谷歌地图搜索结果抓取、基于谷歌地图数据的文本地址转经纬度坐标
技术细节及使用方法

朱潇圣 2020.04.25

技术细节

1. 谷歌地图搜索结果抓取

调用谷歌地图提供的 API（官方文档地址：https://developers.google.com/places/web-service/search）。

API 调用结构如下：

`https://maps.googleapis.com/maps/api/place/textsearch/output?parameters`

此 API 采用标准的网络请求方式。其中：

`output` 字段用于指定结果返回格式，可用值包括 **`json`** 和 **`xml`**。

`parameters` 字段用于指定结果特定限制参数，参数之间通过**`&`**连接。可设置参数包括以下项目：

必需参数 **`query`** 用于指定搜索词。

必需参数 **`key`** 用于指定本次请求归属的 Google Map API key。

Google Map API key 可在 Google Cloud Platform（https://console.cloud.google.com/google/maps-apis/overview）免费申请。

可选参数 **`region`** 用于按行政区划设定搜索界限。

按 ccTLD（顶级域名后缀）格式设定参数值。ccTLD 可用于区分不同的国家或地区，如 cn（中国大陆）、hk（香港）等。

可选参数 **`location`** 和 **`radius`** 用于按指定中心点和搜索半径设定搜索界限。

本组两个参数共同起作用，且与 **`region`** 互斥。

`location` 参数值按<latitude>,<longitude>格式设定。**`radius`** 参数值类型为数值，采用单位为米。

可选参数 **`language`** 用于设定搜索结果优先返回语言。

具体支持的语言和对应的参数值可在 https://developers.google.com/maps/faq#languagesupport 查看。

可选参数 **`pagetoken`** 用于直接指定要抓取的搜索结果页。当此参数指定值时，所有其他可选参数均不生效。

如果单次搜索结果过多导致 Google 地图默认将结果分页返回，则返回的结果中会有下一页结果的 **`pagetoken`**（搜索条件与对应前页相同）。

例如，如果希望按 json 格式抓取搜索词“大学”在以点(22.3,114.2)为中心、半径为 25km 范围的搜索结果，则可以发送下面的网络请求：

`https://maps.googleapis.com/maps/api/place/textsearch/`**`json`**`?`**`key=***`**`&`**`query="大学"`**`&`**`location=22.3,114.2`**`&`**`radius=25000`**

之后，通过返回结果中的 `pagetoken`，可循环抓取后续分页的搜索结果，直到返回的 `pagetoken` 为空或未返回：

`https://maps.googleapis.com/maps/api/place/textsearch/`**`json`**`?`**`key=***`**`&`**`pagetoken=***`**

2. 基于谷歌地图数据的文本地址转经纬度坐标

通过 Python 模块 geocoder 间接调用谷歌地图提供的 API，省时省力。

（geocoding API 官方文档地址：https://developers.google.com/maps/documentation/geocoding）

geododer 模块在目标服务为谷歌地图时的调用格式：

`geocoder.google(<字符串类型文本地址>, key = <字符串类型 Google Map API key>)`

返回的对象中包含 `latlng` 属性，属性值即为输入的文本地址在谷歌地图数据库中对应的经纬度坐标。

使用方法

textsearch 在满足下列要求的 Python 环境中可稳定运行：

Python 3.6.5 & libs(requests 2.18.4 & json & time & re)

通过修改程序中的 `search_words`，可自定义搜索词列表。程序将自动分别对列表中每个搜索词一一抓取搜索结果。

通过修改程序中的 `request_url`，可自定义上述 API 参数。

被注释掉的`#search_result` 为搜索词“麻将”的搜索结果，供调试使用。如需提取搜索结果中的更多字段，可参考此搜索结果。

name2latlon 在满足下列要求的 Python 环境中可稳定运行：

Python 3.6.5 & libs(geocoder 1.38.1 & time)

通过修改程序中的 `location_list`，可自定义待搜索的文本地址列表来源。默认为读取同目录下 positions20200421.txt 中的内容，每行一条文本地址。可按注释掉的列表格式直接在代码中指定文本地址列表。