Python 脚本，用 Python 实现监听后的处理。

正是因为 mitmdump 的存在，使得 python 可以以 requests 外的另一种形式介入，也正是因为这

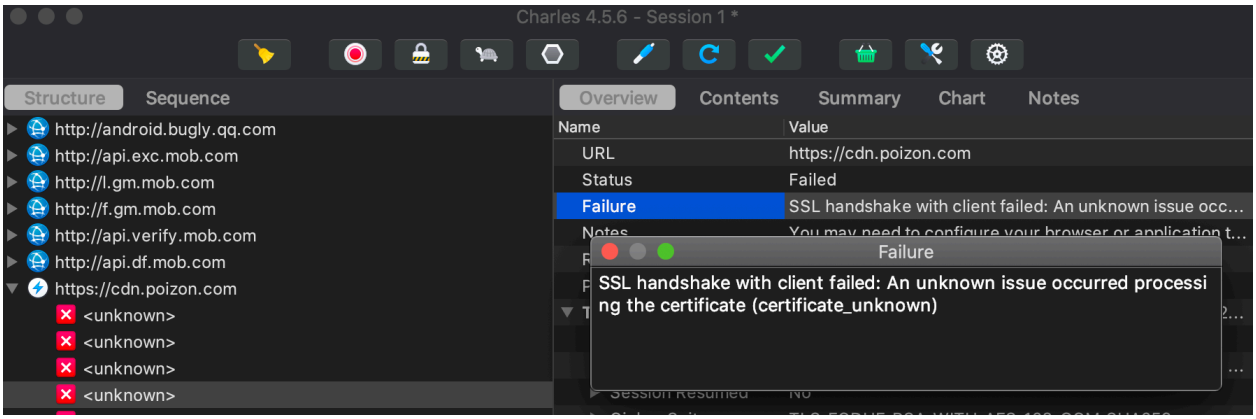

图 2: Charles 页面出现 Unknown

图 3: 在 python 中尝试通过 url 获取数据

一特性帮助我解决了我最大的难题，进而初步完成了任务。

开始之前请确保已经正确安装了 mitmproxy，同时配置好了 mitmproxy 的证书，具体配置要求请自行 Google。整个过程与 Charles 抓包非常类似。图4

在终端操作有些别扭，可以用 mitmweb 在 Web 窗口观察接口细节。图5

这里我已经能看到我需要的信息了。接下来我需要借助 python 把需要的数据提取出来。通过 mitmdump 可以将传输流发送到 python 脚本中，脚本通过过滤 url 得到所需文件，接下俩利用 re 库即可，不熟悉 re 库请自行 Google。

我将获取到的数据保存在了本地数据库中。图6

然而，通过这种方式获取数据最大的问题是我必须手动进入每一个产品页面，然后脚本才会捕捉到产品的 url 并执行存储操作。

有一个成熟的想法是：在浏览产品的界面，把所有产品的 url 线全部保存下来，然后逐个自动访问，这是我们爬 Web 数据的惯用手段。但很遗憾，诸如得物，京东 App 等已经非常成熟的软件，已经不再设置静态的 url 了，简单说就是仅仅知道某一界面的 url 没有任何意义，因为 App 会通过某种加密方法对 url 进行加密，加密的内容除了体现在 url 后的一个参数外，还体现在别的地方（我甚至感觉不在头部文件中，可能与其它的文件产生了关联），由于我没有破解其中玄机，因此只能主观猜测。

因此，如果想实现全自动地抓取数据，还需要借助其他工具。

2.3 Appium 的使用

Appium 是一个跨平台移动端自动化测试工具，可以非常便捷地为 iOS 和 Android 平台创建自动化测试用例。它可以模拟 App 内部的各种操作，如点击、滑动、文本输入等，只要我们手工操作的动作 Appium 都可以完成。图7

