电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

《综合课程设计》报告

2019-2020-2

BACHELOR THESIS

论文题目：《基于微博社交媒体数据的疫情后复工复产及消费热点研究》

学　　院：公共管理学院

专　　业：信息管理与信息系统专业

时　　间：**2020.7.3**

小组成员及分工表

学号	姓名	主要工作
2017120101006	李云开	频繁词集聚类实现
2017120101018	何治健	**K-means** 聚类实现
2017120101026	韩笑	**K-means** 聚类实现
2017120101021	黄淳祯	频繁词集聚类实现

摘　要

随着互联网社交媒体的不断发展，微博逐渐成为研究者就舆论中心与讨论话题的主要研究对象，其中基于微博文本聚类与主题挖掘开始成为新的研究趋势。本文就 2020 年的初期在微博社交媒体数据的疫情后复工及消费热点进行研究。针对文本聚类与 LDA 主题模型的互补特征，综合考虑了中文微博的特殊文本与短文本聚类问题，提出了基于 kmeans 和预设主题的文本聚类和基于类簇和主题的 LDA 主题挖掘相融合的微博检索方法。实证结果表明本文中提出的两种方法不仅能有效地划分微博文本，并且能清晰地挖掘类簇中潜在主题。

关键词： 文本聚类；主题建模；消费；复工复产

ABSTRACT

With the continuous development of Internet social media, Weibo has gradually become the main research object of researchers on the center of public opinion and discussion topics. Among them, text clustering and topic mining based on Weibo have become a new research trend. This article studies the factors affecting the resumption of work and consumption after the epidemic of Weibo social media data in early 2020. Aiming at the complementary features of text clustering and LDA topic model, considering the special text and short text clustering problem of Chinese Weibo, text clustering based on kmeans and preset topics and LDA topic mining based on clusters and topics are proposed Integrated microblog retrieval method. The empirical results show that the two methods proposed in this paper can not only effectively divide the Weibo text, but also clearly excavate the potential topics in the cluster.

Keywords: **text clustering; topic modeling; consumption； work resumptio**

目录

1. 绪论

1.1 问题研究的背景

随着计算机与互联网技术的不断发展，社交媒体以各种形式迅速发展并崛起，其中微博作为一个基于用户关系的信息分享、传播以及获取信息的平台，占据了中国微博用户总量的 57%，以及中国微博活动总量的 87%，具有典型代表性。微博作为一个社交平台，用户既可以在微博上浏览感兴趣的信息;也可以作为发布者,在微博上发布内容供别人浏览。发布的内容一般较短。具有便捷性与传播性的特点。用户发布话题信息的吸引力、新闻性越强，对该话题感兴趣、关注该话题的人数也越多，影响力越大，由此兼具实时性、现场感以及快捷性的特点。

在突如其来的 2020 新冠疫情中，微博广场效应显赫，舆论导向明显，舆情风向变化迅速，是大众发声的第一选择。微博作为一个意见广场，从疫情发酵期开始，转变成为了向公众传递与疫情相关的重要信息,也让最新的疫情与复工消费信息得以迅速形成话题并快速传播。微博作为疫情后的复工复产与消费复苏的舆情中心，影响并传达着大众在疫情后复工与消费的主要讨论内容。但由于微博的搜索采取的方式与传统的关键词搜索方式区别不大，且微博内容本身具有文本短小特征，属于短文本范畴，具有半结构化特点，关键词出现的次数很少但在文本中占比较大，仅仅依靠关键词的比较无法获得较好的检索效果。微博作为新一代的社交媒体，尤其是话题性微博，具有主题密度高，即时性强，具有官方媒体发文和个人社交媒体发文两种主要发文渠道。其中：1. 个人社交媒体发文具有：情感表达强烈而理性评价淡化，口语色彩浓重，观点表达相对隐晦，评价对象省略和非规范性的特点。但所占发文比例最大，文本挖掘与主题分析难度也最大；2. 官方媒体发文具有：着重理性评价，观点和主题较为清晰明确，发文具有规范性。是文本挖掘和主题分析的主要组成部分与微博话题的主要倡导者之一。

微博的主题具有蒲公英式的传播特点，具有随机性，爆炸性，即时性。一个微博主（个人/官博）会发出一个源微博，形成一个大致的讨论话题/主题，其余微博会围绕该主题发表意见，下一轮微博会继续围绕上一轮的微博继续发表对该主题的意见，形成一个具有鲜明主题的微博群，该微博群的大小由网民们对该主题的关注程度确定。

围绕复工复产与消费两个主题，伴随微博蒲公英式的传播特点，每个主题下可能会生成围绕经济，娱乐，政策等多板块跨区域性的话题（topic）。且每个主题与其包含的多个话题由于微博本身特性的存在，致使传统的文本挖掘方式难以致用于微博文本。这就导致了对于微博这种具备特殊特征的文体,传统的搜索与挖掘方式不足以满足对于主题分析的需求，需要选取更为有效的方法对微博信息进行检索与进一步的分析。

1.2 研究意义

对突发事件报道的研究是舆情研究中的重要领域，而疫情、公共卫生事件对政府公信力的影响更是高于普通突发事件，因此相关研究更应受到重视。疫情报道、公共卫生事件报道的研究通常归于突发事件报道研究的范围内。通过对微博文本的主题挖掘与分析，能够帮助政府与研究者更好的确定舆情风向与话题中心，具有典型代表性。通过对微博文本

就疫情后复工及消费热点研究可以更优地研究我国经济在疫情后复苏情况下对大众影响程度较大以及大众所关心的主要因素与舆情状况,方便政府与决策部门更好的制定相关政策与规定来有效听取舆情并做好复工与刺激消费的相关准备。

1.3 文献综述

从早期的文本挖掘研究来看，我国基于文本挖掘的主题研究相对较晚。通过高级数据挖掘和自然语言处理，对非结构化的文字进行机器学习的手法，可以使得从文本中挖掘更具有价值的线索和信息。文本数据挖掘包含但不局限以下几点：主题挖掘、文本分类、文本聚类、语义库的搭建。其中在机器学习和自然语言处理等领域，主题挖掘的主要目的是寻找主题模型

基于当前研究的背景分析可知：首先，文件数量迅速增长，无法仅依靠人工的方式实现对全部文本信息高效阅读和理解，将该流程自动化已经势在必行。其次，主题挖掘可以提高文字重度依赖应用的使用效率和产出影响，比如搜索加索引、文本总结、聚类、分类和情感分析。

从大量文字中找到主题是一个高度复杂的工作，不仅因为人的自然语言具有多层面特性，而且很难找到准确体现资料核心思想的词语。借助主题模型这一无监督技术，用来发现各种文本文档中的主题。这些主题本质上是抽象的，即彼此相关的单词会形成主题。主题模型可以对海量的文本数据进行探索，对词组进行聚类，找到文本之间的相似性，并发现抽象的主题,可以有效查找在社交媒体上基于舆情表达的消费热点并进行有效提取与归纳。

