**三、数据的处理和分析**

**3.1 数据来源**

在数据源的选择中，我们选用来自国家应急管理部以及人民网2020年1月21日至2020年11月25期间的疫情相关新闻，共计4791篇文章，旨在通过对官方报道进行分析得出官方在疫情各个阶段的关注点以及心态。 ****

**3.2 研究方向**

我们希望通过不同主题在疫情发展的不同阶段内的占比，得出官方的态度变化。由于数据没有主题标签，所以我们选择了无监督的聚类算法对文本进行归类，然后提取每个类别的关键词，人工分析得出该类别主题，再为每篇文章进行标注。

**3.3分析方法**

图3.1-官方报道高频词

关于聚类算法，我们选择了K-Means（K均值）聚类算法，并对每篇文章生成tf-idf向量作为文本特征。为了进行可视化展示，我们使用了T-SNE以及PCA两种降维手段对文本特征进行降维。

**3.4分析方法简介**

**K-Means聚类算法**。K-Means算法又称k均值算法。是一种基于样本集合划分的聚类算法，属于无监督学习。

规定个样本之间的距离为欧氏距离平方

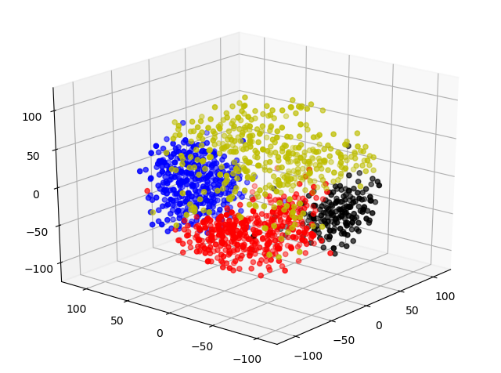
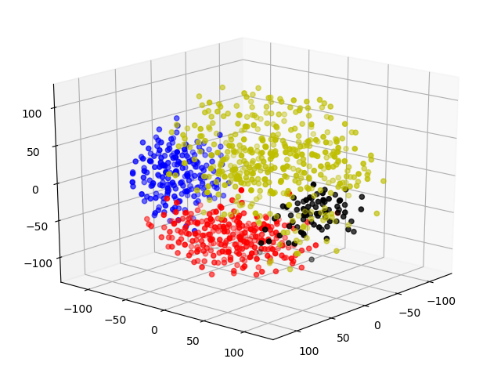
定义样本与其所属类中心之间的距离总和为损失函数，即

其中,是第l个类的均值（或称中心）， 是指示函数，取值为0或1.函数**W(C)**可以用来表示一个类中的样本相似程度。

那么，k均值算法即转为求解最优化问题：

K均值算法通过迭代进行求解。首先选择k个类的中心，将所有样本依次指派至与其最相近的类中，更新添加样本后的样本中心，重复上述过程直至收敛。收敛后的划分结果即为k均值算法的聚类结果。

**3.4分析过程**

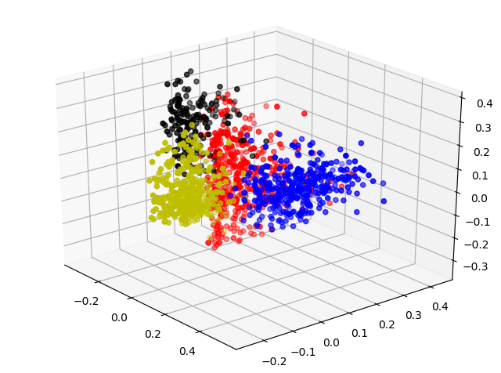
 在具体的无监督训练中，我们随机选取了4477篇报道作为训练集，314篇文章作为测试集，比例大致14：1。对训练集文本切词后计算得到tf-idf文本特征，使用K-Means模型进行训练，在多次尝试后，发现将文本分为4个主题效果较好。可以看出，不同主题的文章特征具有较好的区分度。

