**#Instructions**

* **Please make a copy before you edit it.**
* **Please submit the new copy access link in the** [**submission form**](https://forms.gle/S6GZ48U9F7t6jexHA)**.**

|  |
| --- |
| **2020 Google Girl Hackathon VI Semifinal Round Project Submission** |
| Project Name:安心出行 |
| Group Name: 珍珠奶茶大杯全糖 |
| Group Members: Chinese Name/Pinyin Name  薄德芳/Bo Defang 崔苏苏/Cui Susu 潘海琪/Pan Haiqi 王立岩/Wang Liyan 王明慧/Wang Minghui |
| **Group Member Roles & Responsibility**  Please brief introduce the contribution of each group member: the specific tasks or duties that each member complete.  薄德芳/Bo Defang：前后端开发，文档写作  崔苏苏/Cui Susu：问答模块开发，文档协作  潘海琪/Pan Haiqi：危险指数模块开发，文档写作  王立岩/Wang Liyan：前后端开发，数据采集  王明慧/Wang Minghui：前后端开发和交互 |
| **Brief summary**  Please summarize your problem statement and solution in a short paragraph.  2020年初新冠肺炎疫情爆发以来，全球贸易往来、经济发展和社会生活均遭受严重影  响。各国政府实施强制防控政策，人们大都居家隔离,交通出行都受到了极大的限制，但随着疫情防控形势的日渐好转，复工复产的需求日渐上涨,人口流动不可避免。在中国大陆，年初爆发的疫情曾经得到控制，国内大多数地区已较长时间没有本土病例，但是由于新冠病毒可能长时间存在，例如近期北京又出现了新一轮疫情，因此人们亟需得到有效的日常出行防护建议来避免疫情高危的地区或针对特定的地区做有针对性的防护,从而降低日渐频繁出行的患病风险。  我们选择的主题是帮助人们在疫情期间尽可能多地得到信息。由于新型冠状病毒的人  际传播途径为呼吸道飞沫或接触传播，故减少人群聚集和人口流动是最有效的防控疫情和降  低风险的手段。本作品基于实时人流量和交通信息，区域面积、人口和GDP等数据对一个地区的疫情危险指数进行建模，结合知识图谱以问答系统的形式来为人们的出行提供科学有效的建议。不同于目前中国政府发布的仅与当地确诊病例有关的风险等级和一些机构基于确诊病例时间序列信息构建的风险指数，本作品中的危险指数更多地反映了人群拥挤程度与感染新冠及其他具有类似传播途径的传染病的风险之间的关系。受限于现有数据和国内疫情发展状况，本作品的应用范围目前暂定为北京市，但是本作品可作为一个通用的系统，接入其他城市或者其他类似传染病疫情的数据来拓宽应用范围。 |
| **Problem Statement**  What's the background of the chosen theme? (brief introduction is enough)  What is the specific problem on the chosen sub theme? Please mention the theme your solution caters to. (Multiple selections is acceptable.)  新冠肺炎疫情的爆发彻底改变了人们原有的生活秩序，为防控新冠疫情,各地曾纷纷  采取“封城”措施,加大对人口流动的限制。严格的人口流动限制使服务业遭受重创，特别  是零售、休闲、酒店、娱乐以及交通运输等需要实体互动的行业。由于这些行业在总体就业  岗位中所占的比重超过四分之一，随着企业营收减少，失业率可能大幅增长,从而使经济供  给侧的冲击进一步转变为更大范围的需求侧冲击。因此，为了维护社会稳定,在疫情形势逐  渐好转时,尽快恢复正常的生产生活是必须的。