国内对在疫情下的消费与复工的研究较为突出，其中包括从当前对于疫情后的消费研究来看[1]，疫情已造成相当部分消费群体消费能力阶段性下降，整体消费更趋保守。在经历疫情后，消费者会采取更加理性和保守的姿态应对未来的诸多不确定性因素。另外，经济增速放缓会带来居民收入降低，也会让部分顾客在消费环节中压缩支出成本，消费者居家生活的基本需要和减少外出加强风险防范的疫情限制[2]。随着疫情缓解、复工复业、复产复市的展开以及各类人员返岗就业的增加和之前积压的消费需求,逐步开始恢复和活跃起来。并且更多涉及新一代信息技术对生产生活的融合改造。在对疫情后居民消费习惯的研究[3]中表示疫情会通过增强居民的风险感知、恐惧感和无聊感，提升他们疫情期间的从众性、冲动性和稀缺性消费倾向，除此之外，居民的物质主义倾向和共同居住人数会放大疫情对居民非理性消费的影响。在关于疫情后的复工复产的研究方面[4]，企业的复工决策是风险偏好型,政府的助推政策和企业复工参照点的动态变化等因素都很大程度上会影响企业复工决策的顺利实施。同时还需考虑保障基本生活需要、保障疫情防治、保障就业稳定和保障发展质量[5]。在复工过程中，存在企业收入骤降，复工达产难度高；供应链衔接不畅，融资难融资贵问题加剧；内需紧缩外需中断，行业下行压力加大等问题[6]。但是相对来说，基于社交媒体对消费与疫情的主题研究相对较少，因为其中消费与复工复产的相关民意都会着重表达于网络舆情中。同时,官方媒体也会通过网络渠道来对舆情进行引导，形成蒲公英式的传播效应与相关主题群。因此，针对基于微博社交媒体的文本的热点与主题挖掘就显得格外重要。同时，由于相关研究较少，仍需对主题挖掘方法进行拓展与改进。

首先是由于当前研究的主流热点使用的方法也是基于传统的 TF-IDF 结合 LDA 主题模型的算法与手段，从中对传统短文本进行主题模型细分。其次从研究方式来看，纵观国内外现有的研究成果，理论研究普遍多于实证研究。当前多采用的有通过聚类分析将隐藏状态序列与知识流动模式一一对应的 LDA-HMM 方法[7]；有基于 LDA 与 Self-Attention 的短文本情感分类方法[8]；以及使用融合词频-逆文本频率（term frequency-inverse

document frequency, TF-IDF）和 LDA 的中文 FastText 短文本分类方法[9]；另一方面，由于对突发事件报道的研究是舆情研究中的重要领域，而舆情与经济社会对疫情、公共卫生事件的关注度更是高于普通突发事件。因此更需要创新式的方法对疫情下的消费热点进行提取。本文借鉴当前广泛使用的 LDA 模型，提出了基于预设主题的聚类手段对研究思路进行创新，总结提取出五大疫情下的消费热点以及各自的子热点，对基于消费领域的微博文本进行主题刻画。

整体来说，在经典理论的基础上，本文从原有的主题模型挖掘的研究基础上摸索出了一个具有较好聚类与挖掘效果的模型。但目前研究仍存在不足，在对预设主题上的选择与阈值设置上仍缺乏完整的可约束性，有待于不断补充。相关研究方面，极少有针对微博半结构化短文本下的疫情下消费热点的主题挖掘，在该领域仍等待新一步的研究与实证过程。本文将结合创新式的主题模型，对当前主流社交媒体上的舆情信息中进一步挖掘所研究的消费热点问题并进行实证分析，对于挖掘的主题结果也会进行进一步的解释与总结。

1.4 相关理论基础

1.4.1 频繁词集与支持度的相关定义

定义 1 (支持度)：给定文本集合 D，词集 X 的支持度为同时包含 X 中所有词语的文本数量，其中 X 表示文档中的一系列词，记为 sup (X) 。

定义 2 (频繁词集)：设一个文本数据库 TDB 由一系列的文档构成，每个文档表示为＜tid，X＞ 。其中 tid 为文档的唯一 ID。设 T = {t1，t2，…，tn} 为词的集合，Y 是 T 的第一个非空子集。如果 sup (Y) ≥α，则称 Y 为一个频繁词集，其中α为一个阈值(最小支持度) 。如果对于 Y 的任意超集 Z，均有 sup (Z)＜α，则称 Y 为集合 D 上的最大频繁词集。

1.4.2 雅卡尔（Jaccard）系数：

雅卡尔系数（英语：Jaccard index），又称为并交比（Intersection over Union）、雅卡尔相似系数（Jaccard similarity coefficient），是用于比较样本集的相似性与多样性的统计量。雅卡尔系数能够量度有限样本集合的相似度，其定义为两个集合交集大小与并集大小之间的比例：

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

雅卡尔距离（Jaccard distance）则用于量度两有限集合之间的不相似度，其定义为 1 减去雅卡尔系数。如果 A 与 B 完全重合，则定义 J(A,B) = 1。于是有：

$$0 \leq J(A,B) \leq 1.$$

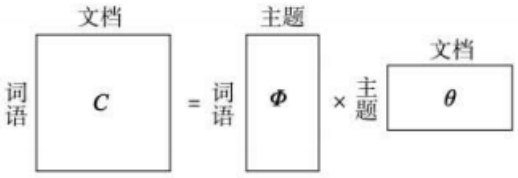
雅卡尔距离（Jaccard distance）则用于量度样本集之间的不相似度，其定义为 1 减去雅卡尔系数，即：

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}.$$

雅卡尔距离是所有有限样本集合间的度量。常应用于：比较文本相似度，用于文本查重与去重；计算对象间距离，用于数据聚类等。

1.4.3 LDA 主题模型

隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）主题模型是近年来主流的概率主题模型，在文本的主题挖掘领域应用广泛。主题是一个抽象的概念，一篇文档的主题潜藏在文档词语中，通过对文档词语的分布规律进行分析，将分布规律具有相似性的词语划分为一个簇，从每个簇中就可以挖掘文本潜在的主题，这就是概率主题模型。随着对概率主题的不断研究，2003 年，Blei 等人提出 LDA 概率主题模型，随后该模型在文本挖掘领域得到广泛应用。LDA 的主要思想是基于文档、主题和词语三个层面所产生的，每篇文档都蕴含有若干权重不同的潜在主题，而每个潜在主题都可表示为文档构成的词语池中部分词语的概率分布。这样就将文档与词语之间的关系转变为文档与主题、主题与词语之间的关系，而文档与词向量的空间矩阵就可以映射为文档与主题的矩阵和主题与词语的矩阵，而 LDA 的目的即是解析出文档与主题的矩阵和主题与词语的矩阵。LDA 模型中文档、主题、词语矩阵转换的关系如图 1 所示。词袋假设认为文档中的词是独立的，文档即是词语的无序组合，而 LDA 模型便是建立在此假设之上的。相比于注重语义语法的文本分析模型，微博文本篇幅较为短小，语法较简单且存在相当多的不规范情况，LDA 主题分析模型在处理微博短文本数据时具有一定优势。