如果使用 Android 设备做 App 抓取的话，需要下载和配置 Android SDK，另外还需要配置环境变量。

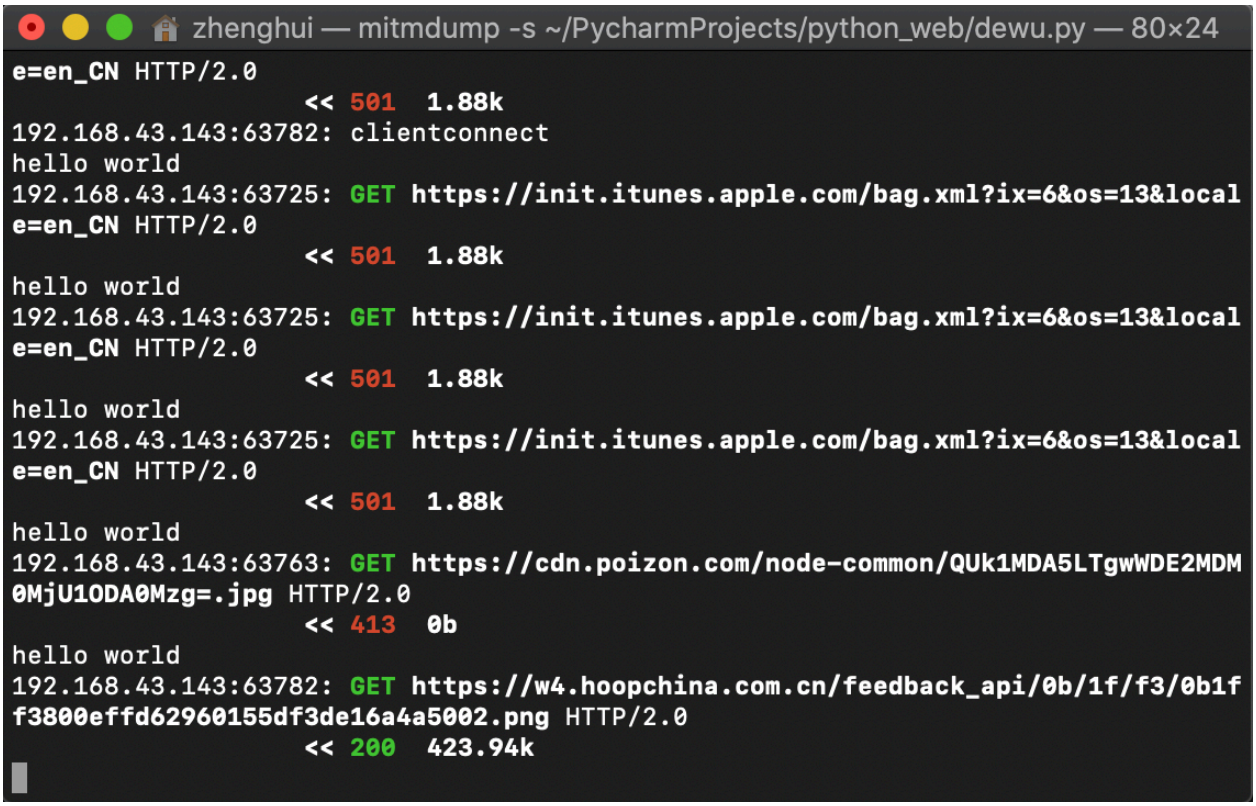

图 4: mitmproxy 捕捉到传输流

如果是 iOS 环境，那么通过 Apple Store 下载的软件都是禁止被 Appium 测试的（Apple Store 中的软件都携带了分发证书，而携带这种证书的应用都是禁止被测试的），只有通过获取 ipa 安装包再重新签名后才可以被 Appium 测试，具体方法请自行 Google。

