



# 基于ProgrammableWeb 的Web服务爬虫系统

**答辩人：刘羽鑫**

**导 师：刘建勋**



# 目录

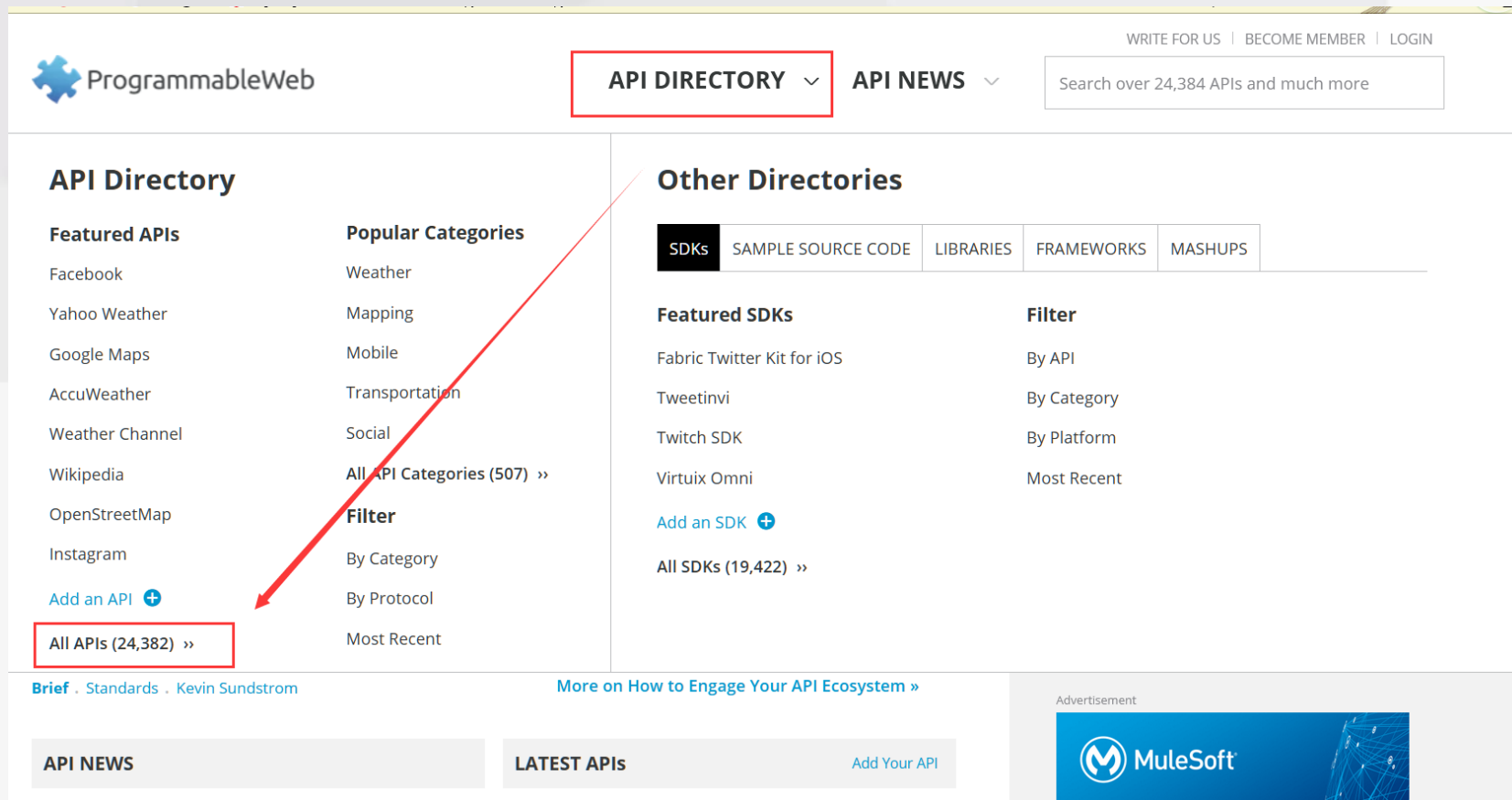
01 项目概述

02 项目使用的技术

03 项目的系统结构

04 项目结果展示

# 项目概述



爬取网站：ProgrammableWeb

爬取的数据来自四个模块：API、Mashup、SourceCode、SDK

数据导入MySQL数据库，使用Java语言编写前后端分离的系统进行展示数据

# 项目使用到的技术

## 爬虫程序

- **BeautifulSoup**
