研究设计

（一）数据来源及处理

首先利用Python编写爬虫程序，爬取人民网留言板网站和微博上关于新冠肺炎疫情的评论文本作为“第三方”的评论数据来源，共爬取了XXXXX数据。

**人民网留言板：**

**观测数据时间：2022年1月1日至今**

**观测数据对象：选取地级行政区领导人（市场、市委书记）的留言板作为爬取页面**

**限定主题相关性：“疫情”、“肺炎”、“病毒”、“新冠”、“口罩”等**

**微博：**

**观测数据时间：2022年1月1日至今**

**限定主题相关性：关于“疫情”、“肺炎”、“病毒”、“新冠”、“口罩”的评论文本**

然后对爬取的文本数据进行了如下处理：（1）首先使用pandas库进行数据的读取，对数据进行去重处理。（2）利用jieba库对每一条文本数据进行分词、自定义词典和停用词过滤等处理。（3）使用空格将分好词后的词语连接起来，再利用jieba和WordCloud库，生成第三方评论文本的词云图。

（二）研究方法使用python和R语言，利用情感词典对第三方的评论文本进行情感分类，并对主要的情感关键词进行聚类可视化；利用Python和jieba库提取评论文本的名词、动词和形容词，绘制拥有不同第三方情感人群整体描述的词云图，接着再只提取动词，绘制拥有不同第三方情感人群会采取何种行为的词云图；利用python以及gensim对第三方评论文本进行LDA主题模型分析.

（三）研究框架首先利用jieba库对评论文本进行分词、自定义词典等处理，利用python和R语言编写程序对评论文本进行情感分类；再对处理好的评论文本，提取名词、动词以及形容词来绘制每一类情感的词云图，从而在整体上描述拥有不同第三方情感人群有哪些特征。然后只提取动词，绘制每一类情感的词云图，从而可以反映出拥有不同第三方情感人群会采取哪些行为和动作。最后，利用python和gensim库，对全部评论文本作LDA主题模型分析，来研究第三方评论中主要关注的话题