2. 研究思路及研究方法

2.1 研究思路

本文希望能通过对微博文本数据的分析来挖掘疫情背景下民众对于复工复产和消费方面所关注的话题。在本研究中选用在文本挖掘及其相关领域具有广泛应用的 LDA 主题模型来挖掘微博文本下的潜在主题。微博文本具有数量大、篇幅短、单篇价值低的特点，为了让 LDA 分析后得出的主题下的词语更加相似，优化概率分布的结果，提出基于文本聚类和 LDA 主题模型相融合的思路，希望通过文本聚类将原微博文档划分为若干簇，再对每个簇调用 LDA 算法来获得潜在主题。该模型的示意图如图 2 图所示：

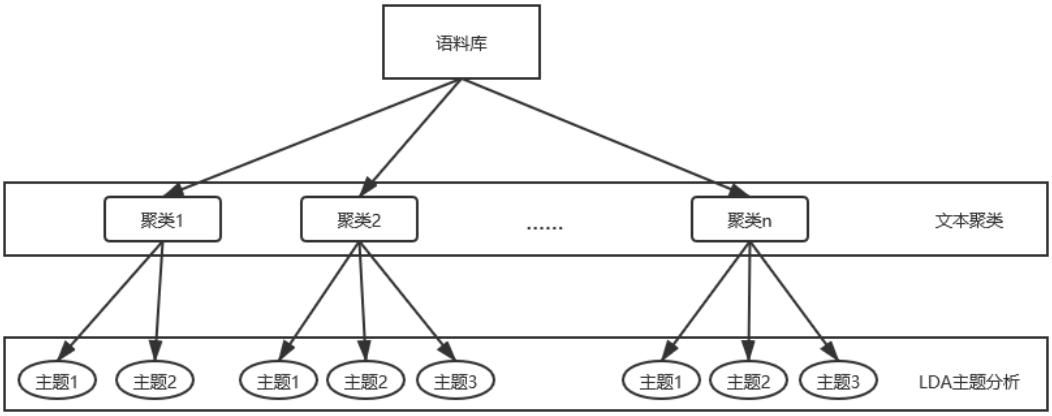


图 2 基于文本聚类和 LDA 相融合的模型

2.2 研究方法

2.2.1 基于 K-Means 算法的文本聚类

文本聚类在主题挖掘领域占据重要作用，其可以在短时间内根据内容将大量文本划分到若干有意义的簇中，方便对文本数据的主题进行挖掘。文本聚类主要是依据著名的聚类假设：同类的文档相似度大，而不同类的文档相似度小。中文文本聚类的常用策略为先将文档进行分词处理，再将处理后的文档转换为向量，最后利用聚类算法对向量矩阵进行聚类，所得结果既为若干个簇，且每个文档都会被划分到一个簇中。

在中文文本聚类中，文档的向量化是重要的一步。2013 年，Mikolov 等提出“Word2Vec”的向量表示方法，2014 年又提出“Doc2Vec”方法。相比于独热编码，Word2Vec 模型可以将高维词向量嵌入到低维空间中，其作为简化的神经网络，在训练时会考虑词语的上下文，通过上下文的词向量来预测目标词的词向量，而 Doc2Vec 则是在 Word2Vec 的基础上，再加上句子向量。Doc2Vec 还会考虑“上下文”，即词语、句子、段落中的语义信息。本文采用 Doc2Vec 模型，发挥其能较好的联系上下文中的语义信息的优势，实现对文本信息的向量化表示。

K-Means 是典型的划分型聚类算法。其算法是在数据点所在空间中选取 n 个聚类中心，每个聚类中心代表一个簇，计算每个数据点与 n 个聚类中心的距离，将数据点划到距离最近的簇中，划分完成后将计算每个簇中数据点位置的均值，将其作为新的聚类中心，重复进行以上过程知道簇中结果不再发生变化，所得各簇即为算法输出的聚类结果。K-Means 算法容易得到局部最优解且受初始聚类中心影响较大，但其内容相对简单高效，在文本聚类中有着广泛的应用。本文中将被使用 K-Means 聚类算法划分的数据点即为上文中所提到的经过 Doc2Vec 模型处理后的文档向量。

2.2.2 基于频繁词集的文本聚类

首先对每个微博文本使用进行分词处理，对应语料库去除无实际意义的虚词和代词，得到每个微博文本的词集。随后使用 FP_growth 对每个微博文本的分词集进行处理。

FP_Growth 算法的步骤如下： 1）第一次扫描数据库，寻找频繁 1-项集，并按照由大到小的顺序排序。2）创建 FP 模式树的根结点，记为“null”。3）根据频繁 1-项集的顺序对数据库中的每条事务数据进行排序，并存储在 FP 模式树中，并建立项头表。4）为每一个频繁 1-项集寻找前缀路径，组成条件模式基，并建立条件 FP 树。5）递归挖掘条件 FP 树，获得频繁项集。

在得到 FP_Growth 挖掘得到的频繁项集后，围绕 FP 树根节点下的一级子节点形成若干个簇。对每个簇设定参数 α，从每个簇中节点计数 count > α 的节点中人工挑选节点词作为预选主题词，每个主题都包含经人工筛选后的若干个词，对每个簇进行挖掘并生成一个主题直至所有簇共形成若干个主题。

随后基于我们的预设主题再对所有预处理后的微博文本词集进行预处理，使用雅卡尔系数（Jaccard index）比较每个微博分词集与每个主题词集的相似度。当微博文本的分词集 A 与主题词集 B 完全相同时，J（A,B）=1。通过 jaccrd 系数将每个微博文本的分词集合放入与其距离最近的主题之内，随后再在每个主题内对其聚类后的微博分词集合们进行 LDA 处理。从基于预设主题簇的角度对微博信息进行主题建模，可以加强聚类后领域特性主题的区分度。每个主题下 LDA 处理得到的 Topic，在 Topic 与主题簇之间和 Topic 与 Topic 之间存在更明确的包含信息与逻辑关系，不同主题簇间的 Topic 区分度更大，彼此间互通逻辑性也越小。通过人工筛选的预设主题可以进一步增强 LDA 对 topic 的挖掘与聚类结果，更好的将复杂信息层次化的显示出来，实现对微博文本在复工与消费层面的主题挖掘。