- **openpyxl**
- **pandas**
- **IP代理**

## 前端

- **Vue**
- **ElementUI**
- **Echarts**
- **Axios**

## 后端

- **SpringBoot**
- **Mybatis-Plus**
- **Redis缓存**
- **MySQL**
- **Maven**

# 项目的组成

## 爬虫程序

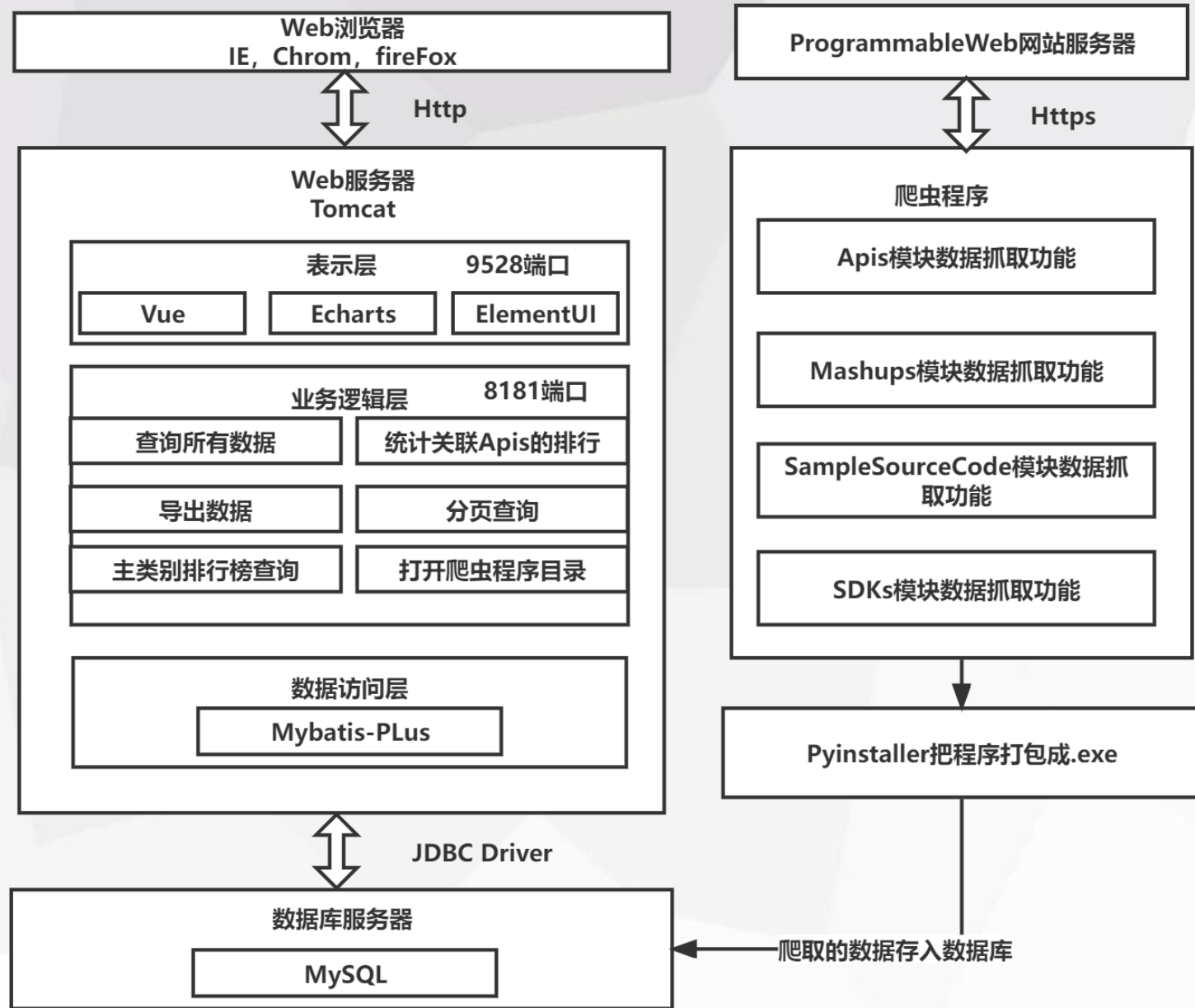
向PraxmableWeb网站发送请求，获取HTML文本，按照一定的规则获取需要的数据，然后把数据导入到数据库中