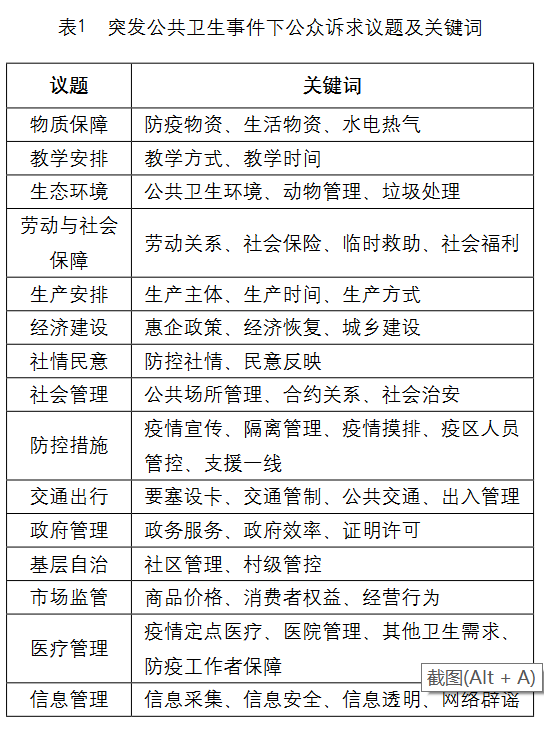
实证分析

1. 诉求内容和情感分类

首先，利用情感分析软件ROSTEA来对突发公共卫生事件中公民诉求内容进行情感分类，按照软件自带规则划分情感倾向（负面、中性、正面）以及情感强度（弱、中、强）

**得到以下表：**

**表1：公民诉求议题和关键词**

****

**表2：公民诉求议题分布**

****

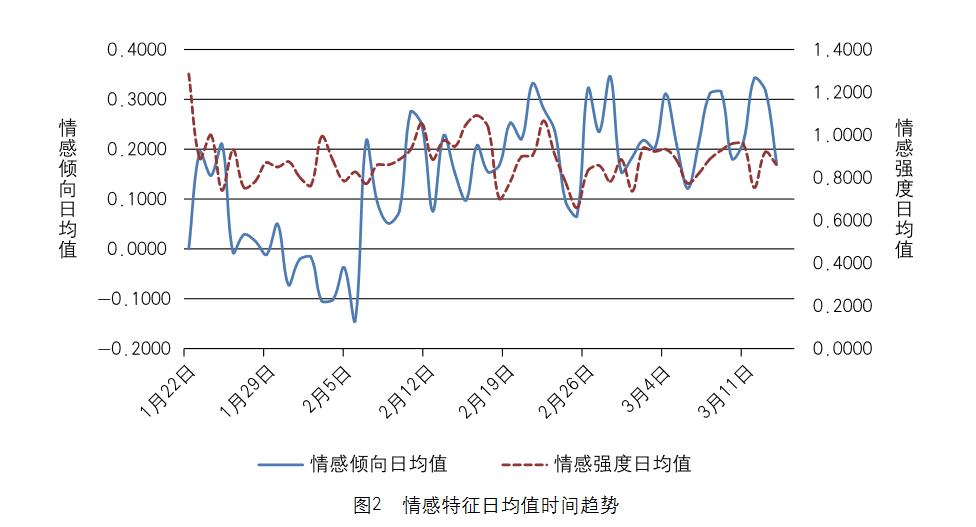
**表3：公民诉求表达特征**

****

**图1：情感聚类图**

****

**图2：情感特征日均值时间趋势**

****

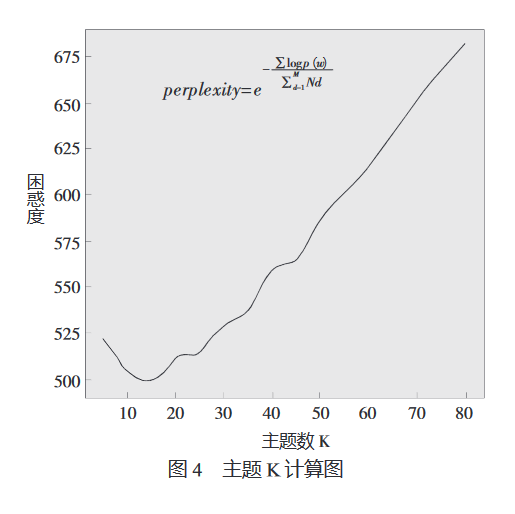
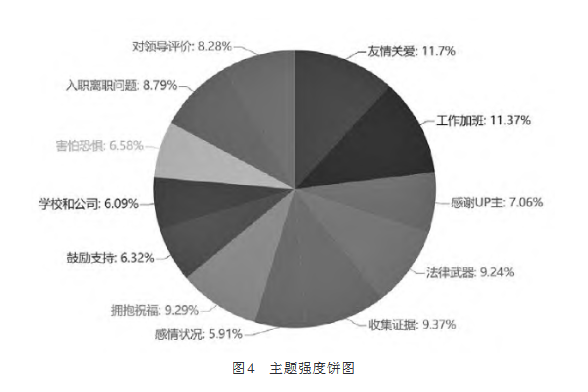
1. 词云图分析

本文运用jieba库，首先，对第三方评论文本进行分词、自定义词典和停用词等处理，根据TF-IDF原理对文本进行关键词提取，然后，再利用Python和jieba、wordcloud库分别对各类情感进行词云分析，用jieba库对评论文本进行分词和自定义词典等操作，然后将分好的词语用空格连接起来，最后利用wordcloud库生成词云图。词云图主要是为了突出文本中词频较高的关键词，词频越高的关键词，其字体显示越大。

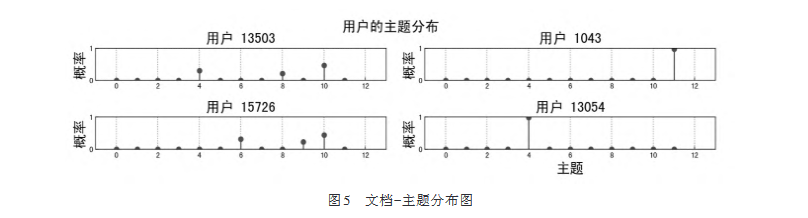
**得到以下图：**

**图1：公民诉求主题词云图**

****

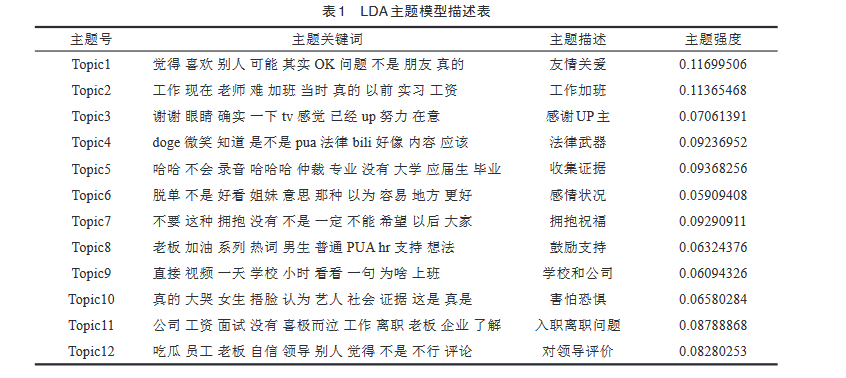
1. LDA主题模型分析
2. **主题数选取，通过K值调整，确定K值范围在（X,Y）范围中，语料库的困惑地较低，性能较好。  
   **
3. **主题强度饼图，根据各主题的强度绘制了主题强度饼状（数据有了的话，这个我自己做也可以）  
   **
4. **文档-主题分布可视化**

