Logo

Description automatically generated

**中山大学政治与公共事务管理学院**

**“政务关注”省级社会调研大赛之第四届公共治理数据分析大赛暨第二届城市管理案例创新大赛初赛方案**

**题目：**

**基于微博平台的老年人群体对于健康码使用现况的数据分析**

队伍编号：161

队伍名称：偶然组队

目 录

[一、选题背景 2](#_Toc54553219)

[二、文献综述 2](#_Toc54553220)

[三、研究问题 4](#_Toc54553221)

[四、研究意义 4](#_Toc54553222)

[1、理论意义： 4](#_Toc54553223)

[2、现实意义： 5](#_Toc54553224)

[五、.数据分析部分 6](#_Toc54553225)

[5.1关键词词云图 6](#_Toc54553226)

[5.1.1算法简析——基于matplotlib的词云图 6](#_Toc54553227)

[5.1.2词云图展示 7](#_Toc54553228)

[5.2 量化分析样本中的词频 7](#_Toc54553229)

[5.2.1 算法简述 7](#_Toc54553230)

[5.2.2词频统计结果 10](#_Toc54553231)

[5.2.3 针对词频的舆论分析 11](#_Toc54553232)

[5.3 LDA主题分类模型 12](#_Toc54553233)

[5.3.1 算法简析 12](#_Toc54553234)

[5.3.2 根据LDA模型得出的主题分析 13](#_Toc54553235)

[5.3.3 基于主题分类进行的舆情分析 14](#_Toc54553236)

[六、研究小结 14](#_Toc54553237)

[七、参考文献 15](#_Toc54553238)

# 一、选题背景

健康码作为数字时代社会治理在非常时期推出的第一大创举，为抗击疫情发挥了不可替代的作用，直接推动了社会和城市管理的现代化进程。然而健康码的传播过程也并非一帆风顺， 作为疫情期间的通行凭证，在保障了大多数人健康出行的同时，却遗漏了很大一部分没有智能手机、不会使用智能手机的老年人群体，这使得他们在必须使用健康码的场景当中面临困境，造成了这其中的“数字鸿沟”问题日益尖锐。

# 二、文献综述

关于数字鸿沟的来历说法众多，其中据Hoffman考证，数字鸿沟一词最早始于Lloyd Morrisett有关信息富人和信息穷人之间所存在的一种鸿沟认识。随着互联网日益广泛的使用，数字鸿沟成为一个笼统的标签或比喻，用来说明人们对信息传播技术，特别是互联网的采纳和使用上存在的差距（金兼斌，2003）。较早的研究是从美国国内开始的。从1995至2000年，美国国家电信与信息局以因特网普及率、个人电脑普及率和电话普及率作为衡量指标，探讨数字鸿沟的相关政策。其后，数字鸿沟的研究从关注信息化水平拓展到经济格局的变化，比如谈世中与赵丽红认为数字鸿沟必将带来最不发达国家的贫困化、边缘化（谈世中,赵丽，2003）。接着，不少学者将数字鸿沟纳入数字经济发展策略的要素之中，从地理经济学的视角，基于互联网，对数字鸿沟的空间表现和成因进行探究。自此，学界逐渐开拓公共政策与数字鸿沟的联系，根据收入水平、网络制度、工业化水平、知识水平、社会环境等维度对数字鸿沟问题进行分析，但大多侧重于数字化技术的应用和传播层面，鲜有涉及年龄群、不同社群、不同教育程度及性别的比较。

数字鸿沟虽然一开始以信息与技术间的主体差异出现，但实际上涉及政治、经济、社会、文化等诸多层面，在政治学、社会学、经济学、哲学等上有非常广阔的理论探讨空间，仅从技术层面的分析不足以显示出当今的社会问题。在我们所探讨的新冠肺炎疫情以及其他重大突发事件面前，面对具有“数字难民”色彩的老年人群体以及其他特殊人群则需要融合其他理论支持。目前我国也有不少学者开始将社群和阶层的概念引入到数字鸿沟的研究当中来，并在对于数字鸿沟的概念介绍时提到了人群之间对信息、网络技术应用程度不同造成的“信息落差“、”知识分割”和“贫富差化”，当下学术界正尝试用社会分化、社会排斥、社会不平等的角度去解释数字鸿沟的内涵。另外这个现象在公共管理中的影响也值得我们去研究和反思，以便于在信息化迅猛发展的时代背景下，为国家治理能力和治理体系现代化提供智力支持。