## 前端项目

负责把获取的数据通过Echarts图表展示，同时提供各种数据的查询，导出等功能

## 后端项目


提供接口数据，根据不同的SQL语句，获取不同的数据，把数据返回给前端



Web服务爬虫系统架构

# 项目结果展示

## Step 1: 爬虫程序数据抓取过程






[LEARN ABOUT APIS](#)

[WHAT IS AN API ?](#)

[API DIRECTORY](#)

[CORONAVIRUS](#)

Search over 24,382 APIs and much more



### Search the Largest API Directory on the Web

Search Over 24,382 APIs

SEARCH APIS


Filter APIs


By Category

☐ Include Deprecated APIs

API Name	Description	Category	Followers	Versions
Google Maps API	[This API is no longer available. Google Maps' services have been split into multiple APIs, including the Static Maps API, Street View Image API, Directions APIs, Distance Matrix API, Elevation API,...]	Mapping	4,039	REST
Twitter API	[This API is no longer available. It has been split into multiple APIs, including the Twitter Ads API, Twitter Search Tweets API, and Twitter Direct Message API. This profile is maintained	Social	2,422	Version

[Coronavirus Developer Resource Center](#)  
COVID-19 APIs, SDKs, coverage, open source code and other related dev resources »

 **Today in APIs**  
Latest news about the API economy and newest APIs, delivered daily:  
 [SUBSCRIBE](#)

 **API UNIVERSITY**  
[FEATURED](#) [LATEST](#)  
**FOR API PROVIDERS**

点击

Versions数据 可以在这个界面爬取

# 项目结果展示

## Step 2:爬虫程序数据抓取过程

Home » APIs » Google Maps

Google Maps API

MASTER RECORD

API Name

Mapping

Viewer

Categories

[This API is no longer available. Google Maps' services have been split into multiple APIs, including the [Static Maps API](#), [Street View Image API](#), [Directions APIs](#), [Distance Matrix API](#), [Elevation API](#), [Geocoding API](#), [Geolocation API](#), [Places API](#), [Roads API](#), and [Time Zone API](#). This page is maintained purely for historical and research purposes.] The Google Maps API allow for the embedding of Google Maps onto web pages of outside developers, using a simple JavaScript interface or a Flash interface. It is designed to work on both mobile devices as well as traditional desktop browser applications. The API includes language localization for over 50 languages, region localization and geocoding, and has mechanisms for enterprise developers who want to utilize the Google Maps API within an intranet. The API HTTP services can be accessed over a secure (HTTPS) connection by Google Maps API Premier customers.

+

 TRACK API

SEE ALL VERSIONS ↓



f

t

in

Description

Versions

SDKs  
(15)

Articles  
(228)

How To  
(7)

Source Code  
(29)

Libraries  
(23)

Developers  
(2575)

Followers  
(4041)

Changelog  
(179)

Google Maps Version History

Filter by Architectural Style:

All

Architectural Styles

Title	Style	Version	Status	Submitted
Google Maps REST API	REST	N/A	Recommended (Active, Supported)	12.05.2005



# 项目结果展示

## Step 3: 爬虫程序数据抓取过程

LEARN ABOUT APIS

WHAT IS AN API ?

API DIRECTORY

CORONAVIRUS

Search

icles (28)

How To (7)

Source Code (29)

Libraries (23)

Developers (2575)

Followers (4039)

Changelog (179)

DEVELOPERS (2575)

Developers Numbers

ADD MASHUP

Name

Mashup / Resource

espocrm

EspoCRM

groundswelldigital

BanksNearMe

Developers Names

Developers Mashup/Resource

LEARN ABOUT APIS

WHAT IS AN API ?

API DIRECTORY

CORONAVIRUS

Search o

icles (28)

How To (7)

Source Code (29)

Libraries (23)

Developers (2575)

Followers (4039)

Changelog (179)

FOLLOWERS (4039)

Followers Numbers

Track this API

Username

ANAS-AMEEN

CARRY-Minati

Erika-Villanueva

T-B

Followers Names

# 项目结果展示

<https://www.programmableweb.com/category/all/mashups?page=0>

<https://www.programmableweb.com/mashup/ behold>

<https://www.programmableweb.com/mashup/ behold/ followers>

```
{'mashups_name': 'Behold', 'related_apis': 'Instagram Graph', 'description': 'Add Instagram fee  
JSON API. No API keys or back end code required. Perfect for SPAs and JAMstack.', 'categories':  
'03.23.2022', 'mashup_app_type': 'Web', 'company': '', 'followers_numbers': '1', 'followers_na  
row: 2
```

