开发环境及工具

* python3.80
* OpenPyXl 模块
* Re 模块
* datetime 模块
* os 模块

概要：运用自然语言处理等技术

（1）为什么会想到这个方式来进行处理?

首先谢谢老师的提问，当时这个项目是因为涉疫信息数据体量大、当前新冠流行病调查暴露出疫情急、数据大、信息来源复杂等问题，常规方法根本无法实现疫情追踪的应急处理和实时研判。同时，由于我们学校这边老师的推荐，让我们踊跃的报名参与，为了提高自己的科研能力，积极参与可了一些项目，所以来共同开发这样的一个平台，中大的导师称之为传染病智慧追踪系统，以实现疫情防控的快速反应和精准追踪

（1）详细介绍你在本项目中的任务?

首先谢谢老师的提问，我在这个项目中的主要任务就是将流调信息中的地名这个命名实体识别出来，但是由于地名的口语化和格式不规范的问题，在当时处理这个问题也是遇到了一些挑战的。

最开始应对的方式是：百度和 jieba 的分词，用于将流调信息中的地名识别出来，但是 jieba 分词的颗粒度太小了，也就是说他会将一个完整的地名切分开，试验下来分词的结果并不是很理想，

之后我们是用自定义库，也就是将流调信息中一些地名加入到我们自定义词典中去，但是效果确实不错，但是工作量太大，所以在实际中并没有得以采用；

之后偶然接触到，所以我们决定在用机器学习中的 NER 中的 LAC 工具来对数据进行处理。LAC 全称 Lexical Analysis of Chinese，是百度中用于自然语言处理的一个词法分析工具，它可以实现实现中文分词、词性标注、专名识别等功能的一个工具。LAC 的原理是，将地名（LOC）和机构名（ORG）识别出来，试验下来发现，精确度还是不够。

之后联想到可以采用字典树来实现的地名搜索，但是实际下来发现查找例如“海南省人民医院”搜索人民医院的话就查找不到这个地名，因为字典树将每一个结点看作是一个字符，那么只能强制从地名最开始查找；所以这个方法但是也并没有采用。

最后我想到了可以采用动态规划求最长公共子序列 LCS，它的原理就是用于比较串 A 和串 B 之间共有元素的长度，计算编辑距离，以此来衡量串 A 和串 B 的相似度，也就是说，它可以用于衡量当前查询的地名和 POI 的地名的差异和相似程度。当时，在我们在 LCS.py 中也实现了将所有的相似度进行从大到小的排序，并返回相似度最大的地名，使得精确度最

大。
（补充：

LCS 计算编辑距离的原理：

首先 LCS 求最长公共子序列是是动态规划 dp 问题，所有可能的情况只有三个，即删除、增加和修改，对于此类问题，我们需要写出动态转移方程。

该算法要求我们按顺序对比字符，若 a 串当前字符与 b 串当前字符相等，那么直接从前一个字符继承修改次数，否则选择删去该字符或替换该字符，或增加该字符，因此分别从两串的上一状态+1 来继承并延续修改次数，对于当前字符，选择最小的一种修改方式赋值，用于构造最终最优解。

命名实体识别：(简称 NER)是 NLP 中的经典任务,即给定一个输入文本,让模型识别出文本中的实体信息。

）

（3）请介绍这个项目的主要流程

首先是需要进行文本预处理

* 文本语料在输送给模型前一般需要一系列的预处理工作，才能符合模型输入的要求，其中包括

1、将全角字符串统一为半角

全角情况下输入一个字符就会占用两个字符，半角情况下输入一个字符只占用一个字符)，我们可以通过正则表达式将其识别并处理。

2、去除或修改指定无用的标点符号和干扰文字

3、将语义近似的词语替换为统一词语，　　我们是通过正则表达式将其识别并处理。

通过分析文本，我们发现统一修改一部分词语，不会影响文本含义，并且可以提高识别效率，简化文本结构。

接下来是地名这个命名实体的识别，

从最开始的用 jieba 分词，自定义库，之后机器学习的 LAC，最后是字典树和最长公共子序列 LCS，让识别地名的效率和精确度都有很大的提升）

（4）该项目是如何进行标准化的?