**过研究文档-主题分布的情况，能够得出突发公共卫生事件中第三方评论文本的主题划分情况，本文 从这些文本中随机抽取 10 篇文档，绘制出其文档-主题分布图（图5 展示了 4 篇文档的文档-主题分布 图，其余详见附件增强出版）。**

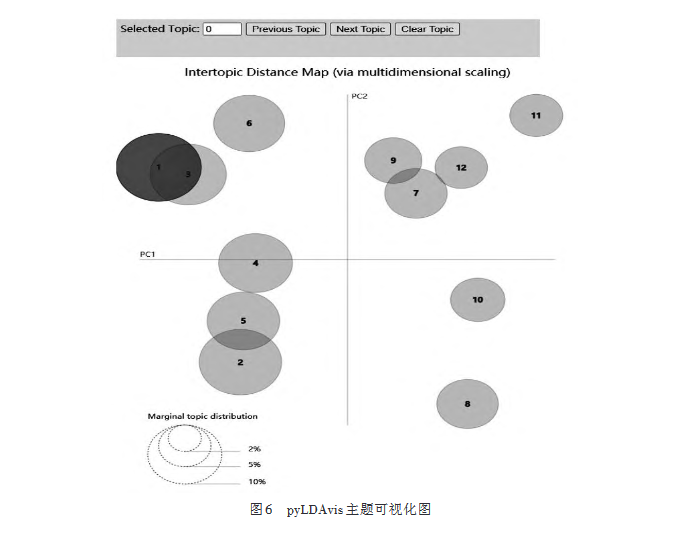
****

1. **各主题描述**

**本文对整个第三方评论文本数据进行 LDA 主题模型分析，运用Python 语言，gensim 和 pyLDAvis 库进 行主题生成和主题的可视化，本文根据主题一致性最终把文本内容生成12 个主题，如表1 所示，并根据主题强度作图 4。**

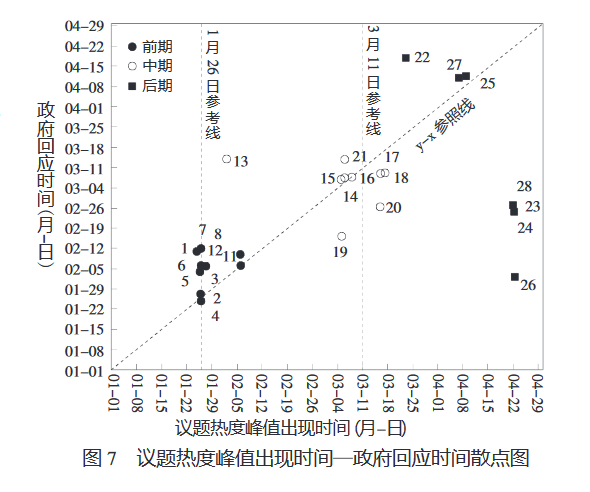
****

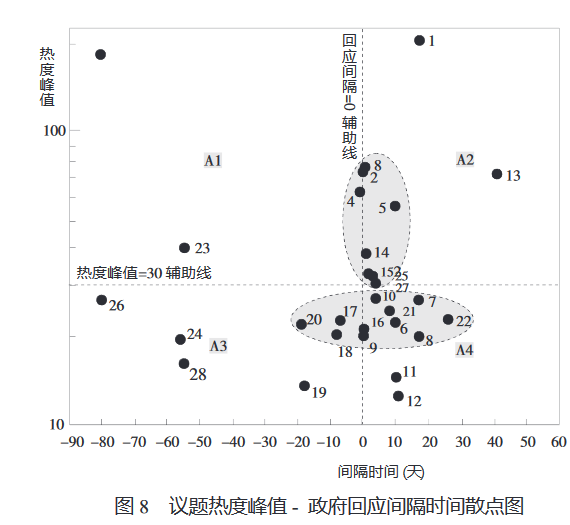
1. **热门主题可视化**

**以上文分析中主题强度较大的主题 1（友情关爱）、主题2（工作加班）、主题5（收集证据）为例，进行 主题可视化，运用pyLDAvis 库对 LDA 模型生成的 12 个主题进行可视化，如图6 所示（图6 仅展示主题 1 的可视化图，其余热门主题可视化图见附件增强）。图中的 12 个圆圈代表每一个主题，圆圈里标的数字 表示所对应主题的序号，圆圈的大小表示包含该主题文档的数目，圆圈越大表示包含该主题的文档数越 多。圆圈之间的远近代表主题之间相关性的大小 ，两个圆圈离得越近表示两个主题关联性越强。**

1. **主题-词分布**

****

1. **议题热度峰值出现时间-政府回应时间散点图  
   **
2. **议题热度峰值-政府回应间隔时间散点图**

****