同时，传染病专家称新冠病毒目前仍在全球  扩散，几乎可以确定它不会消失。事实上在本次比赛初赛提交后，中国大陆的东北和北京相继出现了小规模的疫情爆发。这些事实告诉我们在很长时间内人类不可能恢复不需要任何防护的生活，这就对人们在恢复正常人口流动后的自我防控提出了很高要求。  我们选择的主题是帮助人们在疫情期间尽可能多地得到科学有效的信息。具体为帮助  人们在出行前了解目的地的是否为感染新冠或其他通过飞沫传播的疾病风险较高的地区，以及提供有效的防护建议，从而降低患病风险。我们希望能基于实时人流量和交通信息，人口密度和GDP等数据来评估某个具体的地区的出行危险指数并通过构建知识图谱以问答系统的形式来为普通用户人群提供科学有效的出行建议，从而指导用户针对不同的地区来进行针对性的防护，减少人群的聚集，抑制疫情的发展并降低患病的风险。受限于现有数据和国内疫情发展状况，本作品的应用范围目前暂定为北京市，但是本作品可作为一个通用的系统，接入其他城市或者其他类似传染病疫情的数据来拓宽应用范围。 |
| **Problem Factor**  What is the underlying reason behind the problem  新型冠状病毒和其他很多呼吸道疾病的人际传播主要途径为飞沫传播，人群聚集和人口流动是此类疾病蔓延的重要原因之一。2020年初在中国武汉发生的疫情由于遭遇春运这一人群聚集和人口流动程度极大的时期导致一月底至二月初国内疫情迅速扩散和爆发，而中国政府执行如“武汉封城”等一系列限制人口流动的措施两周（新冠潜伏期）后国内新增确诊病例数开始平稳并阶段性下降。从累计确诊病例数量看，与湖北相邻省份及东南沿海经济发展水平较高、人口密度大和人口流动大的省份累计确诊数在国内排名靠前。基于以上事实，本作品尝试基于实时人流量和交通信息这两种可以一定程度上反映人群聚集程度和人口流动性的数据，结合人口密度和GDP等，参考政府和机构现有的疫情风险评估办法来计算某个具体的地点的出行危险指数并通过构建知识图谱以问答系统的形式来为普通用户人群提供科学有效的出行建议，从而指导用户针对不同的地区来进行针对性的防护，减少不必要的人群聚集和人口流动，抑制疫情的发展并降低患病的风险。不同于目前政府发布的仅与当地确诊病例有关的风险等级和一些机构基于确诊病例时间序列信息构建的风险指数，本作品中的危险指数更多地反映了人群拥挤程度与感染新冠及其他具有类似传播途径的传染病的风险之间的关系。受限于现有数据和国内疫情发展状况，本作品的应用范围目前暂定为北京市，但是本作品可作为一个通用的系统，接入其他城市或者其他类似传染病疫情的数据来拓宽应用范围。 |
| **Use Cases**  Describe specific use cases that illustrate the problem/opportunity.  本系统根据实时人流量和交通信息，区域面积、人口和GDP等数据，构造了一个危险指数来反映在一个地点感染新冠病毒风险的高低。同时，采集数据并构建面向疫情出行的知识图谱，结合深度学习算法以问答系统的形式为人们的出行提供科学有效的防护建议。本项目以网站形式呈现，我们使用了Google公司的产品Angular框架进行前端开发（另使用了Google公司的产品TensorFlow框架进行了深度学习算法开发，完成一个基于BERT的问答系统，但由于时间原因为整合到系统中，详情见Caveats）。  在本系统中，用户可进行两种操作，一是点击地图选择目标地点并查询出行危险指数，二是输入自然语言问句查询具体的出行防控建议。实例如图1所示。通过这两种操作，用户可评估出行的风险和获取防控措施，保障自身的安全。    图1实例 |
| Detailed Design  Adapted from front-end design section description and backend design section  Description. How does the overall design work? Details about the algorithms, schemas, encodings, etc., that will be used should be included here.  Guideline: If something's going to be 100 lines of code or more, it should be described here. You should also describe every major data structure and algorithm you use.  **主要技术路线**  本作品主要包含以下部分:   1. 危险指数计算   随着新冠肺炎疫情数据的持续发布，我们每日都可以从国家卫生健康委员会和地方卫生健康委员会的渠道获得全国各地确诊的累计人数。如何从现有发布数据中正确看待不同区域的疫情风险，成为公众经常关注的问题。现有的对地区疫情风险进行定性或定量的描述的指标主要有中华人民共和国国家卫生健康委员发布的疫情风险等级[1]和清华大学Aminer团队发布的新冠肺炎病毒区域性风险指标[2]，前者是对地域、时间、疫情数据来对一个地域进行定性的风险评估，后者是使用疫情数据的时序信息进行建模得出具体的数值。此外我们查到了一篇文献以公共安全的“三角形”模型为指导，考虑了突发事件即病毒本身的危险性、承灾载体即易感人群的脆弱性、应急管理的有效性 3个维度因素，建立了综合多个维度的疫情风险评估方法[3]。  然而，以上三种方法均无法实时反映人群聚集带来的感染新冠或其他呼吸道疾病的风险。本作品使用了腾讯位置大数据[4]提供的实时定位数据来反映人群拥挤程度和高德地图WEB服务接口[5]提供的交通路况数据来反映人口流动性，同时结合区域面积、人口和GDP等数据来构造危险指数，具体算法流程如图2所示。    图2 危险指数计算流程  具体算法步骤如下：   1. 计算易感人群的脆弱性   根据上文提到的参考文献，我们认为易感人群指暴露在外的易感染人群，其脆弱性与易感人群的空间密度及暴露程度有关，某一地区的 GDP 越高，人民生活越活跃，人群接触越多。此处计算易感人群的脆弱性计算公式如下：  （1），  单位为（万人×万亿元/），其中地区常住人口数与地区面积的比值为人口密度。本作品从政府网站收集了北京市各区的人口密度和GDP数据进行计算，若未来可获得更多的数据，我们可以对更多的地区进行计算。   1. 计算应急管理的有效性   根据上一步提到的参考文献，应急管理的有效性可用医疗资源相关的指标来衡量，在该文献中作者使用了累计治愈率来反映此指标，公式如下：  （2）。  本作品从清华大学提供的API[6]获取了北京市的累计确诊人数、累计死亡人数和现存确诊人数，由此计算出治愈出院人数。受限于数据，应急管理有效性只精确到了北京市，具体限制为北京市的累计死亡人数并没有精确到区级，所以无法计算出精确到各区的治愈出院人数。   1. 获取实时人流量和交通信息   为了使本作品中的危险指数可以评估人群聚集带来的风险，我们在计算过程中引入了实时人流量和交通路况。本作品中使用的实时人流量为腾讯位置大数据提供的实时定位数据，由于腾讯位置大数据API只对企业开放，故本作品从网页前端获取实时定位数据，从中根据经纬度选择离我们所需要查询的地点最近的区域的定位数据作为实时人流量。  关于交通路况，我们查阅了相关文献，认为景区和一些城市公共地点附近的人群聚集程度与其附近的交通拥堵程度是正相关的[7]，而交通拥堵程度的直接反映是道路平均车速，且平均车速与拥堵程度呈反比。因此，我们利用高德地图开放平台WEB服务提供的交通路况查询API来获取实时交通路况，具体为所查询地点经纬度为中心，半径为一公里内的路况。该API不仅提供实时道路平均车速，还提供定性的对交通是否拥堵的判断，本作品认为调用该API返回的数据若显示道路畅通，则平均车速视为经验值60（城市道路车速上限通常为60）以排除道路上少数车辆异常车速的干扰。   1. 图谱构建与问答系统   问答系统部分的主要功能是用户可通过输入自然语言问句的方式获取想要的信息，在本系统中主要为出行防控措施，也可查询地址、联系方式等。本模块主要包含图谱构建和问答两部分，主要流程如图3。    图3 基于AC自动机的问答系统流程   1. 面向疫情出行的知识图谱构建   本作品构建知识图谱所需数据使用了高德地图POI查询API进行采集，从高德地图POI分类中选择了餐饮服务、购物服务、生活服务、医疗保健服务、住宿服务、风景名胜、交通设施服务、公司企业共8类场所，获取了约十万条数据，并将以上类别的所有地点信息进行融合， 通过JSON进行数据规范化。同时，本系统根据网络上能查到的各类防控建议对将八大地点类型划分了三级防控建议，该数据同样进行JSON数据规范化。  