<https://www.programmableweb.com/mashup/ motics>

<https://www.programmableweb.com/mashup/ motics/ followers>

```
{'mashups_name': 'Motics', 'related_apis': 'Google AdMob', 'description': 'Motics helps you sta  
Mopub. You have all your advertising data in one place.\nMotics has real-time data, view your  
date range.\nfeature:\n- Monitor app monetization with real-time reports.\n- Set goals and com  
impressions, fill rates, revenue and more from Admob.\nSupported ad networks:\n- Admob.\n- Mop  
.com', 'categories': 'Advertising###Mobile', 'submitted_date': '03.22.2022', 'mashup_app_type'  
'0', 'followers_names': ''}
```

row: 3

<https://www.programmableweb.com/mashup/ daily-crypto>

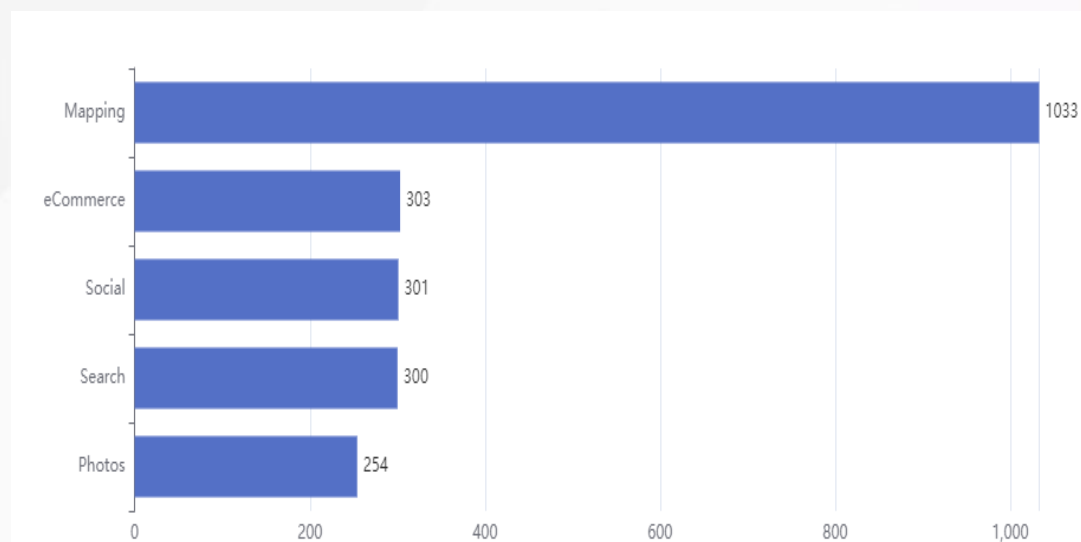
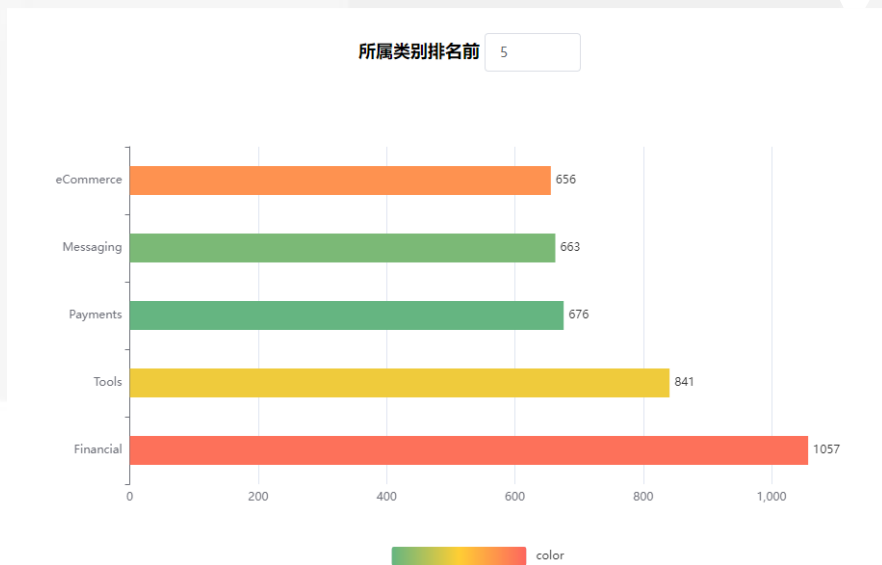
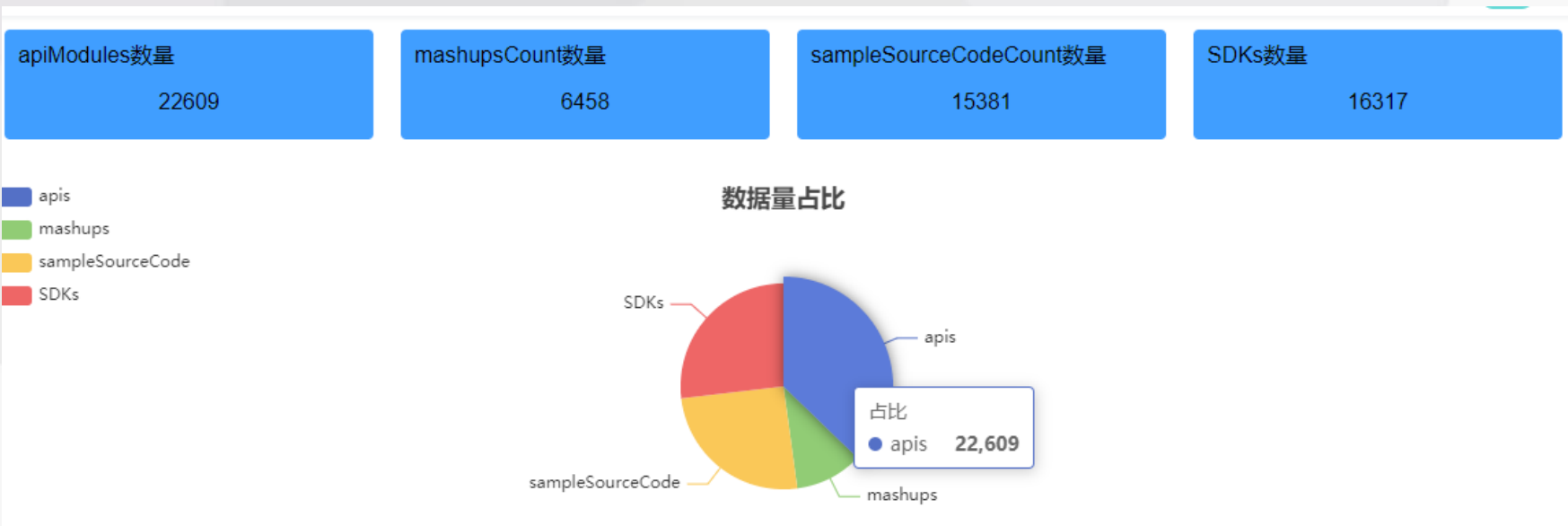
<https://www.programmableweb.com/mashup/ daily-crypto/ followers>