3. 实证研究

3.1 数据收集与清洗

本文的样本来自新浪微博。使用新浪微博的高级搜索功能，可以得到特定时间内特定关键词下的热门微博。热门微博相比普通微博发布者更具权威，关注者也更多，同时紧跟热点话题，具有更强的可靠性。因此，我们的数据采集策略便是使用 Python 爬虫定位、获取特定时间段内特定关键词下的热门微博。

具体过程上，利用了 Python 的爬虫常用包 request 和 BeautifulSoup，前者用来获取请求的网页信息，后者用来对 HTML 代码进行解析、处理。整个采集分两步，第一步是采集每个关键词下每一天的全部热门微博页链接，即通过构造微博 URL 链接找到特定日期特定关键词下的热门微博的第一页，根据第一页的页码获取当天全部热门微博页的链接。第二步是根据收集到的页面链接依次获取页面信息，从中提取本研究需要的微博内容。

最终，我们收集了 2020 年 2 月 1 日到 2020 年 4 月 30 日“复产”、“复工”、“消费”三个关键词下的全部热门微博，其中“复产”关键词下收集到微博 26940 条；“复工”关键词下收集到微博 53001 条；“消费”关键词下收集到微博 22814 条。

在对数据进行清洗前，我们首先对“复产”和“复工”微博进行了合并，方法是取二者的并集，对二者重复的部分进行去除，这样比直接以“复产复工”为关键词进行搜索获取到的微博数量更多、也更加全面。然后对“消费”数据集和合并好的“复产复工”数据集进行清洗，去除部分错误微博，最终得到“复产复工”微博 61013 条，“消费”微博 22777 条，其时间分布图 3 所示：

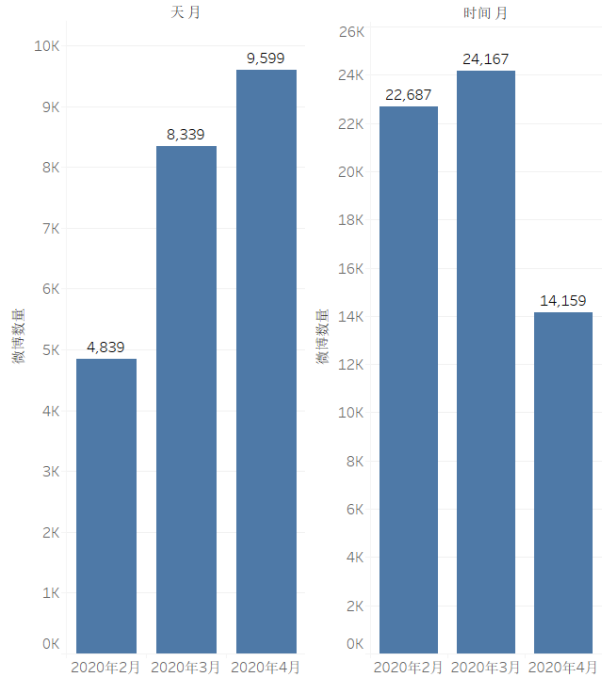


图 3“消费”微博数量时间分布（左）和
“复产复工”微博数量时间分布（右）

使用新浪微博高级搜索功能单次获取的微博数量有 50 页（约 1000 条）的数量上限，但由于我们在采集的过程中，一次只搜索一天内特定关键词下的微博数量，所以没有出现单次高级搜索微博数量到达上限的情况（虽然理论上有这种可能，但实际操作中没有遇到）。这对我们有两方面的意义，一方面，这保障了我们收集到的是特定时间内特定关键词下热门微博的全集，不存在某一天热门微博过多而未采集完整的情况；另一方面，我们可以根据不同时间段内特定关键词下微博数量来估计该关键词的时间热度变化。如图 3 所示，针对“消费”热门微博，可以看到其数量在 2020 年 2 月——4 月呈现出持续走高的趋势，这反映了随疫情好转大众对消费的讨论热情逐渐增强。针对“复产复工”热门微博，可以看到其数量在 3 月达到峰值，2 月其次，4 月最少，这同样不难理解，2 月份疫情比较严重，但关于“复产复工”的话题讨论仍然具有很高热度，而随 3 月份疫情逐渐得到控制，“复产复工”讨论的热度达到了顶峰，4 月份“复产复工”已经稳步实现，热度也就开始下降了。

3.2 疫情后“复工复产”热点分析

对于“复产复工”关键词下所得的 61013 条数据进行基于 K-mean 聚类的 LDA 主题分析，首先利用 Doc2vec 算法将每条微博向量化后，然后利用 K-mean 算法进行微博文本聚类，再对每个聚类进行 LDA 主题分析，通过多次观察、比较聚类结果，我们发现当 K 取 5 时，主题数量取 3 时，LDA 主题分析的效果最好，此时共得到 5 个主题簇。对每个主题簇挑选关键词，按照其簇内主题特点，将其分别命名为“复工复产热点”、“宅家热点”、“抗疫举措”、“复产复工保障”和“复工复产中的疫情防控”。

Topic1	
交通	0.0031
有序	0.0030
铁路	0.0025
消防	0.0025
车辆	0.0023
线路	0.0022
秩序	0.0020
道路	0.0020
助力	0.0018
区域	0.0017

Topic2	
开学	0.0032
有序	0.0028
复学	0.0027
口罩	0.0025
儿童	0.0024
形势	0.0024
学校	0.0023
居家	0.0023
人群	0.0022
距离	0.0021

Topic3	
总书记	0.0054
市场	0.0043
贷款	0.0042
农业	0.0039
餐饮企业	0.0037
数据	0.0033
政策	0.0029
会议	0.0028
精准	0.0028
金融	0.0026

图 4 “复工复产热点”主题簇

观察“复工复产热点”主题簇（图 4），可以看到三个主题个有明显的侧重，同时又包含共性，主题一明显侧重于交通，主题二侧重于复课，而主题三则侧重于国家政策。交通是返工复工、生产运输的关键；复课是学生和家长时刻关注的重点；国家政策更是涉及复产复工的最终进展，因此这三个主题共同刻画出了复产复工中最为牵动人心的热点话题，反映出随疫情好转复产复工工作的重点所在。

观察“宅家热点”主题簇（图 5），可以看到主题一涉及明星公益，主题二涉及税务信息，主题三涉及居家娱乐。这三个主题很明显都是三个月中居家已久的普通民众关心的话

题，主题一和主题三不用多说，主题二涉及的税务信息来自于 2020 年 4 月份集中展开的个人所得税退税。尽管该簇不直接涉及复产复工的话题，但它反映的疫情期间普通民众的兴趣所在仍然具有意义，结合“复产复工热点”主题簇，我们可以得到“复产复工”话题下更加全面的民众在不同时期、不同领域的关注点，这些关注点其实一定程度上反映出民众对“复产复工”的需求，对于我们最终确定、反向检验具有十分重要的作用。