考虑到流调信息获取不全或不一致等现象普遍存在，我们是采用 NLP 中命名实体识别的方式，读取后台数据库信息进入到选择界面以得到准确的、标准化数据输入，实现命名实体的统一；

对于地理信息，需要调用百度地图 APP 获得其标准化数据输入，实现流行病调查活

动轨迹追踪过程地理信息的标准化。

利用百度地图 API 接口以及全国各地的公共交通信息，定位密切接触者的地理位置，并进行流行病溯源分析。这大大提高了防控队伍的信息捕捉能力，为预测高危地区和潜在高危地区提供了精准依据。

（5）你们开发的这个项目，你认为比现有的项目的优点在哪里？

首先谢谢老师的提问，首先这个项目：

1、用的尽可能标准化数据输入

考虑到流调信息获取不全或不一致等现象普遍存在，我们是采用 NLP 中命名实体识别的方式，读取后台数据库信息进入到选择界面以得到准确的、标准化数据输入，实现命名实体的统一；

对于地理信息，需要调用百度地图 APP 获得其标准化数据输入，实现流行病调查活动轨迹追踪过程地理信息的标准化。

利用百度地图 API 接口以及全国各地的公共交通信息，定位密切接触者的地理位置，并进行流行病溯源分析。这大大提高了防控队伍的信息捕捉能力，为预测高危地区和潜在高危地区提供了精准依据。

2、 同时，我们有一个自己的腾讯云服务器，可以提供实时的 web 服务，能支持实时控制系统工作并提供数据的相关处理，并实时更新相关信息。

3、 同时，因为这个项目是高校合作项目，然后我们这边目前所做的工作是前三步，也就是数据的预处理，命名实体的识别，地名的标准化，然后中山大学同学那边是拿到我们的数据之后再利用知识图谱等技术可视化处理，同时该项目是搭建了网络平台，在该网站上输入某个地方，在地图上就会显示该疫情的严重程度，我认为是一个比较有意义的一个项目。

（5）你们这个模型是如何实现的，用到了什么方法？

首先谢谢老师的提问。这个项目是通过数据预处理，然后进行命名实体识别，并提取出有用信息，例如航班号、地点等信息，方便后期实现疫情防控的、精准追踪。再将地名利用百度 API 进行经纬度的转换，获取标准化的数据，然后中山大学拿到这份数据之后再利用知识图谱等技术可视化处理，同时该项目是搭建了网络平台，在该网站上输入某个地方，在地图上就会显示该疫情的严重程度

（6）老师问：该项目的成果是什么？

首先感谢老师的提问，由于在大二的时候，我们参与该高校合作项目的队友，在自然语言处理知识只有一个初步的认识，在这个项目中贡献更大的是中山大学的同学，同时在当时，后面的项目是交给中山大学，我们这边是没有消息的。但是我认为更为重要的是从此次项目中，初步了解了如何查找查找文献，如何进行

外刊的阅读，同时当时我们是每周开一次组会，从组会中，见识到了一些更为优秀的人，也明白了自己和他们还有很大的差距，所以自己会继续努力，追求更高的目标

（7）老师问，你认为这个项目给你最大的收获是什么？

首先谢谢老师的提问。我认为此次项目我最大的收获是，学会走出了自己的舒适圈：起初我特别担心自己能力不够，害怕自己不能胜任这份任务，但是为了挑战自己和提升自己的科研能力，这次项目确实算作是一个不错的经历。于是在大二时，从 0 开始学习自然语言处理的一些知识，更为重要的是从此次项目中，初步了解了如何查找查找文献，如何进行外刊的阅读，同时当时我们是每周开一次组会，从组会中，见识到了一些更为优秀的人，也明白了自己和他们还有很大的差距，所以自己会继续努力，追求更高的目标。

（8）你认为什么是自然语言处理？

NLP（Natural Language Processing，自然语言处理）是计算机领域以及人工智能领域的一个重要的研究方向，它研究用计算机来处理、理解以及运用人类语言，能够用于文本分析，情感分析

文本分类：计算机能够采集各种文章， 进行主题分析，从而进行自动分类

- 机器翻译：计算机具备将 种语言翻译成另一种语言的能力
- 情感分析：计算能够判断用户评论是否积极
- 舆论分析：计算机能够判断目前舆论的导向
- 智能问答：计算机能够正确回答输入的问题
- 文摘生成：计算机能够准确归纳、总结并产生文本摘要
- 知识图谱：知识点相互连接而成的语义网络

（补充：什么叫做语义网络：
语义网络通俗上理解为由语意构成的网络，该网络中，计算机可以理解人类的语言，同时人类可以和语义网络对话，该网络我认为是对于现有网络的一个突破和之后的一个热点所在。

（9）你认为自然语言处理有哪些应用？（即可以用于干什么）

见上