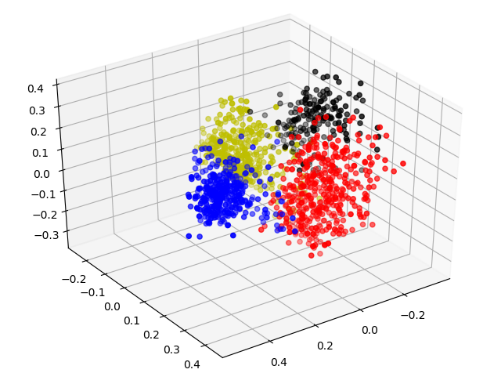
**黄色-抗疫相关**

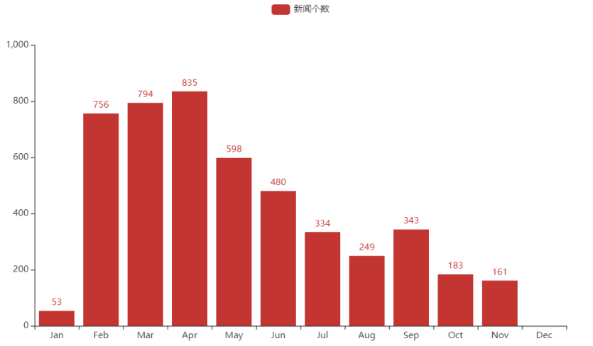
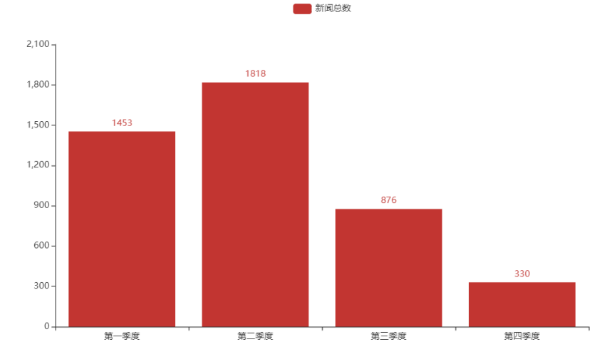
**蓝色-经济相关**

