对外API文档 - 通用信息抽取

功能介绍

通用信息抽取旨在从常见的不同领域下的大量文本中抽取出其语义相关的实体。本服务提供的是基于 **UIE** 进行 **zero-shot** 的通用信息抽取模型，使用的数据集为 **“CLUENER2020”** ，包含从新浪新闻检索到的约74万篇新闻文章，拥有来自不同领域的14个新闻类别，包括金融、股票、教育、时尚、体育、游戏、娱乐等，目前支持抽取的实体schema为 **“地址（address）”，“书名（book）”，“公司（company）”，“游戏（game）”，“政府（government）”，“电影（movie）”，“姓名（name）”，“组织机构（organization）”，“职位（position）”，“景点（scene）”**，由于是zero-shot，所以后续的迭代支持对schema进行任意拓展和修改。

在线服务接口定义

服务路由

general_info_extraction

入参说明

字段	类型	说明
input_data	List of String	常见的不同场景的文本内容

入参示例

```
input_data = [
    '''6月15日，河南省文物考古研究所曹操高陵文物队公开发表声明承认：“从来没有说过出土的珠子是墓主人的''',
    '''诸侯曰类宫。”东汉蔡邕的《明堂丹令论》解释为：“取其四面环水，圆如壁，后世遂名辟雍。”魏晋南北朝、'''
]
```

返回说明

字段	类型	说明
地址	list of string	"" 为未抽取到相关信息

书名	list of string	"" 为未抽取到相关信息
公司	list of string	"" 为未抽取到相关信息
游戏	list of string	"" 为未抽取到相关信息
政府	list of string	"" 为未抽取到相关信息
电影	list of string	"" 为未抽取到相关信息
姓名	list of string	"" 为未抽取到相关信息
组织机构	list of string	"" 为未抽取到相关信息
职位	list of string	"" 为未抽取到相关信息
景点	list of string	"" 为未抽取到相关信息

返回示例


```
[
    {
        "姓名": [
            "曹操"
        ],
        "机构组织": [
            "河南省文物考古研究所曹操高陵文物队"
        ]
    },
    {
        "书名": [
            "明堂丹令论"
        ],
        "姓名": [
            "蔡邕"
        ]
    }
]
```


📌 以下为内部内容

效果说明

本服务支持从常见的不同领域下的文本中抽取出其语义相关的实体，目前实体的schema为 **“地址（address）”，“书名（book）”，“公司（company）”，“游戏（game）”，“政府（government）”，“电影（movie）”，“姓名（name）”，“组织机构（organization）”，“职位（position）”，“景点（scene）”** 共十类；数据集 **“CLUENER2020”** 包含了从新浪新闻检索到的约74万篇新闻文章，拥有来自不同领域的14个新闻类别，包括金融、股票、教育、时尚、体育、游戏、娱乐等，在dev上进行zero-shot的效果如下：

	precision	recall	f1-score	support
address	0.48	0.24	0.32	373
book	0.05	0.01	0.02	154
company	0.14	0.00	0.01	378
game	0.53	0.33	0.40	295
government	1.00	0.01	0.02	247
movie	0.05	0.05	0.05	151
name	0.79	0.62	0.70	465
organization	0.27	0.72	0.39	367
position	0.62	0.24	0.35	433
scene	0.44	0.60	0.51	209
micro avg	0.42	0.32	0.36	3072
macro avg	0.44	0.28	0.28	3072
weighted avg	0.48	0.32	0.32	3072

数据集参考链接：https://github.com/CLUEbenchmark/CLUENER2020

参考分类句子（不同场景下的文本内容）：

1. 娱乐：

```
句子：中博传媒与新加坡家乐电影合作，也非常明确地体现了中博传媒的国际视野。拍摄《危险关系》
"label": {"movie": {"《危险关系》": [[36, 41]]}, "company": {"中博传媒": [[0, 3],
```

2. 金融：

```
句子：宁波正是中国人民银行力推的金融ic卡多应用的一个试点城市。这种金融ic卡在宁波被称为“市
"label": {"address": {"宁波": [[0, 1], [37, 38]]}, "government": {"中国人民银行":
```

3. 司法：

```
句子：应用法学研究所所长胡云腾日前在北京表示，最高法正对此起草相关司法解释，希望能尽快出台
"label": {"address": {"北京": [[15, 16]]}, "organization": {"应用法学研究所": [[0
```

4. 体育：

```
句子：欧冠中4比0大胜波尔图来得十分酣畅淋漓，阿德巴约和范佩西已经联手在各项赛事中打入11球，
"label": {"organization": {"欧冠": [[0, 1]], "波尔图": [[8, 10]]}, "name": {"阿德
```

临时API地址

http://192.168.50.121:8989/general_info_extraction

请求代码示例

```python
import requests
import json

if __name__ == "__main__":

    url = 'http://192.168.50.121:8989/general_info_extraction'  # 服务地址
    input_data = [
        '''6月15日，河南省文物考古研究所曹操高陵文物队公开发表声明承认：“从来没有说过出土的
        '''诸侯曰类宫。”东汉蔡邕的《明堂丹令论》解释为：“取其四面环水，圆如壁，后世遂名辟雍
    ]
    resp = requests.request("POST", url, data=json.dumps(input_data))
    print(json.dumps(resp.json(), ensure_ascii=False, indent=4))
```