之前已经用 iOS 做了很多工作了，此时再换成 6.0 以下的版本重新来我我有点接受不了，我更愿意接受半自动地获取我需要的数据……

实际上我觉得这个软件最原始的目的应该是用于移动端开发的测试的，作为一个移动端开发零经验的工程水水来说直接掌握这个软件没有太大必要，这一部分还是留到后面吧~

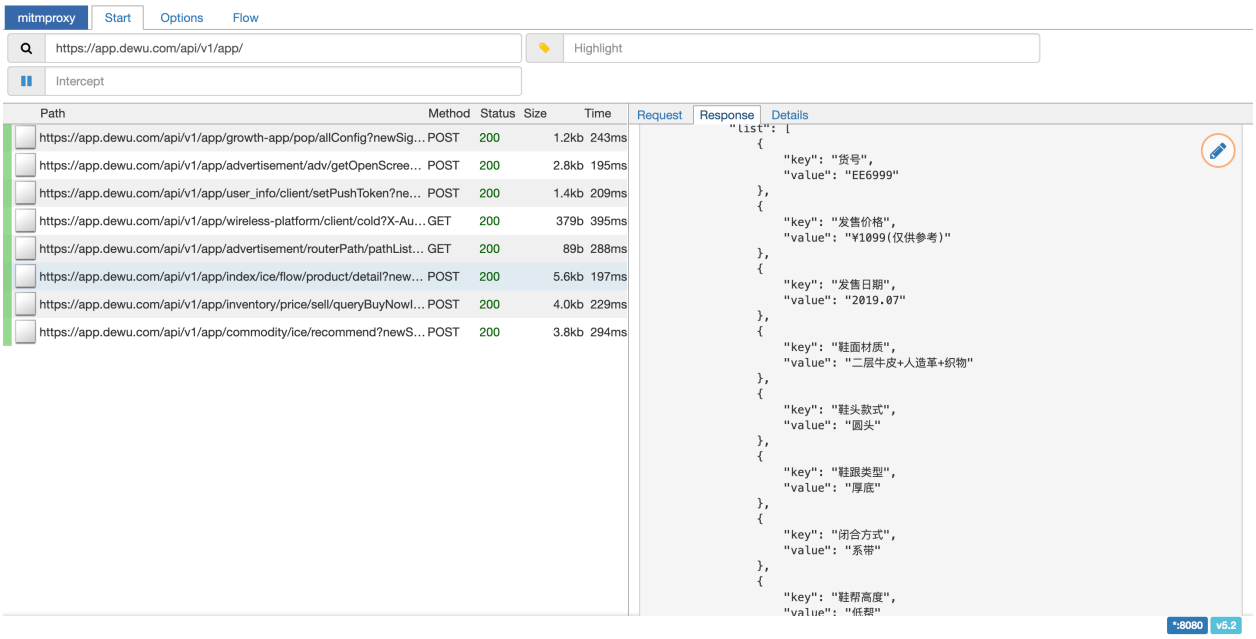

图 5: mitmweb 界面

Id	Brand	Title	Price	Img_url	功能性	适用季节	配色	鞋头款式	鞋帮高度	鞋底材料	鞋跟类型	风格
2182	Vans	Vans Authentic Golden Cc	31900	https://cdn.poizon	防滑	null	黑	圆头	低帮	牛津底	平跟	潮流,街头
2354	Nike	【张艺兴同款】Nike Air Fc	63900	https://cdn.poizon	null	春,夏,秋,冬	白	圆头	低帮	null	厚底	潮流,街头,
2357	Nike	Nike Air Force 1 MID 07 男	66900	https://cdn.poizon	包裹性	春,夏,秋,冬	白	圆头	中帮	塑胶底	平跟	运动,复古
3271	adidas ori	adidas Superstar J White	42900	https://cdn.poizon	null	null	白	圆头	低帮	null	平跟	null
8707	Vans	Vans Sk8 Hi black 黑白 板	41900	https://cdn.poizon	null	null	黑/白	圆头	高帮	null	平跟	null
8708	Vans	【刘耀文同款】Vans Old Skool Black 黑白休闲板鞋		oizon	null	春,夏,秋,冬	黑/白	圆头	低帮	牛津底	平跟	嘻哈,潮流,
10437	Nike	【虞书欣同款】Nike Air Fc	56900	https://cdn.poizon	null	春,夏,秋,冬	白	圆头	低帮	塑胶底	平跟	街头,复古,
10797	Vans	Vans Authentic Black Whi	29900	https://cdn.poizon	防滑	null	黑/白	圆头	低帮	橡胶底	平跟	街头
10798	Vans	Vans Classic Slip-On 黑白	30900	https://cdn.poizon	轻便	春,夏,秋	黑/白	圆头	低帮	牛津底	平跟	复古,街头,
10801	Vans	Vans Old Skool Navy 海军	31900	https://cdn.poizon	防滑	null	海军蓝	圆头	低帮	牛津底	平跟	街头
10903	Nike	Nike Air Monarch 4 White	43900	https://cdn.poizon	增高	春,夏,秋,冬	白/银	圆头	低帮	橡胶底	厚底	运动,复古
13017	Vans	【权志龙同款】Vans STYL	45900	https://cdn.poizon	轻便,透气	春,夏,秋,冬	深蓝色	圆头	低帮	橡胶底	平跟	运动
26063	Nike	Nike M2K Tekno 女款 白 蓝	86900	https://cdn.poizon	null	春,夏,秋,冬	白色	圆头	低帮	橡胶底	厚底	运动,潮流
30587	Vans	Vans SK8-HI 蓝黑	42900	https://cdn.poizon	防滑	null	蓝色/黑色	圆头	中帮	橡胶底,牛津	平跟	街头,嘻哈
41615	adidas ori	adidas originals Sambaos	40900	https://cdn.poizon	null	null	黑/白	圆头	低帮	null	平跟	null
43363	adidas ori	【易烊千玺同款】adidas o	66900	https://cdn.poizon	减震	null	黑	圆头	低帮	橡胶底	厚底	null
45331	Vans	Vans Sk8-Hi Reissue CAP	41900	https://cdn.poizon	null	null	黑/白	圆头	高帮	null	平跟	null
