**新冠肺炎疫情控制的知识图谱框架**

**项目计划书（初稿）**

**一、项目目标**

**背景：**

2020年新春伊始，新型冠状病毒肺炎（简称新冠肺炎）给我国带来非常重大的损失，也严重影响了人民的生产生活。在信息时代，互联网上，特别是移动互联网（微信、微博、门户网站等）传播了大量信息。但是这些信息是离散的，相互独立的；如通过知识图谱技术，将这些离散的文本信息提取出实体（关键词），并自动提取和推理出这些实体之间的关系，则可以将疫情控制相关的知识系统化、可视化的管理起来。本项目的输出后续也可用于疫情信息的管理。

**目标：**

基于2019年12月以来官方及大型新闻门户正式发布的新冠肺炎疫情相关信息和数据，开发出一套小型知识图谱，实现疫情相关知识图谱查询，可视化等功能。

**项目输出：**

1. 一套基于疫情控制知识图谱生成和查询软件；
2. 大数据及软件专业实训教案；

**二、项目框架**

**程序架构：**

图片包含 屏幕截图

描述已自动生成

程序分成数据源、知识图谱生成、知识图谱查询三个大模块，每个模块下又分成几个小模块。

**开发工具：**

本软件基于python开发，使用面向对象开发方法，并使用了tkinter、gephi、neo4j、scrapy、jieba、json等开源组件。

软件可跨平台使用，后续可扩展支持BS架构，或做成微信小程序等支持网页和移动终端访问。

**三、项目计划及分工**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **工作内容** | **完成日期** | **工作量估计** | **输出** | **责任人** |
| 完成项目计划书 | 2020-2-6 | 2天 | 项目计划书初稿；  完成任务分工（责任到人）  给出微博爬取用户名清单，过滤字典 | 叶老师 |
| 完成微博爬虫模块 |  | 7天 | 完成官微文本爬取 | 邱海龙、郑天安 |
| 完成网页爬虫模块 |  | 10天 | 完成多个官方网站文本爬取 | 江勇、杨源斌 |
| 完成主程序框架搭建 |  | 4天 | 完成主程序框架搭建 | 叶老师 |
| 完成关键词提取模块 |  | 5天 | 使用jieba分词完成关键词提取词典 | 学生 |
| 完成图谱数据库模块设计 |  | 5天 | 完成图关系数据库框架设计 | 叶老师 |
| 完成程序查询界面开发 |  | 5天 | 在程序主窗口实现查询功能 | 学生 |
| 完成关系图谱显示功能开发 |  | 13天 | 基于gephi组件，完成关系图谱显示功能 | 学生 |
| 完成知识图谱核心模块 |  | 15天 | 项目 | 叶老师 |
|  |  |  |  |  |

**模块说明：**

|  |  |  |  |
| --- | --- | --- | --- |
| **程序模块** | **子模块** | **主要功能** | **功能描述** |
| 数据源 | 微博爬虫模块  WeiboSpider | 爬取指定微博用户的微博文本，并实现关键词过滤，保存到指定csv文件 | * 输入：微博用户名，指定时间段，关键词 * 功能：自动爬取该用户的原创微博，并保存包含关键词的微博文本到指定文件 * 输出：微博用户名-时间段-关键词.csv文件 |
|  | 网页爬虫模块  HtmlSpider | 爬取指定网站的网页文本，实现关键词过滤，保存到指定csv文件 | * 输入：网址，指定时间段，关键词 * 功能：自动爬取该用户的原创微博，并保存包含关键词的微博文本到指定文件 * 输出：微博用户名-时间段-关键词.csv文件 |
| NLP自然语言处理 | 分词模块  WordSplitter | 用jieba分词模块对爬取的网页和微博文本进行初步分词处理 | * 输入：csv文件夹，过滤字典 * 功能：使用jieba分词组件对网页文本csv文件进行分词处理，自动过滤掉过滤字典中的词语 * 输出：关键词词典 |
| 知识图生成 | 关键词分类  WordCategory | 对关键词进行分类，识别 | * 输入：关键词词典 * 功能：使用NLP组件对关键词进行分类，识别出词性（名词、动词、语气词等等） * 输出：关键词分类 |
| 知识图生成 | 知识图谱生成算法模块  KG\_Algorithm | 根据文本和关键词构建知识图谱实体、关系、属性三元组 | * 输入：关键词分类，csv文件 * 功能：根据文本和关键词生成知识图谱三元组 * 输出：关键词及三元组关系 |
|  | 知识图数据库管理模块  KG\_Database | 用net4j封装生成知识图谱数据库管理模块 | * 输入：生成算法生成的三元组 * 功能：管理现有的知识图谱数据库，实现三元组增加、删除、修改、查询的功能 * 输出：net4j中存储图三元组 |
| 查询 | 程序图形界面  Client\_UI | 主窗口中包含了知识图谱查询和管理功能 |  |
|  | 知识图谱可视化界面 | 知识图谱可视化界面，比较美观、简洁的形式显示关键词查询结果 |  |