mashups_name	related_apis	descript
Tripotria Hotel Search	Expedia	Tripotria
Yoke	Zapier	Yoke is
Flex Mail	TEC Mailing	Flex Mai
Perfecto	Bike Index	Perfecto
Tripotria	Expedia	Tripotria
Local Weather	ipinfo.io IP geolocation##OpenWeatherMap	Local We
MIT Visual Traceroute	ipinfo.io IP geolocation	Visual I
Dinero.no	zanox##Finansportalen	Dinero.n
Your Automated China Investment Advisor	Highcharts##Highcharts Highstock	Your Aut
Open Data Companion (ODC)	CKAN##CKAN Ireland##Brazil CKAN##CKAN Czech Republic##CKAN Ital	The Oper
Kissmetrics on Shopify	Shopify Admin##KISSmetrics	Kissmetr
Segment	Pipl	Segment
Jabify	Telegram Bot	Jabify i
Posto	Resfly	Posto is
99designs Tasks Slack Bot	99designs Tasks##Slack Real Time Messaging	The 99de
Swyft	GTFS Data Exchange	Swyft is
Elsy Arres Flights	Pyton Flight Portal	Elsy Arr
Boom Cellar	Highcharts Highstock	Boom cel
Spotisoma	Spotify Metadata	Spotison
Api Expert - MyMemory Language Translator	MyMemory	MyMemory
Api Expert - World Weather Forecast	World Weather Online	World We
Api Expert - Yelp Local Business Search	Yelp Fusion	Yelp Loc
Api Expert - Stack Overflow	Stack Exchange API	Stack Ov
Api Expert - Indeed Job Search	indeed	Indeed J