本系统利用上述数据进行图谱构建，其中包括8类节点，具体指地点、类别、商圈、三级防控建议；4个地点属性，具体指经度、纬度、地址和电话；5个实体关系，具体指地点与类别关系、地点与商圈关系、类别与三级防控措施关系。本系统利用Neo4j进行图谱存储。   1. 字典匹配 2. WEB前后端   本作品以网站形式展现，网站采取前后端分离的开发模式，为人们提供出行目的地的危险指数查询和出行防护措施建议的功能。前端使用Angular进行开发，后端使用了Python语言和Django框架，前后端交互通过Restful API实现，存储部分使用的是Neo4j数据库。前端部分实现了点击地图查询和用户手动输入查询，主要使用了高德地图的JavaScript脚本，调用高德地图的一些原生的API来获取地址信息。同时为了用户的隐私安全，本系统不会收集用户的个人信息，用户无需登录便可直接使用网站的相关功能。网站架构如图4所示。    图4 网站整体架构 |
| **Caveats**  You may need to describe what you did not do or why simpler approaches don't work. Mention other things to watch out for (if any).   1. 危险指数计算及等级划分准确性   本作品中的危险指数计算尝试引入实时人流量和交通拥堵情况来反映人群聚集带来的新冠及其他具有类似传播方式的疾病的感染风险，但是囿于医学和社会学知识及数据，我们设计的计算方法略显简单，且没有足够的数据和时间去验证其准确性。同理本作品中的危险等级划分由于缺乏参考的标准，划分的阈值主要依靠的还是直观上的经验判断，难以做到非常精确。但是本作品引入实时人流量和交通路况来计算实时新冠危险指数的方法可为政府和其他拥有足够的数据的机构提供参考思路，帮助他们建立更好的疫情风险评估方式。   1. 构建问答系统的必要性   由于目前数据暂时不足，本系统中的知识图谱存放的实体和属性数量有限，目前问答系统的对话效果与用户自行选择给定地点进行数据库查询效果类似，但是本作品的目标是构建一个通用的系统，当能获取更多的数据和信息时，采用对话系统效果更佳。   1. 高德地图开放平台API额度限制   本系统使用的经纬度查询、交通路况查询等模块均使用了高德地图开放平台WEB服务API，然而高德地图API免费额度有限，故我们上传的源代码中没有包括调用这些API需使用的key，也没有部署以提供公网访问，防止盗用和恶意查询。此处提供一个key供评委运行系统时使用：   *  e5927609d1b678ae42bd7d6bfb9c4687  1. 代码命名规范   由于开发时间仓促，多人参与，故代码没有统一命名规范，现主要有下划线命名和大驼峰命名法。   1. 问答系统模型   由于时间原因，我们只完成了基于AC自动机的问答系统与web前后端的连接。我们另外单独完成了一个基于深度学习模型的问答系统，尝试解决基于AC自动机的问答系统存在的用户输入的地名需要与数据库中完全一致才能得到结果的问题。该系统与上述基于AC自动机的问答系统在图谱构建环节完全一致，流程如图5所示，区别之处在于增加了实体识别和属性链接，并增加了地名校正环节。    图5 基于BERT的问答系统流程   * 实体识别   由于地点类实体是用户容易输入不规范的实体，基于字典的实体检索容易导致检索失败，从而直接影响实体识别效果，因此本系统采用基于BERT的方式进行实体识别来弥补字典方法的弊端。基于BERT的实体识别旨在找到问句中询问的实体名称，也就是各类别地点名称。  由于本系统属于特定领域内的问答，缺失公开可用的语料库，因此，我们基于规则构造语料库，用于BERT地点实体训练。具体构造语料库的规则为：通过收集的北京市地点实体添加前缀、后缀的方式，构造问句、三元组和答案。通过此方式构造了33498组问答对。问答对样例如图6。  本系统构造实体识别数据集部分根据三元组-实体反向标注问题，给数据集中的问题打标签。本系统采用BIO的标注方式，用B-LOC, I-LOC标注问题中的地点实体。  本系统使用了BERT模型进行实体识别。BERT中蕴含了大量的通用知识，本系统利用预训练好的BERT模型，再通过上述构造的标注数据进行微调，从而获得此特定领域下的NER模型。在构造属性链接数据集时，由于问答关系只有三个，本作品采用NLPCC ICCPOL 2016 KBQA数据集实现BERT的文本相似度匹配，从而实现属性链接。