健康码背后所代表的依赖于大数据的数字治理是能够有效帮助城市、国家的一种资源。疫情期间以健康码为代表的数字抗议给全国的疫情的监控和管理都带去了便利，由地方政府最早开始使用后来在国务院统一协调下由阿里巴巴公司和腾讯公司共同参与到了开发运营当中，具有覆盖全员、全域的特点（方兴东、严峰，2020）。大数据作为国家治理现代化的科技型技术存在严重的“脱嵌”状况，现有的国家治理体系不能够完全适应信息技术的快速变迁（唐皇凤，2014）。随着现阶段疫情走向平缓期，健康码也走入了常态化，在健康码之前，并没有如此大规模的在全国范围内推行的数字化治理实操，同时中国大陆又是最早推出健康码的，因此缺乏可借鉴的相关经验经验。

# 三、研究问题

本文基于对微博平台上有关”老年人群体”与“健康码”相关的数据梳理及分析，旨在探索疫情期间中国大陆老年人健康码的使用现况，试图对老年人群体在健康码使用中面临的数字鸿沟问题进行初步解释性探究，为解决数字技术与公共传播在城市管理当中的冲突可能的新方向做出贡献。

# 四、研究意义

## 1、理论意义：

使老年人必须接受、适应、掌握健康码，但另一方面却没有给老年人以足够的“强制性”的保障——譬如宣传政策、指导使用等等。本研究将研究视野创新性地放到了非常时期的公共治理与数字资源建设当中，在信息化时代的潮流下，探讨了老年人群体和健康码使用之间的数字鸿沟问题。拓展了数字鸿沟主体的范围，弥补了针对老年人群体研究的空缺；从突发公共卫生事件为切入，在公共治理领域探究数字鸿沟问题，摆脱了以往过于关注数字鸿沟中的技术差异问题，同时能够为城市管理和发展提供现实指导意义。健康码使用中出现的“数字鸿沟”不仅是对于该现象的一种研究发展也是未来数字化城市管理当中不可避免的议题。

## 2、现实意义：

老年人在信息化社会如何体面地舒适地生活？他们的权利如何得到全面保障？这些都要求我们在城市政务管理当中付诸行动。反思在信息化迅猛发展的今日，本研究所关注的老年人所面临的数字鸿沟问题，能够帮助他们融入新的数字化环境、重新放心自在的出行。技术的本意是提升效率和质量，其根本落脚地是为人类服务，让生活更美好。因此老年人在健康码面前的数字鸿沟问题也警醒了我们，现阶段社会信息化越加发达，为何不仅没有给老年人的生活带去便利，更造成老年人群体与社会生活之间越加严重的疏离。社会的信息化发展与人文生活的关怀相融合任重道远。

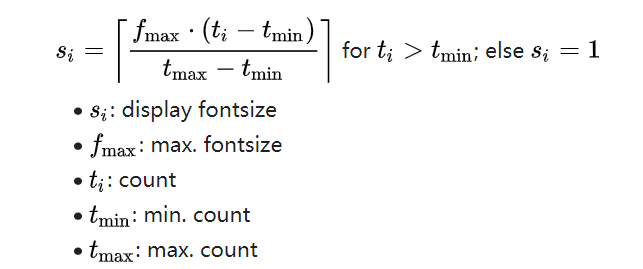
从“扫一扫”到“健康码”再到“文明码”，数字技术与公共传播的冲突始终存在，如何平衡技术发展提高效率和以人为本，需要更多的探索和思考。健康码作为数字科技时代智能传播的最佳出行，为未来城市管理的发展展现了新的图景。此次健康妈的使用就是一个宝贵的案例，我们在关注它与弱势群体之间关系的同时也势必能够为重新审视城市管理中的数字资源建设做出有益的思考。本研究也希望可以通过对“老年人与健康码”问题的研究为累积政府部门处理突发重大事件的经验、提升公共服务能力和管理能力以及确保公共服务质量和公众满意度做出努力。