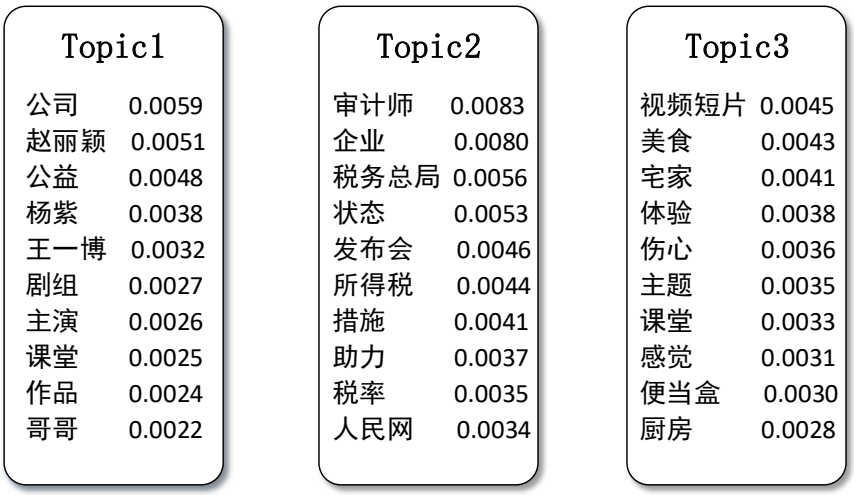


图 5“宅家热点”主题簇

如果说前两个主题簇具有的联系及自身特点为我们确定“复工复产”热点内容从一般民众的角度提供了指引方向，那么“抗疫举措”、“复产复工保障”和“复工复产中的疫情防控”这三个主题簇则具有更强的联系并从国家运行的角度为我们确定“复工复产”热点内容提供了答案。

观察“抗疫举措”主题簇（图 6），主题一是抗疫的动员与宣传，主题二是疫情的溯源与控制，主题三是疫情中的物资调度与生产维系。不难发现，该主题簇反映的是抗疫早期国家控制疫情的举措，此时虽然“复产复工”虽已提上日程，但抗击疫情仍然是主要任务，也就是说，疫情的控制是复工复产的前提，疫情发展是“复工复产”最重要的热点内容。接着观察“复产复工保障”主题簇（图 7），该主题簇涉及的信息非常多，主题一涉及各种公共机关，反映的是“复产复工”对消防、公安、用电等公共服务方面提出的挑战；主题二涉及道路、邮政等方面，反映出的是“复产复工‘对交通方面的要求；主题三涉及国家政策于补贴，反应的是”复产复工“离不开国家政策的指导和国家补贴的扶持。这些主题相当于国家以实际举措给出”复产复工“的对策，其对策的切入点便是我们要找的”复产复工“的具体热点内容。最后来看“复工复产中的疫情防控”主题簇（图 8），主题一涉及原材料的运输，主题二涉及进出口的疫情检查，主题三涉及政策实施的力度。三个主题综合来看其实就是在疫情特殊背景下“复工复产“的实施策略，即在绝对遵守抗疫政策的前提下有序开展”复产复工“过程，而主题一重点关注原材料运输毫无疑问是极为具有远见的，北京疫情的”复发“很大程度上就是食材原料运输过程中抗疫检测的疏漏。

