第八届“泰迪杯”数据挖掘挑战赛——

南都大数据研究院特别赛题：疫情通报文本中涉疫地点的自动提取

一、问题背景

随着我国“智慧城市”建设进程的不断推进，如何用“大数据”为公共服务提质加码是大数据研究者一直在不断为之努力的课题。自 2020 年初新型冠状病毒疫情发生以来，从中央到地方各级政府已形成常态化的信息发布模式，进一步推动了以“知情权”为基础的政府信息公开及数据开放的进程。应用大数据技术分析疫情通报信息，构建疫情发展模型，可以对传染源、传播速度、传播路径、传播风险等进行评估和预测。精细的疫情通报信息颗粒度，有助于获取疫情流散特征；完整、透明的疫情分析，可以探究疫情通报信息的质效；梳理确诊病患居住小区或逗留场所，追踪建立个体关系图谱，可以定位疫情传播路径，防控疫情扩散。本赛题提供收集自网络的官方通报和媒体报道网址，要求参赛者建立文本分析模型，自动提取包括确诊病例所在城市、行政区、小区或逗留场所等的涉疫地点分布信息。

二、解决问题

疫情期间全国有 400 多个城市公布了新冠肺炎确诊病例所在的居住小区或逗留场所等具体位置信息，这是公众非常关心的话题。从流行病学调查角度来看，分析新冠肺炎确诊病例居住小区或逗留场所，对寻找密切接触者有很大的意义，有助于社区进行有针对性地防控，使公众能够更好进行个人防护。在这些疫情通报或媒体报道的网页中，确诊病例的常住地点或逗留场所多嵌入在自然语言文本中，并以页面正文、页面内嵌文本和截图等多种形式表现。要从这些信息来源中及时分析涉疫地点的分布情况，首先要从包括自然语言文本和图片等非结构化数据中提取有关信息并转化为结构化数据的形式。这项工作以往主要通过人工对文本中的特定信息的查找与分类来实现，工作量大并且效率低下，缺乏时效性。

表 1 所示的附件 1 提供了南都大数据研究院收集的官方通报和媒体报道的网址，以及已经提取的涉疫地点数据。请参赛者利用自然语言处理和文本挖掘技术，从表 2 所示的附件 2 提供的 URL 对应的疫情通报中提取如附件 1 所示的包括城市、行政区、具体位置等涉疫地点的相关信息，并保存为文件“广东新冠肺炎确诊病例小区或场所信息.xls”。 对信息提取模型采用 F1-score 做为评价指标。

提取示范 1：

- 从化区（1个）
  城郊街新雍丽酒店
- 增城区（6个）
  荔城街中海城市广场
  新塘镇尚东阳光
  新塘镇太平洋花园
  新塘镇翡翠绿洲
  新塘镇西洲村豪门公寓
  新塘镇西洲村

A	B	C	D	E	F
province	city	district	place	source	link
广东省	广州	从化区	城郊街新雍丽酒店	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	新塘镇西洲村	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	新塘镇西洲村豪门公寓	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	新塘镇翡翠绿洲	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	新塘镇太平洋花园	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	新塘镇尚东阳光	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	广州	增城区	荔城街中海城市广场	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg

提取示范 2：

汕头市
新增3例
所住小区披露

据汕头市卫生健康局官微今天消息，2月2日，汕头市新增新型冠状病毒感染的肺炎确诊病例3例。截至2月2日24时，汕头市累计报告确诊病例17例，其中危重1例，重症2例，普通病例14例，无死亡病例。

新增病例具体情况如下：

病例15：女，32岁，现住汕头市龙湖区龙腾熙园，其丈夫（病例13）为确诊病例。1月23日随丈夫自驾车从武汉到汕头，25日隔离观察，31日出现咳嗽症状并入院就诊，目前病情稳定。

病例16：女，12岁，武汉人，1月23日随父母自驾车从武汉到汕头，居住在汕头市金平区东兴路11巷。29日入院就诊，目前病情稳定，其父母已隔离观察。

病例17：男，54岁，武汉人，1月21日一家四人自驾车从武汉到汕头南澳旅游，入住南澳县蓝海豪景，26日出现头痛乏力症状，31日入院就诊，目前病情稳定，其余3人已隔离观察。

广东省	汕头	龙湖区	龙腾熙园	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	汕头	金平区	东兴路11巷	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg
广东省	汕头	南澳县	蓝海豪景	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ1OaFofrnxWPtHREg

三、数据说明

表 1：附件 1 数据结构及示例

省份	城市	行政区	具体位置	数据来源	网址
广东省	广州	从化区	城郊街新雍丽酒店	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ10aFofrnxWPtHREg
广东省	广州	增城区	新塘镇西洲村	南方都市报	https://mp.weixin.qq.com/s/Y2tnLZ10aFofrnxWPtHREg
广东省	广州	增城区	新塘镇大墩村	南方都市报	https://mp.weixin.qq.com/s/N76fZ6R4zeiB6Zqa3TYaxg
广东省	深圳	南山区	鸿丰大酒店	南方都市报	https://mp.weixin.qq.com/s/KkYpkD4oV997_ZPNRa0vHg
广东省	珠海	金湾区	三灶镇万科城市花园	南方都市报	https://mp.weixin.qq.com/s/KkYpkD4oV997_ZPNRa0vHg

表 2：附件 2 疫情通报网址样例

网址
https://mp.weixin.qq.com/s/Y2tnLZ10aFofrnxWPtHREg
https://m.mp.oeeee.com/a/BAAFRD000020200203257725.html
http://wjw.gz.gov.cn/ztzl/xxfyyqfk/yqtb/content/post_5655251.html
http://static.nfapp.southcn.com/content/202002/04/c3062247.html?colID=17078
https://m.mp.oeeee.com/a/BAAFRD000020200203257514.html?layer=4&share=chat&isndappinstalled=16
http://wx.hznews.com/data/hzby_2019-nCoV/index_bl.php

附录：

请仔细阅读以下说明：

1、关于赛题数据

①建模数据：2020 年 4 月 25 日 9:00:00 公布。

②测试数据：2020 年 5 月 9 日 9:00:00 公布。

2、提交作品

①命名方式：论文命名为“南都特别赛题”，附件命名为“作品附件”，测试结果命名为“作品测试结果”。

②论文及附件内请勿出现队号、学校、学院、队员以及指导老师相关任何信息，否则视该作品为无效作品。

③请参赛队于 2020 年 5 月 8 日 16:00:00 之前在竞赛官网“提交作品”处提交论文（PDF 版，大小不超过 50M）及附件（包含论文正文（Word 版）、过程数据、程序、热点问题表、热点问题留言明细表的压缩包，大小不超过 200M）。

3、公布测试数据，提交测试结果

2020 年 5 月 9 日 9:00:00 准时公布测试数据，请在“赛题与数据”页面对应的题目右下方下载测试数据,并于 2020 年 5 月 10 日 9:00:00 前在“提交测试结果”页面提交测试结果。