51276	adidas ori	adidas originals Ozweego	85900	https://cdn.poizon	防滑,轻便	春,秋,冬,夏	灰黑/黄	圆头	低帮	橡胶底	平跟	运动
65084	adidas ori	adidas originals Superstar	42900	https://cdn.poizon	null	null	白/黑	圆头	低帮	null	平跟	null
71010	adidas ori	adidas originals Superstar	38900	https://cdn.poizon	防滑,轻便	春,夏,秋,冬	白/黑	圆头	低帮	橡胶底	平跟	运动
72885	Vans	Vans Sk8-Hi 38 DX 白蓝	52900	https://cdn.poizon	轻便,包裹	春,夏,秋,冬	白/蓝	圆头	高帮	橡胶底	平跟	运动
78848	adidas ori	adidas originals Yeezy Bo	246900	https://cdn.poizon	null	null	黑	圆头	低帮	null	平跟	null
79375	adidas ori	adidas originals Yeezy Bo	161900	https://cdn.poizon	减震,防滑,	null	灰	圆头	低帮	null	平跟	null
1011657	adidas ori	adidas originals Yeezy Bo	161900	https://cdn.poizon	减震,防滑	春,夏,秋,冬	灰蓝	圆头	低帮	橡胶底	平跟	运动
1015288	Vans	Vans Style 36 白灰	41900	https://cdn.poizon	null	null	白灰	圆头	低帮	null	平跟	null
1018043	adidas ori	adidas originals Yeezy Bo	153900	https://cdn.poizon	耐磨	春,夏,秋,冬	黑蓝	圆头	低帮	橡胶底	平跟	运动
1033656	adidas ori	【王霏霏同款】2020款-ac	221900	https://cdn.poizon	减震,防滑,	null	黑白	圆头	低帮	橡胶底	平跟	运动,潮流
1034249	Vans	【吴亦凡同款】Vans Era 服	57900	https://cdn.poizon	防滑	null	黄/黑/蓝/	圆头	低帮	橡胶底	平跟	嘻哈
1035433	adidas ori	adidas Yeezy 700 V3 Arza	198900	https://cdn.poizon	减震,防滑	null	蓝灰	圆头	低帮	橡胶底	厚底	运动
1075922	Nike	Nike Blazer Mid '77 Vintag	98900	https://cdn.poizon	包裹性,防	null	白彩	圆头	中帮	橡胶底	平跟	街头
1125718	adidas ori	【丁禹兮同款】adidas orig	62900	https://cdn.poizon	防滑	null	灰/粉	圆头	低帮	橡胶底	平跟	运动
1141224	Nike	Nike Air Force 1 Hi LX \"G	99900	https://cdn.poizon	防滑,耐磨,	春,夏,秋,冬	白/蓝	圆头	高帮	橡胶底	平跟	运动,街头
1141239	Nike	【王源同款】Nike Air Forc	83900	https://cdn.poizon	防滑,耐磨,	春,夏,秋,冬	黑/白/蓝	圆头	低帮	橡胶底	平跟	街头
1142680	Nike	【周雨彤同款】Nike Blaze	73900	https://cdn.poizon	包裹性,防	null	白蓝红	圆头	中帮	橡胶底	平跟	街头
1148190	Nike	Nike Air Force 1 '07 Skelet	104900	https://cdn.poizon	防滑,耐磨,	春,夏,秋,冬	黑/橙	圆头	低帮	橡胶底	平跟	街头
1154329	Vans	Vans SK8-Hi 蓝紫	44900	https://cdn.poizon	防滑,耐磨	春,夏,秋,冬	蓝紫	圆头	高帮	橡胶底	平跟	街头,潮流