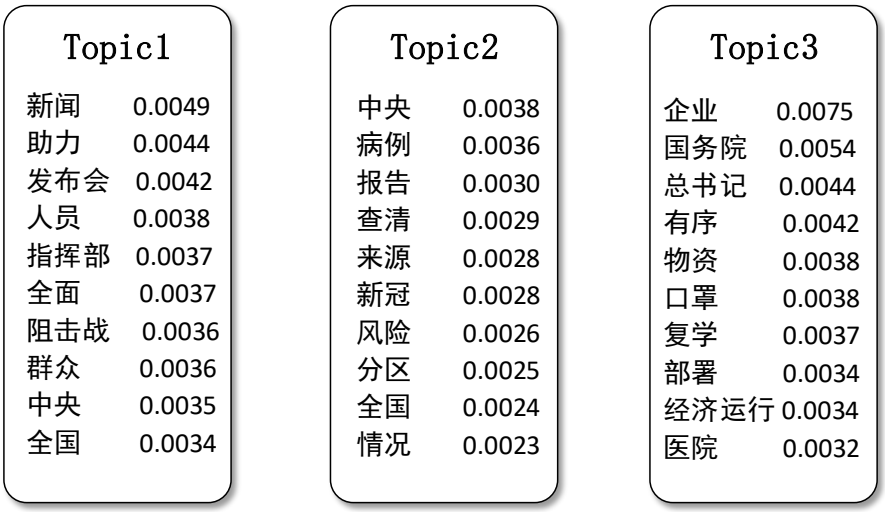


图 6“抗疫举措”主题簇

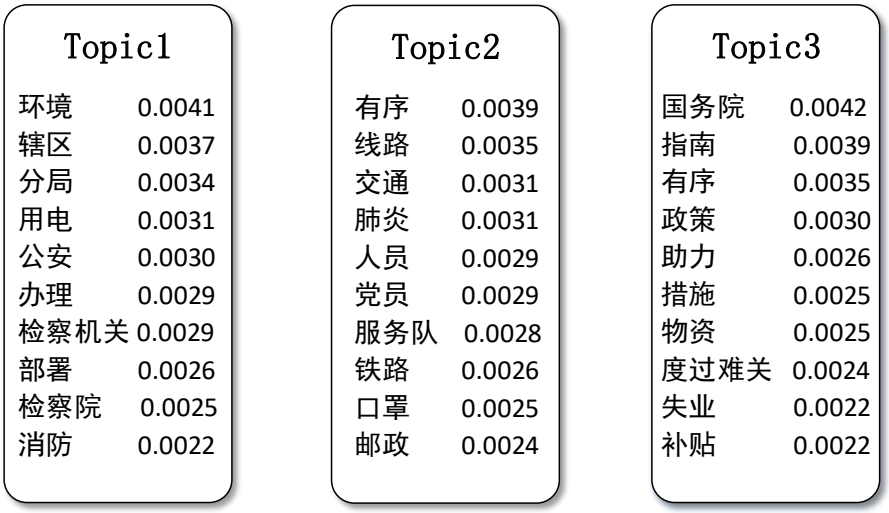


图 7“复工复产保障”主题簇

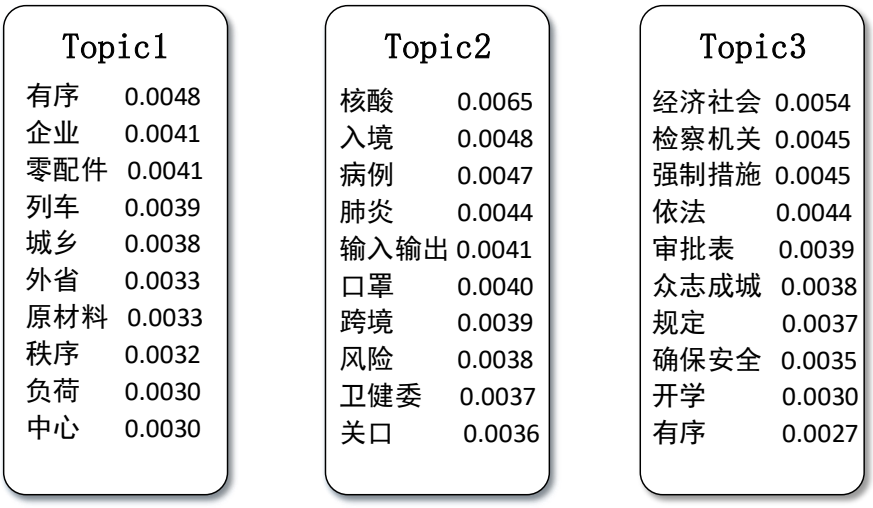


图 8“复工复产中的疫情防控”主题簇

通过对五个主题簇的分析，我们可以对“复产复工”及其热点内容做个总结。就微博“复产复工”话题本身而言，疫情期间大众的关注点集中在两个方面，一是疫情期间的居家热带话题，主要包括明星公益、重大政策；二是复产复工本身的热点，包括交通出行、

复产复学以及国家帮扶;从中我们可以看出大众眼中复产复工最为重要的便是交通保障以及国家政策。就微博“复产复工”热点内容而言，我们从国家举措的角度找寻主要内容，主题簇所体现的国家举措具有不同地切入点，一是疫情早期国家的抗疫动员和抗疫行动;二是国家为“复产复工”所做的具体保障，包括各政府机关的协调、交通运输的协调、国家政策的补贴；三是国家在“复产复工”中继续抗疫的举措，重点是原材料的运输以及抗疫检测。从中我们可以归纳出“复产复工”最引起关注的热点内容是疫情的发展，国家抗疫初期的不遗余力以及恢复生产过程中的检疫措施证明了这一点;“复产复工”的又一大热点内容因是国家机关的协调，各部门都需要为“复产复工”提供不同的保障；同时，不管是从大众还是国家的角度来讲，交通也是 “复产复工”的热点之一；此外，国家政策的支持也同样是“复产复工”的一大热点，最终结果如图 9 所示。

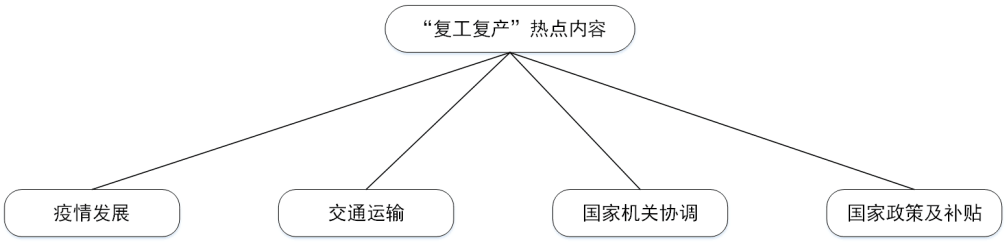


图 9“复产复工”热点内容

3.3 疫情后“消费”热点分析

对于“消费”关键词下所得的 22777 条数据进行基于频繁词集聚类的 LDA 主题分析。为了便于后续分析，对微博文档集合进行分词，共得到 86421 个词 (包括单个词语) ，最高词频达到 5016 次，而词频在 4 以下的词语占语料库的 75. 3%。这是因为微博的发布者为普通用户，微博中的词语为日常用语，比较少的使用术语，因此微博的长尾特征非常明显。本文选取词频为 10 以上的词语为特征词 (不包括单个字词语) ，且只保留动词和名词。

通过前文介绍的基于频繁词集的文本聚类方法，我们在挖掘频繁项集后得到了 1000 余个频繁项集，将挖掘出的频繁项集按一定顺序排列，意思相近的频繁项集划分在一起形成主题，这些频繁项集对应的文档在聚类后，最终形成了 “旅游”、“医疗”、 “娱乐”、“餐饮”、“购物”五个主题簇。下面对这五个主题簇进行 LDA 主题建模分析。

Topic1	
政策	0.0223
陇南	0.0202
河北	0.0201
江苏	0.0178
四川	0.0156
南京	0.0126
浙江	0.0124
游客	0.0122
景区	0.0120
交通	0.0119

Topic2	
半价	0.0253
景区	0.0233
门票	0.0210
机票	0.0195
酒店	0.0150
火车票	0.0133
下单	0.0115
交通	0.0113
网友	0.0112
数据	0.0111

图 10 “旅游”主题簇

从挖掘的结果中可知，对于“旅游”主题簇，存在 2 个潜在主题，图 10 中只展示了每个主题的一部分信息。根据不同的主题之间的相关性并不强可以说明基于类簇的 LDA 主题模型挖掘得到的主题，其关键词准确率高，很容易根据关键词得出相关主题。这里根据各个主题的高频关键词可以发现，Topic 1 侧重于热门旅游地点的主题，热门的地区在旅游的话题中被大量提及，这对于这些地区疫情后的旅游业的复苏具有重大意义。Topic 2 中我们可以提取“门票”“半价“等关键信息，以及 Topic2 中反复提及与交通成本概念相关的词语联想到疫情期间旅客数量急剧下降，航空公司本着尽量减少亏损的原则，卖一些低价票也十分合理，从而得出结论：相关景区门票的优惠与交通成本的降低，将会刺激疫情后人们的消费。

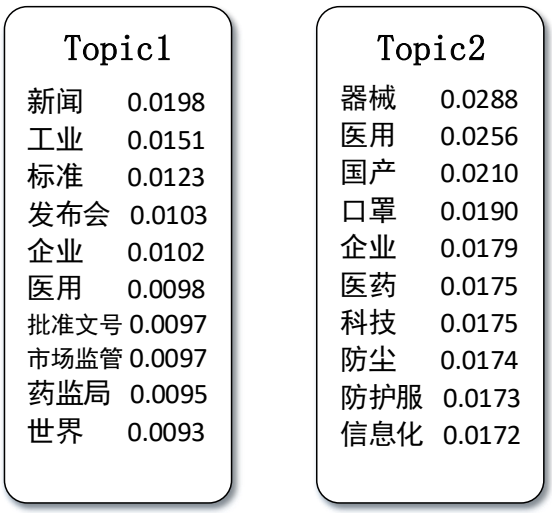


图 11 “医疗”主题簇

对于“医疗”主题簇（图 11），Topic1 中的高频词提及了官方的发布会以及市场监管标准。 4 月中下旬，商务部派出工作组赴广东等 10 省市，指导地方落实属地责任，强化质量监管。公安部、海关总署、市场监管总局、药监局等执法部门加大违法案件查处力度，公布了一批国内市场查处的非医用口罩质量不合格产品和企业清单。这一部分主题与疫情后日趋严格的医疗标准制定与监管力度相结合。Topic2 的高频词侧重于物资，反映出人们在疫情期间需要购置防疫物资，如日常生活中离不开口罩，企业复工需要添置相应的医疗器械以及防护服等从而带动消费。我国在疫情情况较稳定后，在保障国内防控和复工复产需求的基础上，防疫物资出口规模持续扩大，有效支持了全球抗击疫情。