**四、程序设计**

**处理流程图：**

**用户界面（例子）：**

图片包含 屏幕截图

描述已自动生成

知识图谱可视化界面可使用gephi客户端生成比较炫酷的可视化呈现。

图片包含 屏幕截图, 电子产品

描述已自动生成

**后续扩展**：

微信公众号信息爬取、知识提取算法更新，使用机器学习来提高知识提取准确性。

**五、数据源**

|  |  |  |  |
| --- | --- | --- | --- |
| **编号** | **类别** | **信息来源** | **链接** |
| 1 | 微博 | 健康中国（国家卫健委） | <https://weibo.com/jiankangzhongguo?is_hot=1> |
| 2 | 微博 | 健康四川官微（四川省卫健委） | https://weibo.com/u/3754192351?is\_all=1 |
| 3 | 网页 | 湖北卫健委疫情发布网站 | http://wjw.hubei.gov.cn/fbjd/ |
| 4 | 网页 | 国家卫健委疫情发布 | http://www.nhc.gov.cn/xcs/yqfkdt/gzbd\_index.shtml |
|  | 微博 | 四川发布 | <https://weibo.com/sichuantianfu?is_hot=1> |
|  | 微博 | 广东发布 | <https://weibo.com/guangdongfabu?is_hot=1> |
|  | 微博 | 湖南省政府 | <https://weibo.com/hunangov> |
|  | 微博 | 湖北发布 | <https://weibo.com/hubeifabu?is_hot=1> |
|  | 微博 | 香港政府 | <https://weibo.com/newsgovhk> |
|  | 微博 | 黑龙江 | <https://weibo.com/hljfb?is_hot=1> |
|  | 微博 | 辽宁发布 | <https://weibo.com/u/5537781788?is_hot=1> |
|  | 微博 | 山西发布 | <https://weibo.com/shanxifabu?is_hot=1> |
|  | 微博 | 浙江发布 | <https://weibo.com/u/5131766197?is_hot=1> |
|  | 微博 | 江苏省 | <https://weibo.com/weibojiangsu?is_hot=1> |
|  | 微博 | 北京发布 | <https://weibo.com/bjfbt?is_hot=1> |
|  | 微博 | 天津发布 | <https://weibo.com/tjszfxwbfb?refer_flag=1005050006_&is_hot=1> |
|  | 微博 | 吉林发布 | <https://weibo.com/jlxwb> |
|  | 微博 | 内蒙古 | <https://weibo.com/neimenggu?is_hot=1> |
|  | 微博 | 山东 | <https://weibo.com/shandongfabu?is_hot=1> |
|  | 微博 | 上海 | <https://weibo.com/shanghaicity> |
|  | 微博 | 福建 | <https://weibo.com/u/5033508400> |
|  | 微博 | 安徽 | <https://weibo.com/ahfb> |
|  | 微博 | 江西 | <https://weibo.com/jxfbt?is_hot=1> |
|  | 微博 | 河南 | <https://weibo.com/hnrbwb?is_hot=1> |
|  | 微博 | 广西 | <https://weibo.com/gxrb2013?refer_flag=1001030103_&is_hot=1> |
|  | 微博 | 海南 | <https://weibo.com/u/5186902466?is_hot=1> |
|  | 微博 | 西藏 | <https://weibo.com/u/2620622835> |
|  | 微博 | 陕西 | <https://weibo.com/shaanxigov> |
|  | 微博 | 甘肃 | <https://weibo.com/u/1937187173> |
|  | 微博 | 青海 | <https://weibo.com/qinghaifabu> |
|  | 微博 | 宁夏 | <https://weibo.com/nxgat> |
|  | 微博 | 新疆 | <https://weibo.com/xjfb> |
|  | 微博 | 澳门 | <https://weibo.com/u/6190786689?is_hot=1> |
|  | 微博 | 台湾 | <https://weibo.com/u/1888725725> |
|  | 微博 | 香港 | <https://weibo.com/newsgovhk> |
|  |  |  |  |