# 五、.数据分析部分

## 5.1关键词词云图

### 5.1.1算法简析——基于matplotlib的词云图

词云中标签的字体大小取决于其出现率。对于诸如网志类的词云，频率例如对应于分配给该类别的网志条目的数量。对于较小的频率，可以直接指定字体大小，从一个到最大字体大小。对于较大的值，则应进行缩放。在线性归一化中，权重ti描述符的映射到1到f的大小比例，其中tmin 和 tmax 用于指定可用权重的范围。



由于每个描述符索引的项的数目是根据通常分布幂律，对于较大的值的范围中，对数表示是有意义的。词云的实现还包括文本解析和过滤出无用的标签，例如常用词，数字和标点符号。

在本次项目中，我们认为词云图对于文本数据的密集分布，能够给人更加直观的感受。

### 5.1.2词云图展示



图4.1.2：关键字“健康码”的词云图展示

## 5.2 量化分析样本中的词频

### 5.2.1 算法简述

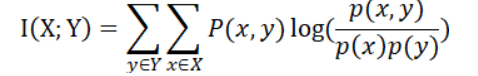
1）命名实体识别问题方法

命名实体识别（NE）问题指识别文本中具有特定意义的实体，其任务在于识别出待处理文本中三大类（实体类、实践类和数字类）、七小类（人名、机构名、地名、时间、日期、货币和百分比）。当前处理自然语言的主流是基于大规模语料库的统计方法，而根据使用的机器学习方法的不同，可以将基于大规模语料库的统计方法划分为以下四种：

2）未登录词的解决方法(搜索引擎，基于语料)

汉语中存在很多没有被收在分词词表中但必须切分出来的词，包括各类专有名词、新词等等，从而会产生未登录词的难题。各类专有名词由于命名实体识别的广泛使用，问题已经较好的得到解决。

对于新词，我们采用的是N-Gram+凝聚度+自由度的方法。首先将文本结果处理，去掉文本中各种无用的符号，然后通过N—Gram分词后统计的词频和频率来找到词的边界，N—Gram即交叉切分算法，指拆分成n个字符算一个关键字，n元即n个汉字来组成一个关键词，设置好最长字串的长度，利用N—Gram得到拆分好的字串字典，并统计字串的词频以及频率。

然后通过凝聚度、自由度来判断该字串是不是一个新词。凝聚度用互信息来表示，通过计算平均互信息来表示不同字典的相关程度，I越大则x、y的相关性即凝聚性就越大。而对于自由度，我们使用信息熵来构建文本片段的左右熵，其中熵越大，证明该片段与所有字段关系的不稳定性越高，越有可能作为独立的片段来使用。通过提前设定好词语长度的最小值、凝聚度以及自由度的阙值，对分好的字串字典进行过滤，输出结果，便可得到新词。

3）CRF解决命名实体识别（NER）流程总结：

训练阶段：确定特征模板，不同场景（人名，地名等）所使用的特征模板不同，对现有语料进行分词，在分词结果基础上进行词性标注（可能手工），NER对应的标注问题是基于词的，然后训练CRF模型，得到对应权值参数值

识别过程：将待识别文档分词，然后送入CRF模型进行识别计算（维特比算法），得到标注序列，然后根据标注划分出命名实体

### 5.2.2词频统计结果



图1：各词语在“健康码”关键字的3000份微博样本中出现的频率



图2：各词语在“老人”+“健康码”关键字的1500份微博样本中出现的频率

### 5.2.3 针对词频的舆论分析

由上图可以看出，在以“老人”与“健康码”为关键字的微博样本中，大爷、徒步、坐火车、回老家、打工、检测、部门等词是高频词汇。其中，有些是单一新闻引起强烈反响，从而引发讨论热潮；有些则是零散信息汇总在一起时，大众普遍关注的一些要点。这共同反映了舆论在相关事件上的注意焦点。下面是对部分高频词汇的原因分析：

“大爷”：新闻出自“#无健康码徒步千里大爷已回老家# 一位来自安徽亳州的大爷徒步千里走到浙江打工的视频，牵动着万千网友的心...”这则报道。由于没有手机无法出示证明自己健康状况的健康码，因此大爷无法乘坐任何公共交通工具，只能选择通过步行的方式由安徽前往浙江打工，半个月徒步千里每日露宿公园。这一报道深刻地反映了健康码系统的运行过程中，我国缺少对于一些信息技术资源匮乏的人群在政策上的合理照顾和疏导。