首先从数据集中获得 4373 个关系 RelationList；然后通过语料库中的问题及三元组构造样本，一个样本由“问题+关系+Label”构成，原始数据中的关系值置为 1；从 RelationList 中随机抽取五个属性作为 Negative Samples。    图6 问答样例   * 地名校正     图7 地名校正流程图  在开发过程中我们发现知识图谱中存储的地名通常为全称，且地点过于详细，例如人们常说的“天安门”，数据库中存储的是“天安门广场”及若干景区附近的相关地点如“天安门服务部”、“天安门国际旅行社”等，而用户输入时通常只会输入口语中的俗称，这给地名查询增加了很大难度，无法完全对应并返回唯一值。若简单地选择经纬度最近的地点，则在一些特殊情况下可能出现错误，例如我们发现数据库中与“谷歌北京”经纬度最近的地点是不相关的“拜博口腔”，原因是仅根据经纬度只能计算地表距离，无法计算空间距离，而商业写字楼通常为多层建筑且有不同公司进驻，所以可能出现经纬度相近而完全不是同一个地点的情况。因此我们设计了一种结合经纬度和地名设计了一个校正流程，算法流程图如图7。  该校正流程首先使用经纬度对地点进行范围缩小，保证备选地点在用户查询的地点附近，然后使用编辑距离算法对用户查询的地名与备选地点的编辑距离值并进行排序，选择最小值对应的地名，若最小值对应的地名只有一个，则直接返回，否则再次通过经纬度选择距离最近的地点。编辑距离算法[8]是一种动态规划算法，它可用于计算两个字符串经过插入、添加、删除操作变成相同字符串所需的操作次数。该算法可解决地名简称与全称的匹配，例如用户输入的问句中只包含“天安门”，数据库中存放的只有“天安门广场”、“天安门服务部”、“天安门国际旅行社”、“时光漫步怀旧主题酒店(天安门广场店)”等，通过编辑距离算法计算“天安门”与其它字符串的编辑距离，选出最小值对应的字符串即可选择出“天安门广场”。然而数据库中包含“天安门”的地点数量庞多，为了减少计算量，我们选择在计算编辑距离前先通过经纬度缩小范围，范围设定为10个地点。 |
| **Final Result**  Please share what you achieve in this project. |
| **References and appendices**  Any supporting references, diagrams or demos that help depict your solution.  Any public datasets you use to predict or solve your problem.   1. 卫生健康委.疫情风险等级查询[EB/OL].http://bmfw.www.gov.cn/yqfxdjcx/index.html. 2. AMiner.新冠肺炎病毒区域性风险指标[EB/OL].https://covid-dashboard.aminer.cn/algorithm/?lang=zh. 3. 袁宏永, 陈涛, 黄丽达,等. 全国及湖北省新冠肺炎综合风险指数分析及比较[J]. 科技导报, 2020(4):21-28. 4. <https://heat.qq.com/> 5. <https://lbs.amap.com/api/webservice/summary/> 6. <https://github.com/AMinerOpen/covid_dashboard/blob/master/API.md> 7. 基于景区周边交通指数和区域客流饱和度指引智能交通出行决策的应用.基于景区周边交通指数和区域客流饱和度指引智能交通出行决策的应用[EB/OL].https://www.doc88.com/p-41799098719690.html,2018-06-29. 8. 百度百科.编辑距离[EB/OL].https://baike.baidu.com/item/%E7%BC%96%E8%BE%91%E8%B7%9D%E7%A6%BB. |