**红色-国际相关**

**黑色-脱贫攻坚**







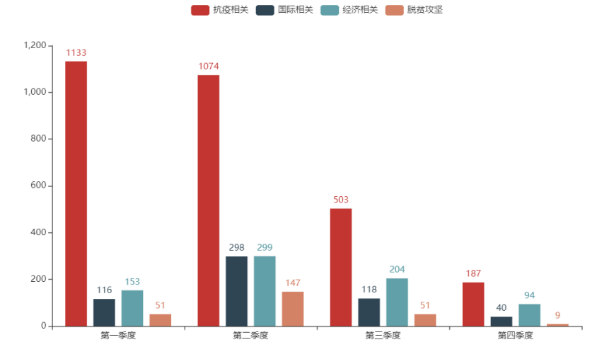
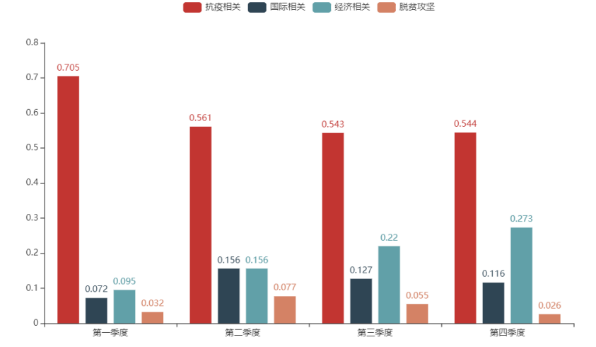
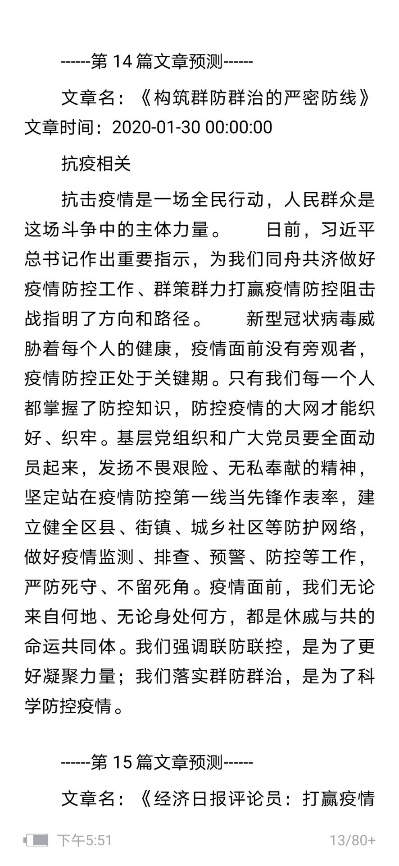
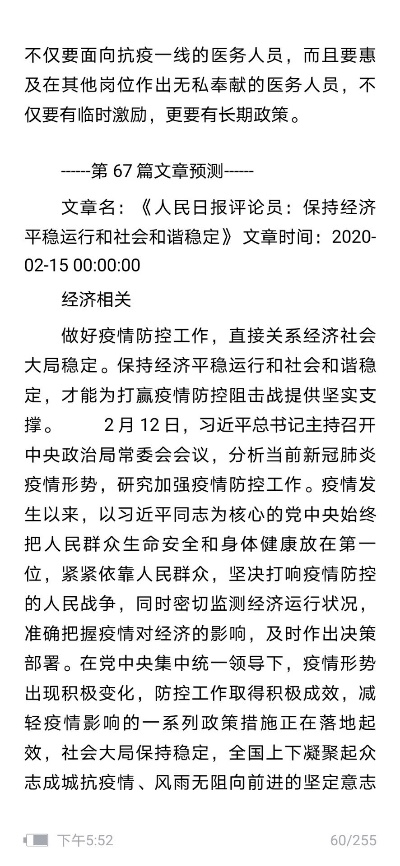
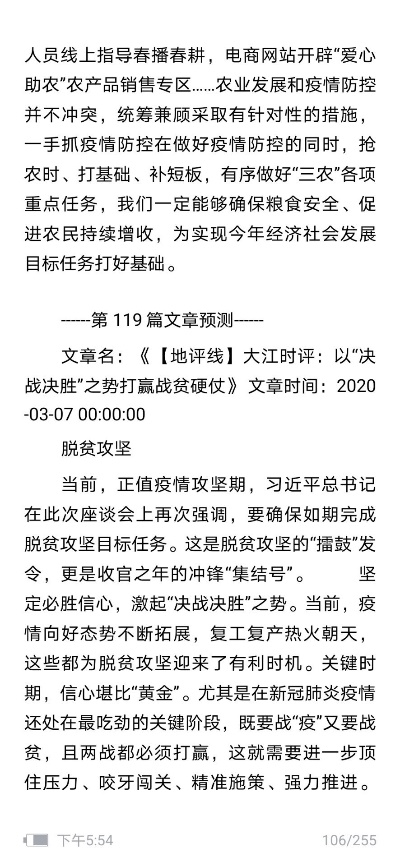


图3-7 每季度新闻主题占比变化

图3-6每季度新闻主题变化

从分类结果中可以看出，抗击疫情一直是官方报道中最侧重的部分，而随着国内疫情得到控制，复工复产、产业补贴等经济相关的报道占比逐渐上升。全球疫情形势、国际合作等国际相关报道占比也有所提升。而作为全面建成小康社会的收关之年，疫情形势下的脱贫攻坚也在官方的关注之中。考虑到筛选过的数据大都与疫情强相关，对经济、国际、脱贫等内容有较大滤去作用。我们认为这样的结果足以说明官方在疫情期间的关注变化符合实际情况。

**3.5模型的预测效果**

 对测试集的主题预测效果保存在predict\_res.txt文件中。经过评定后认为该模型有着较好的分类效果。