图 6: 获得的数据

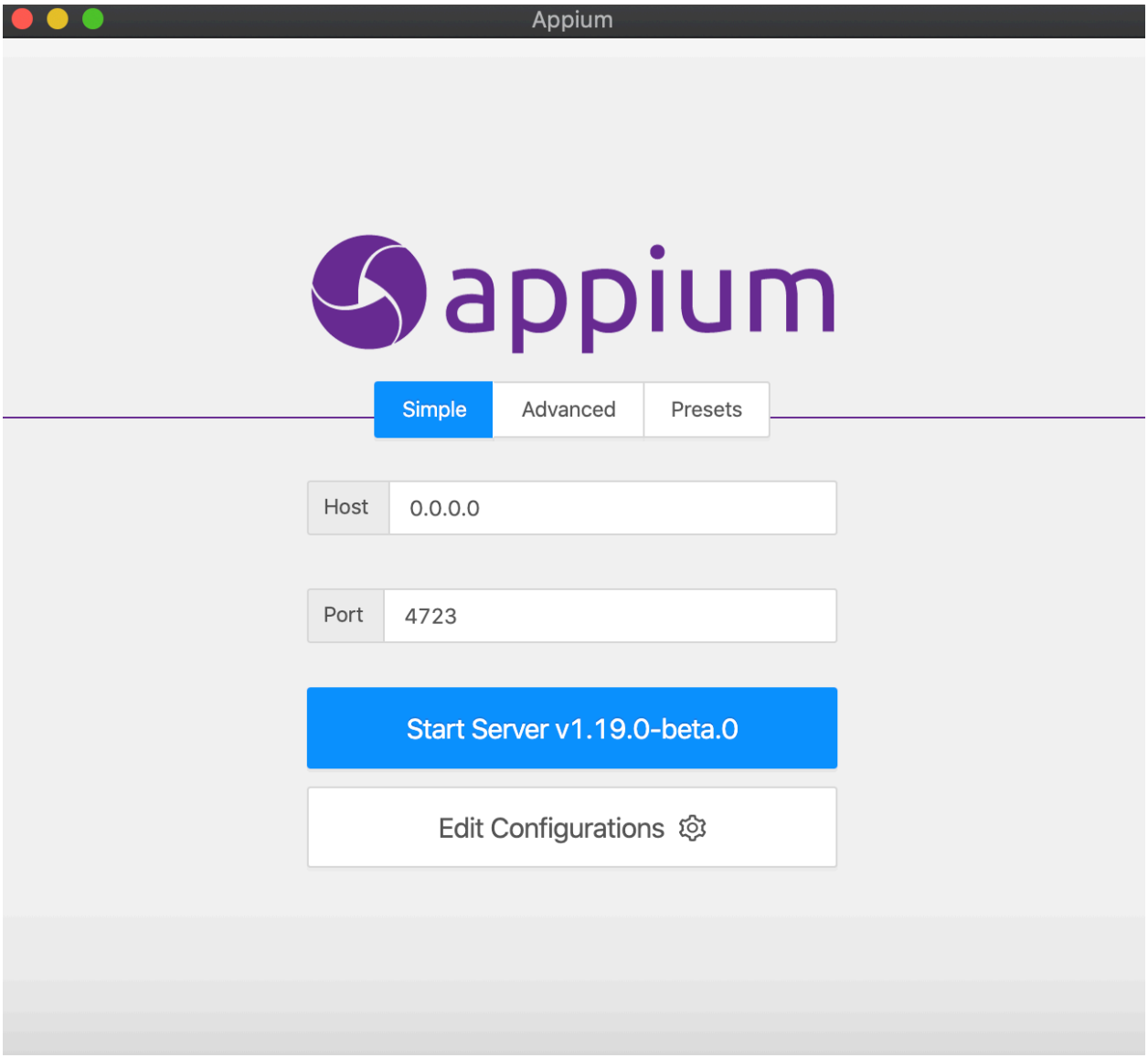

图 7: Appium 主界面