

# 地址处理相关（调用地图API和翻译API）

---

## 任务目标

### 需要解决的问题

1. 已有的酒店信息中的地址信息不全或者不同数据源的信息格式不一（例如详细程度和省略程度不一样）。
2. 针对后续需要进行的酒店聚类任务，需要足够规范的地址以及经纬度等信息来进行聚合。

### 目标结果

生成包含原始数据，翻译数据和规范化数据的json文件。

### 解决思路

采用高德地图的api，将原本未处理的地址信息发送给api接口，记录下返回的分层后的规范化的信息。

1. 首先需要注册一个高德地图的开发者账号，每个免费账号都有一定的额度。
2. 使用request模块来发送请求给api接口（需要附上注册账号的key），设置好参数之后接受返回的数据。
3. 由于原始的地址中可能会出现中英混合的情况（高德api只接受中文），在直接发送数据之前还需要调用翻译api来处理原始数据，这里我用的是百度翻译api（需要付费）。
4. 对返回的数据进行最后的处理和整合。

## 具体实现

### 阶段一，数据翻译

#### 代码获取

代码在Github上，叫“Baidu\_Text\_transAPI.py”。作用是对地址进行翻译，生成一个json文件，里面储存了翻译api的返回结果（注意，这里的文件里数据是没有酒店名称的，但是是有顺序的。在后续处理中，我们需要按顺序匹配对应的酒店）

#### 代码使用

在“Baidu\_Text\_transAPI.py”需要修改以下内容：

1. 身份认证信息（注册百度翻译开发者账号后获得）

```
appid = ''
appkey = ''
```

2. 打开的原始数据的位置和名称。
3. 储存的翻译后的文件名称和路径。

在python环境下运行代码（需要安装json和requests库）

## 代码结果

生成的文件格式如下：

```
[
  {
    "from": "en",
    "to": "zh",
    "trans_result": [
      {
        "src": "酒店地址",
        "dst": "酒店翻译后的地址"
      }
    ]
  }, ...
]
```

然后我还有写一个数据处理的代码，叫“save\_location.py”（你可以在Github上获得），它可以将上面的结果进一步整合。

修改代码里的文件读取路径之后，运行代码的结果将会是：

```
[
  {
    "Hotel": "铜锣湾迷你精品酒店",
    "original_address": "铜锣湾新会道8号",
    "processed_address": "铜锣湾新会道8.号"
  },
  {
    "Hotel": "香港九龙酒店",
    "original_address": "尖沙咀弥敦道19-21号",
    "processed_address": "尖沙咀弥敦道19-21号"
  },
  {
    "Hotel": "香港Casa",
    "original_address": "九龙油麻地弥敦道487-489号",
    "processed_address": "九龙油麻地弥敦道487-489号"
  },
  {
    "Hotel": "香港君怡酒店",
    "original_address": "尖沙嘴金巴利道28号",
    "processed_address": "尖沙嘴金巴利道28号"
  }, ...
]
```

这个数据文件将会用于下一个阶段。

## 阶段二，地图API的调用

## 代码获取

代码在GitHub上，叫"MAP.ipynb"。作用是对翻译处理后的地址进行规范化，生成的文件中会含有原始数据，翻译数据和最终的分层后的地址。

## 代码使用

代码的类型是ipynb，需要能够运行的环境，并且有安装json和requests库。

使用前，需要修改一些设置：

1. 读取的翻译文件的路径（即上一阶段生成的文件）。
2. 目标国家和省份（注意，对于中国境外的地址来说，高德api不提供免费服务，所以没有权限）

```
# 修改payload，这个数据结构将会发送给api
params = {"key":api_key,
          "address":proaddress,
          # "country":country,
          "province":province
        }
```

运行代码，生成结果。

## 注意事项

1. 可能会出现返回失败的情况，请注意是不是额度不够还是本身查询失败。
2. 数据中有些时候分层会比较详细，有些时候可能会是空的。

## 代码结果

数据结构如下：

```
[
  {
    "Name": "香港九龙酒店",
    "original_address": "尖沙咀弥敦道19-21号",
    "processed_address": "尖沙咀弥敦道19-21号",
    "formatted_address": [
      {
        "formatted_address": "香港特别行政区香港特别行政区油尖旺區弥敦道19号-21",
        "country": "中国",
        "province": "香港特别行政区",
        "citycode": "1852",
        "city": "香港特别行政区",
        "district": "油尖旺區",
        "township": [],
        "neighborhood": {
          "name": [],
          "type": []
        }
      }
    ],
  }
]
```

```
    "building": {
      "name": [],
      "type": []
    },
    "adcode": "810005",
    "street": "弥敦道",
    "number": "19号--21",
    "location": "114.176651,22.292835",
    "level": "门牌号"
  }
]
},...
```