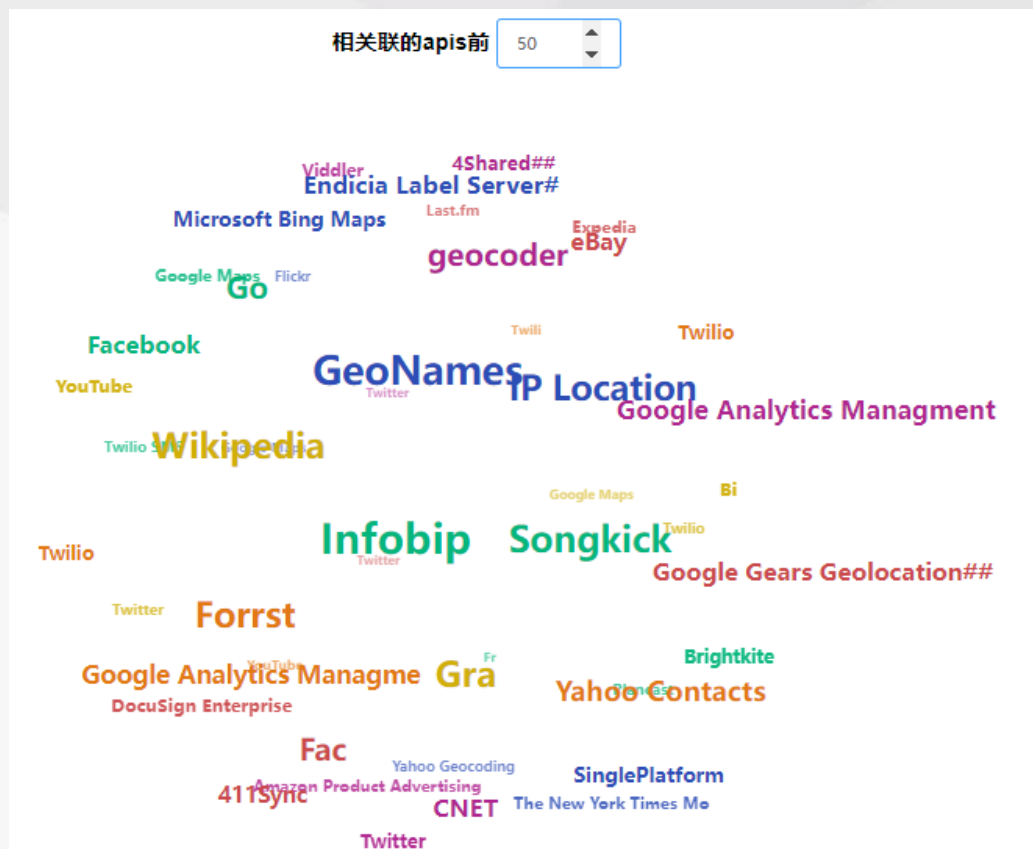
## 程序运行过程

## 爬虫数据

# 项目结果展示



# 项目结果展示



A###B###C分隔截取得到A



A###B###C分隔截取得到B, C

# 项目结果展示

导出所有数据

请选择

按照搜索

搜索

序号	sampleSourceCodeName	description	relatedApis	relatedPlatformLanguages	categories	
1	Rollideo API request wit...	Sample cod...	Rollideo	Python	Video	Fek
2	Geoapify Address Autoc...	Learn how t...	Geoapify	JavaScript	Location##...	Oct
3	Validation Service for C...	se this to di...	CO2 Offset	Node.js	Environme...	Sep
4	Firmalyzer IoTVAS API i...	This applic...	Firmalyzer Io...	Lua###Python	Security###...	Sep
5	IoTVAS NSE Script for n...	This is a N...	Firmalyzer Io...	Lua	Security###...	Sep

共 15381 条

5条/页

< 1 2 3 4 5 6 ... 3077 >

前往 1 页

## 功能

- 可将所有数据导出为Excel表格
- 根据不同的字段进行模糊分页查询
- 根据不同的字段进行数据排序

数据查询界面

# 总结

## 爬虫程序

存在一些冗余代码，正则表达式定义的规则不是那么好，导致了需要其他的操作弥补

## 前端项目


页面的布局样式有待提升，有许多爬虫的数据没有得到渲染

## 后端项目

自定义SQL过于复杂，导致执行时间过长，同时登录接口的实现代码过于简单

## GitHub仓库地址

<https://github.com/IntelligentServiceLab/Web-Service-Crawler>



谢谢