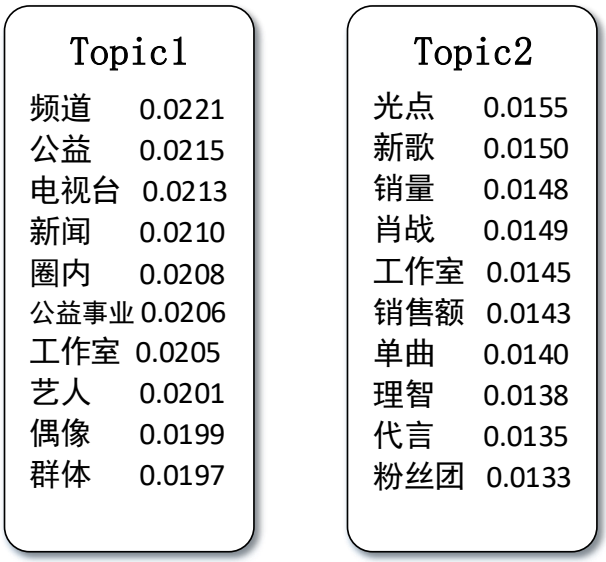


图 12 “明星”主题簇

娱乐明星方面（图 12）更多的焦距在了娱乐圈中的偶像与艺人。Topic1 中反映出公益事业在娱乐明星主题簇占据了很大的比例。黄晓明夫妇等一大批文艺工作者参与到公益和捐赠活动中来，这种公益结合明星效应的方式引起了大众的关注。Topic2 关注更多的是青年演员歌手肖战的专辑热卖带动了消费，肖战的专辑《光点》卖出 3000 多万份，销售额破亿。许多追星女孩年级不大还没有消费的能力，但在一些资本公司无休止的引导下，选择疯狂地消费购买专辑，也引发了人们对于理性追星、理性消费的思考。在著名外刊《经济学人》也有部分篇幅对相关事件所引起消费热潮的分析。

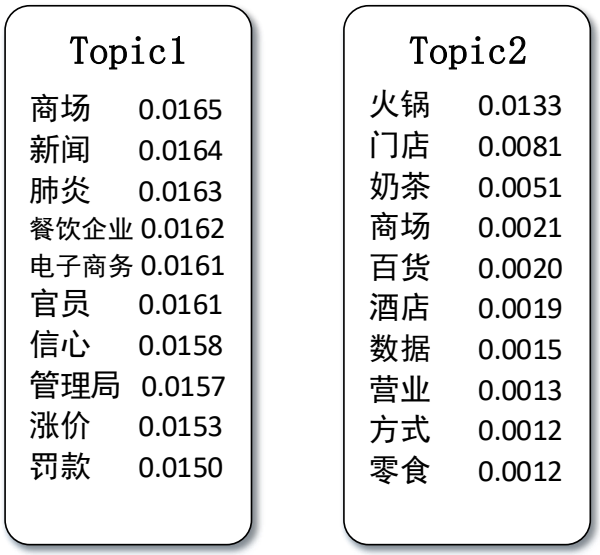


图 13 “餐饮”主题簇

餐饮业是国家商品零售业的重要组成部分，主要为国民经济的发展提供社会生活服务。图 13 中 Topic1 中反映出人们更多的关注是宏观政策的发布，疫情期间很多地方要求餐饮门店不可营业，但随着形势的好转，地方性支持政策陆续出台，不少企业也响应政策，支持餐饮行业。这些政策或多或少地刺激着人们的消费。Topic2 中出现的是各类餐饮业的具体消费种类，这些也是网民常常在微博提及且讨论的。民众的消费离不开日常的餐饮

食品，这也是在情理之中。

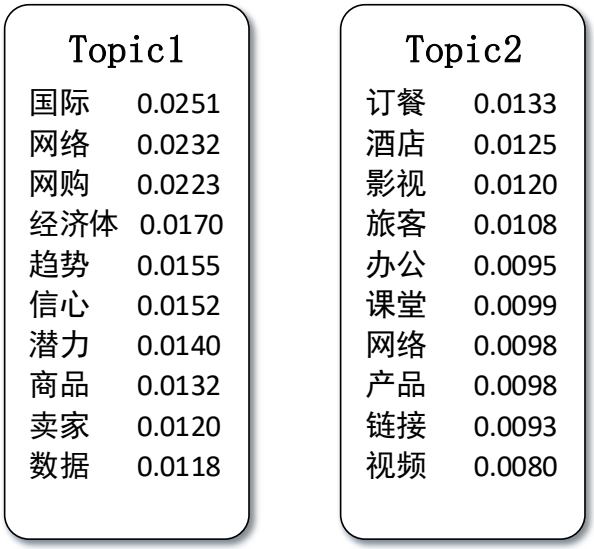


图 14 “购物”主题簇

人们的日常消费同样离不开购物，疫情之后也有很多人预言中国消费者将会进行“报复性消费”。但如图 14 中 Topic1 所示，人们谈论点主要聚焦于网购的方式。由于大多数商家在节日期间暂停营业，网购囤货成为中国百姓应对疫情生活的主要购物形式。得益于中国电商成熟的供应链管理和物流能力，大众的网购需求在隔离期间得到了较为充分地满足。Topic2 体现为具体的购物选项，除了常规的订餐、酒店等之外，引起注意的是课堂也开始进入公众的视野，因为疫情期间学校实行网课，相应衍生了许多收费的网络课堂，所以课堂也逐渐成为了人们教育投资消费的一大选项。

通过对“旅游”、“医疗”、“娱乐”、“餐饮”、“购物”五个方面研究疫情期间民众消费热点的分析，我们可以对“消费”热点内容进行总结。旅游方面：不少地区在旅游簇中被提及，同时提到的还有门票半价等相关的优惠政策以及低廉的交通成本，这对于疫情后的旅游业的复苏具有一定提示意义。医疗方面：民众对于医疗市场监管标准的关注度较高，并且民众需要购置防疫物资，成为日常消费的一部分。娱乐方面：明星与公益事业相结合展现了良好的效果，得到了公众关注的同时也带动了一部分消费如购置抗疫物资等。同时一些专辑新歌的市场也良好，粉丝经济对于消费的影响也日益增大。餐饮方面：民众关注政策对于餐饮行业的支持的同时，日常仍然讨论着美食，对于疫情期间在家中的民众来说，需要学习制作美食也是导致其讨论度较高的原因之一。购物方面：网购囤货成为中国百姓应对疫情生活的主要购物形式。从最后得出的图 15，我们可以看出，“消费”话题下广泛的讨论不仅有政策的调控，还有市场的调节。但无论是哪个方面，疫情带给消费的影响都是深远且复杂的。