| **Appendix: Instructions**  Please provide a user guide for your source code and give instructions to implement your demo.  代码目录及主要文件说明：   * 问答系统部分   https://github.com/Estherbdf/SafeGo/KBQA\_Bert/  bert\  bert程序代码，里边是python文件  Data\  NLPCC\ 原始语料库（问题、三元组、答案）  NER\_Data\ NER任务数据集  Sim\_Data\ 属性链接任务数据集  - construct\_dataset.py 生成NER\_Data的数据  - construct\_dataset\_attribute.py 生成Sim\_Data的数据  ModelParams\  cased\_L12\_H768\_A12\ bert预训练模型  Output\  TensorFlow训练模型输出文件夹  以下是基于Bert的实体识别模块代码  - run\_ner.py main函数  - lstm\_crf\_layer.py  - conlleval.py  - colleval.pl  - run\_ner.sh  以下是基于Bert的属性链接（句子相似度）的代码  - run\_similarity.py main函数  - args.py run\_similarity配置  以下是问答模块的代码  - kbqa\_predict.py 问答main函数  - answer\_search.py 根据question和entity进行cypher查询  - min\_distance.py 编辑距离算法，实体间最小距离计算   * 危险指数计算部分   <https://github.com/Estherbdf/SafeGo/cal_risk/>  - cal\_risk.py 计算危险指数main函数  - get\_area.py 获取区域面积（限北京市）  - get\_district.py 获取所在区域（限北京市）  - get\_gdp.py 获取所在区域GDP值（限北京市）  - get\_population.py 获取区域常住人口数（限北京市）  - get\_key.py 获取高德WEB服务key（暂未上传至github，需要在引号中手动填入）  - get\_location.py 获取经纬度  - get\_speed.py 获取实时平均车速  - get\_visitors.py 获取实时人流量  - load\_cases.py 读入北京市累计确诊、治愈、死亡病例数  - update\_cases.py 每日更新累计确诊、治愈、死亡病例数  - cases.json 北京市累计确诊、治愈、死亡病例数等数值，每日更新   * 网站前端部分   https://github.com/Estherbdf/SafeGo/  webapp/  - setiings.py 项目的配置文件  - urls.py URL配置文件  safego/  - search.py 风险查询和措施查询的逻辑代码，项目主要代码  - urls.py URL配置文件  templates/  - search\_form.html 首页页面  - map-geo.html 按地图查询的页面  statics/ 存放img,css,js等静态文件   * 系统部署流程   这是一个为人们在新冠疫情期间提供出行建议的网站  ## 环境需要  neo4j-4.1.0， python3.x  ## 运行说明  1.在webapp根目录下安装所需的依赖  ```  pip install -r requirements.txt  ```  2.在cal\_risk目录下的get\_key.py中加入高德开发者key  ```  def get\_key():  return 'key'  ```  3.在KBQA\_AC目录下，对answer\_search.py和build\_graph.py修改neo4j的数据库和密码  4.将数据库graph.db.dump导入到本地neo4j中  a.首先关闭neo4j服务，将data/databases目录下的neo4j目录删除  b.然后执行一下命令导入数据库  ```  neo4j-admin load --from=PATHFORgraph.db.dump --database=neo4j --force  ```  4.在SafeGo的目录下，启动服务器  ```  python manage.py runserver  ```  5.打开浏览器访问页面  http://127.0.0.1:8000/safego/search-form/ |