[北京市卫生健康委员会](http://www.nhc.gov.cn/wjw/dfzfwz/201906/97fb8e5a7bcb4c68996024add52d9bbc.shtml)

[天津市卫生健康委员会](http://wsjs.tj.gov.cn/)

[河北省卫生健康委员会](http://wsjkw.hebei.gov.cn/)

[山西省卫生健康委员会](http://wjw.shanxi.gov.cn/)

[内蒙古自治区卫生健康委员会](http://wjw.nmg.gov.cn/)

[辽宁省卫生健康委员会](http://wsjk.ln.gov.cn/)

[吉林省卫生健康委员会](http://wsjkw.jl.gov.cn/)

[黑龙江省卫生健康委员](http://wsjkw.hlj.gov.cn/)

[上海市卫生健康委员会](http://wsjkw.sh.gov.cn/)

[江苏省卫生健康委员会](http://wjw.jiangsu.gov.cn/)

[浙江省卫生健康委员会](http://www.zjwjw.gov.cn/)

[安徽省卫生健康委员会](http://wjw.ah.gov.cn/)

[福建省卫生健康委员会](http://wjw.fujian.gov.cn/)

[江西省卫生健康委员会](http://www.jxhfpc.gov.cn/)

[山东省卫生健康委员会](http://wsjkw.shandong.gov.cn/)

[河南省卫生健康委员会](http://www.hnwsjsw.gov.cn/)

[湖北省卫生健康委员会](http://www.hbwsjs.gov.cn/)

[湖南省卫生健康委员会](http://wjw.hunan.gov.cn/)

[广东省卫生健康委员会](http://www.nhc.gov.cn/wjw/dfzfwz/201906/dea72b93576f469f8a42439b6f1890ef.shtml)

[广西壮族自治区卫生健康委员会](http://wsjkw.gxzf.gov.cn/)

[海南省卫生健康委员会](http://wst.hainan.gov.cn/swjw/index.html)

[重庆市卫生健康委员会](http://wsjkw.cq.gov.cn/)

[四川省卫生健康委员会](http://wsjkw.sc.gov.cn/)

[贵州省卫生健康委员会](http://www.nhc.gov.cn/wjw/dfzfwz/201906/b03d088c9a674b9ebcb8435f0f4239df.shtml)

[云南省卫生健康委员会](http://ynswsjkw.yn.gov.cn/)

[西藏自治区卫生健康委员会](http://www.nhc.gov.cn/wjw/dfzfwz/201906/2c5c21bc3ad847958d8c16069660f762.shtml)

[陕西省卫生健康委员会](http://sxwjw.shaanxi.gov.cn/)

[甘肃省卫生健康委员会](http://wsjk.gansu.gov.cn/)

[青海省卫生健康委员会](https://wsjkw.qinghai.gov.cn/)

[宁夏回族自治区卫生健康委员会](http://wsjkw.nx.gov.cn/)

[新疆维吾尔自治区卫生健康委员会](http://www.nhc.gov.cn/wjw/dfzfwz/201906/3282d1caa6494ac29ec748af0eaf1625.shtml)

六、版本控制

本项目使用github做为版本控制工具，多人协作开发模式：

**七、参考文档**

1、Python编码规范：<https://www.runoob.com/w3cnote/google-python-styleguide.html>

2、Python GUI编程（Tkinter）：<https://www.runoob.com/python/python-gui-tkinter.html>

3、什么是知识图谱：<https://www.jianshu.com/p/2a38bdcc4e3f>

4、知识图谱-用Python代码从文本中挖掘信息的强大数据科学技术：<https://baijiahao.baidu.com/s?id=1650236353203867652&wfr=spider&for=pc>

5、Python爬虫爬取微博用户信息及微博内容：

<https://blog.csdn.net/asher117/article/details/82793091>

6、Python爬虫爬取微博数据：<https://www.jianshu.com/p/c4ef31a0ea8c>

7、Python爬虫入门：<http://c.biancheng.net/view/2011.html>

8、Json教程：<https://www.runoob.com/json/json-tutorial.html>

9、Git版本管理，看这一篇就够了：<https://www.jianshu.com/p/0e9d07ec76f9>

10、python面向对象开发：<https://blog.csdn.net/weixin_44239490/article/details/86357989>

11、W3school python面向对象开发：<https://www.runoob.com/python3/python3-class.html>