3. 实证研究

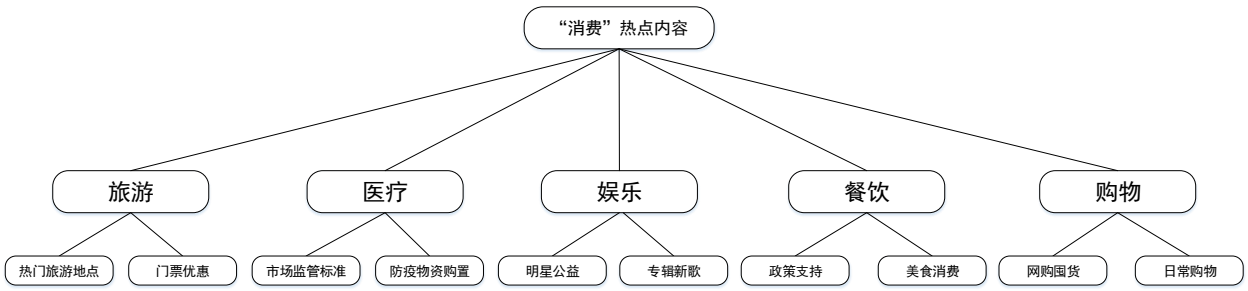


图 15 “消费”热点内容

4. 结束语

4.1 总结与建议

微博文本已经广泛应用于各类研究，基于微博文本进行的LDA主题分析也屡见不鲜，但由于微博文本的特殊性，传统的方法往往不能取得理想效果，因此，我们提出了基于K-mean 聚类的 LDA 主题分析和基于频繁词集的 LDA 主题分析两种方法，这两种方法都是在 LDA 主题分析前对微博文本进行划分，以解决 LDA 在处理大量复杂文本时得到的主题含义模糊的问题，最终我们对“复工复产”、“消费”两类微博文本分别采用以上两种方法，取得了良好的效果。

由于我们采用的 LDA 主题分析的处理不同，所得到的结果也有所差异，就“复工复产“而言，其文本数量较多，没有预先设置主题簇，因此得到的主题簇能够真实反映该主题簇所代表话题的重要程度，而对微博热点的获取也更多的是从整体性、全局性的角度上分析得到的；就”消费“而言，由于其文本数量相对较少，K-Means 聚类效果并不明显，于是我们根据词频合理的得到了若干个主题簇，这样做虽然面临丢失少量信息的风险，但所得到的信息更加细致，也更具参考性，总体而言，两种方法各有其利弊，但的确都在大量微博文本的基础上给出了”热点“的参考。

基于实验结论，我们可以提出如下建议：

1. 在“复工复产“方面，政府发挥着决定性的指挥协调作用，因此，政府一方面要做好内部机关抗疫防疫、恢复生产的准备、协调工作；一方面要出台相应政策，要稳定大众的的心态，避免因疫情带来的恐慌，减少复工复产带来的混乱，要明确补贴政策，对相关企业、失业人群做好安置，要严格把控交通运输中的疫情监测工作，坚决避免疫情二次复发的情况。

2. 在“消费”方面，政府同样起着举足轻重的作用。政府应该加大对于景区优惠政策的调整，从而刺激旅游消费，旅游热门地区也应提前做好准备，在保证安全的前提下适当开放景区。在其他方面，政府也要加强对市场的监管力度与对餐饮行业的扶持力度，让产品安全有保障，商家有缓冲的空间，民众可以放心地消费。

4.2 研究不足

(1) 数据选取不够全面。本文研究数据的选取范围为 2 月 1 日至 4 月 30 日微博关键词为“消费”、“复工”和“复产”的热门微博，但以这些词为关键词也并不能确保能完全收集到疫情期间全部的与消费或复工复产有关的数据，因此本实验所得到的结果也具有一定的局限性。

(2) K-Means 算法的聚类结果与预设的聚类中心数有非常大的关系，而本实验采用多次实验之后主观选取合适聚类中心数的方式，会使结果带有一定主观性。

(3) 本文基于频繁词集的聚集方法具有一定的主观性。本文在频繁词集的选取上，依靠以往文献与经验对词集进行了筛选，剔除了与消费关联度较低的词集，同时在选择意思相近的频繁词集形成主题簇的过程中也是根据经验选取的，因此基于频繁词集的微博文本聚类结果也带有一定主观性。

参考文献

[1]王先庆,矫萍.疫情防控下常态化提振广东消费的对策建议[J].广东经济,2020(06):36-41.
[2]王可山,郝裕,秦如月.农业高质量发展、交易制度变迁与网购农产品消费促进——兼论新冠肺炎疫情对生鲜电商发展的影响[J/OL].经济与管理研究:1-11[2020-07-09].https://doi.org/10.13502/j.cnki.issn1000-7636.2020.04.003.
[3]金晓彤,宋伟,赵太阳,姚凤.公共卫生事件对居民非理性消费行为的影响[J/OL].西安交通大学学报(社会科学版):1-16[2020-07-09].http://kns.cnki.net/kcms/detail/61.1329.C.20200616.1902.006.html.
[4]胡越秋,王军,董泽华.新冠肺炎疫情防控期间企业复工决策分析——基于行为经济学视角[J].统计与决策,2020,36(05):157-160.
[5]丁任重,李俞,李标.新冠肺炎疫情下如何复工复产:基于产业链视角[J].财经科学,2020(05):65-76.
[6]徐玉德.全球疫情冲击下中小企业面临的挑战及应对[J].财会月刊,2020(12):114-118.
[7]张瑞,董庆兴.基于LDA-HMM的知识流动模式发现研究[J].情报科学,2020,38(06):67-75.
[8]陈欢,黄勃,朱翌民,俞雷,余宇新.结合LDA与Self-Attention的短文本情感分类方法[J/OL].计算机工程与应用:1-8[2020-06-24]
[9]屈渤浩. 基于改进FastText的中文短文本分类方法研究[D].辽宁大学,2018.
[10]王永恒，贾焰. 海量短语信息文本聚类技术研究 [J] .计算机工程，2007，33 (14) : 3840.
[11]唐晓波,房小可.基于文本聚类与 LDA 相融合的微博主题检索模型研究[J].情报理论与实践,2013,36(08):85-90.
[12]BLEID，NGA, JORDAN M. Latent dirichlet allocation [J] . Journal of Machine Learning Research , 2003(3).