研究方向：新冠轨迹文本中的时空信息识别

**一、概述：**从新冠轨迹文本中识别中文地名与时间信息，并实现地图的展示。

**二、举例：**

**原始文本：**

在北京市新型冠状病毒肺炎疫情防控工作第250场新闻发布会……刘某，居住在昌平区北七家镇宏福苑小区。10月18日上午步行到 白各庄菜市场购物，到宏福苑社区居委会交物业费、到供热公司交取暖费，下午到社区卫生所做理疗。19日上午步行去白各庄菜市场购物，……

**输出：**

**（1）地名识别：**北京市，昌平区北七家镇宏福苑小区，白各庄菜市场，宏福苑社区

**（2）时间识别：**10月18日，19日

**地图显示：**



**三、具体实现（参考）：**

**1. 从新冠肺炎确诊病患活动轨迹（**[**https://xw.qq.com/act/fytrace**](https://xw.qq.com/act/fytrace)**）公布的患者轨迹数据中，建立新冠轨迹时空标注数据集。**

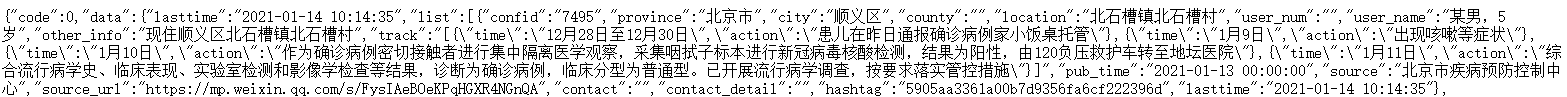
**1.1 数据爬取。**

数据接口：<https://pacaio.match.qq.com/virus/trackList?page=&num=>，根据每条轨迹给出的url（source\_url字段）爬取原始网页中对应的文本

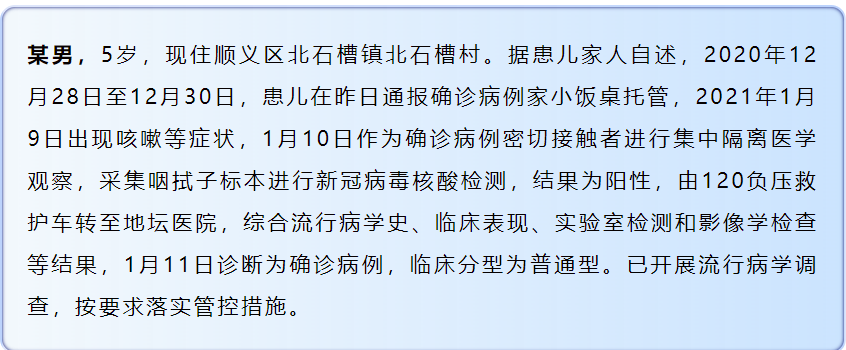
**1.2 根据接口中提供的所有记录，找出原始文本并映射地点（province，city，country，location等字段）和日期（time等字段）出现的位置。以自动处理为主，人工校对为辅的形式完成。**

**举例：**

（1）某记录：



爬取url（https://mp.weixin.qq.com/s/FysIAeBOeKPqHGXR4NGnQA）中的文本，url无效则忽略：

原始文本：

**1.3 字段标注：将原始文本处理成带标签的形式。具体标签见github中的参考资料，至少保留时间和日期的区分，不用标人名和机构。**

**举例：**

**Source：**某男，5岁，现住顺义区北石槽镇北石槽村。据患儿家人自述，2020年12月28日至12月30日，(略)

**Target：**O O O O O O O O B\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC I\_LOC O O O O O O O O O B\_T I\_T I\_T I\_T I\_T I\_T I\_T I\_T I\_T I\_T I\_T O B\_T I\_T I\_T I\_T I\_T I\_T I\_T O(略)

**2. 参考现有研究，设计中文时空信息识别算法，在数据集上进行评估。**

参考技术路线：https://github.com/jiangnanboy/albert\_lstm\_crf\_ner

**3. 对每条记录中识别出来的地名利用地图接口展示，实现轨迹点可视化系统。**