“打工”：健康码作为监测和控制疫情的数字化治理方式，在限制人口流动的同时，也对外来务工人群的就业问题造成了阻碍。如果健康码的运作出了问题，比如出现误判等等，很可能就造成结构性的失业情况，对我国的经济形势将是一个不小的考验。

“回老家”：按照中国传统，回家就是春节时期的要务。但是今年适逢新冠疫情的爆发，在火车站因为健康码而无法乘车的归乡者数不胜数，这也体现了个体在特殊形势下的困境。

## 5.3 LDA主题分类模型

### 5.3.1 算法简析

在本次项目中，我们采用词袋模型，假设了文档之间相互独立，文档中的词汇之间相互独立。假设我们的词典中一共有 V 个词 v1，v2.. ，那么最基础的模型就是认为是按照如下的游戏规则产生文本的。

a. 只有一个骰子，这个骰子有V面，每个面对应一个词，各个面的概率不一；

b. 每抛掷一次骰子，抛出的面就对应的产生一个词；如果一篇文档中N个词，就独立的抛掷n次骰子产生n个词；

c. 在 LDA 模型中一篇文档生成的方式如下:

1) 从狄利克雷分布中取样生成文档 i 的主题分布 *θi*

2) 从主题的多项式分布中取样生成文档 i 第 j 个词的主题 *Zi,j*

3) 从狄利克雷分布中取样生成主题的词语分布 *φi,j*

4) 从词语的多项式分布 *φi,j* 中采样最终生成词语 *wi,j*

### 5.3.2 根据LDA模型得出的主题分析



图4.3.2 微博上相关样本中各主题的归类

### 5.3.3 基于主题分类进行的舆情分析

由上图可知，在微博搜索关键词“健康码”可得到文本的主流观点当中“健康”、“防控”、“防疫”占比最高（79%）。

由此可知，微博文本倾向于关注宏观意义上的中国新冠疫情的处置，和中国卫生健康的状况。这反映微博网民和媒体密切关注新冠肺炎的最新近况，这可能是由于网民更在意且希望国家表现出处置疫情的强硬手腕和坚定的维护秩序的立场所导致的，这些可给国民增加民族自信心和底气。

此外，微博文本还倾向于“老人”，“肺炎”，“服务”主题（59%），这说明了健康码对于那些不习惯电子事务的群体，尤其是老年人群体来说造成了不小的困扰。另一方面，也体现出我国在大政方针执行之下的覆盖面还不够广，在抗疫大背景下的具体细节还有很多待完善的空间。

# 六、研究小结

本研究以老年人群体作为主体，定位于疫情期间公共管理领域，研究社会中存在数字鸿沟问题，基于对微博平台上有关”老年人群体”与“健康码”相关的数据梳理及分析，探索疫情期间中国大陆老年人健康码的使用现况。在数据分析中首先以词云图的方式对问题进行直观的体认，接着分别以词频和主题类型进行微观和宏观层面的趋势和特点分析，以其对老年人群体在健康码使用中面临的数字鸿沟问题进行初步解释性探究，描画出数字鸿沟与老年人群体间的冲突，并从数字技术与公共管理的融合中寻找解决方案。

# 七、参考文献

[1] 谈世中,赵丽红.“数字鸿沟”与最不发达国家的贫困化、边缘化[J]. 求是, 2003(11):57-59.

[2] 王亦斌. 中国地区数字鸿沟的成因和影响：1997-2007[D].浙江大学,2011.

[3] 方兴东, 严峰. “健康码”背后的数字社会治理挑战研究[J].人民论坛·学术前沿，2020:1-14.

[4] 金兼斌. 数字鸿沟的概念辨析 [J]. 新闻与传播研究.2003 (01): 75-79.

[5] 唐斯斯,刘叶婷. 以“数据治理”推动政府治理创新 [J]. 中国发展观察.2014 (05):32-34.

[6] 唐皇凤,陶建武. 大数据时代的中国国家治理能力建设[J]. 探索与争鸣, 2014(10):54-58