基于Python的《新闻联播》文本数据分析报告

中国社会科学院大学 庄妍[[1]](#footnote-1)

**摘要：**本文研究对象为自2009年6月26日至2020年7月8日共11年的《新闻联播》标题和内容的共计61712条新闻的全文本数据。本文基于Python中的pandas、numpy、jieba、gensim、scikit-learn、pyecharts、seaborn等库，搭建基于潜在狄利克雷分配（LDA）模型提取主题、词频-逆文本频率指数(tf-idf)并利用指K均值聚类算法(K-Means)进行文本聚类、调用了百度飞桨生态下的中文NLP开源工具Senta构建双向长短期记忆网络（Bidirectional LSTM）模型进行情感预测、利用Pyecharts绘制不同地区被报道次数的地图。采用以上人工智能算法和数据分析工具，深度解构《新闻联播》作为国内收视率最高、影响范围最广的新闻类节目的语言特点，并挖掘其作为官方宣传媒体在经济、政治、文化、外交等多领域的隐含宏观信息和导向。

**关键词：**《新闻联播》、文本数据挖掘、主题模型、自然语言处理

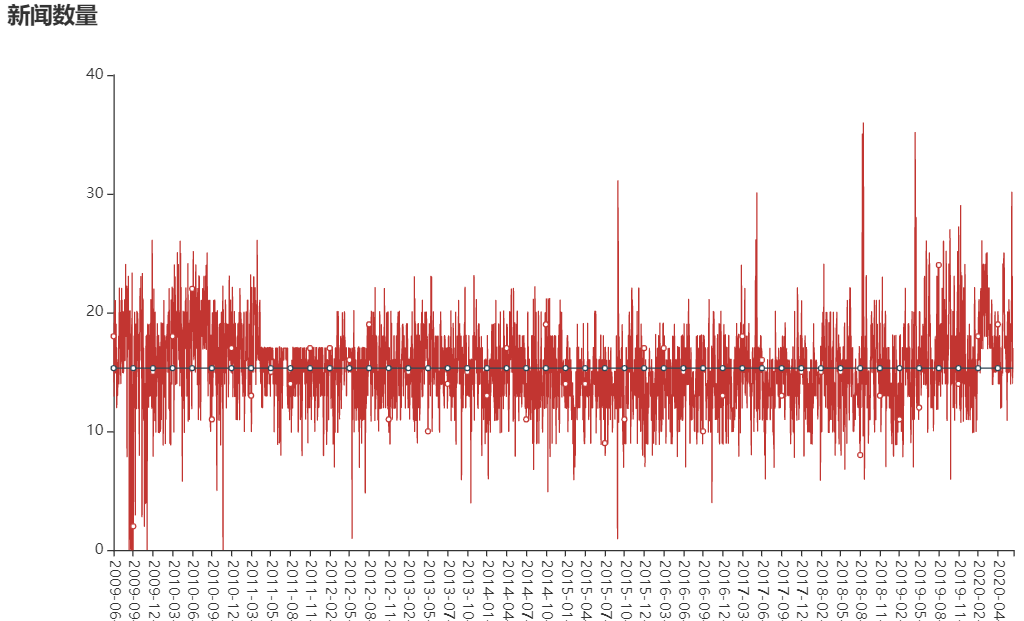
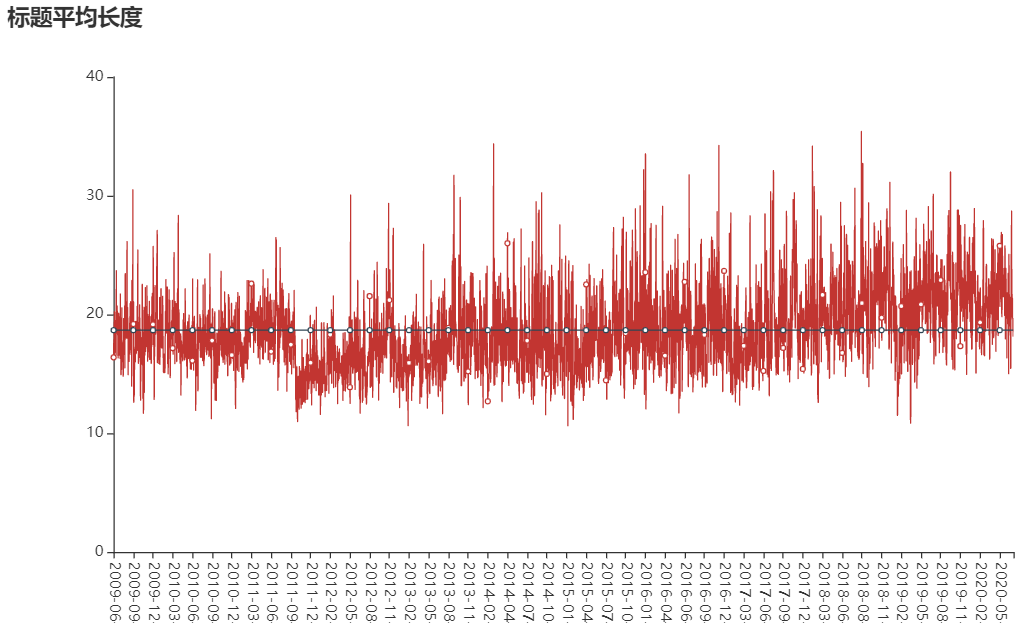
央视网2020年初新闻类节目收视盘点结果显示，《新闻联播》的收视率为11.15%，收视份额为34.83%，单期最高收视率超15%，单期最高份额超45%，位居全国新闻节目中收视第一名。[[2]](#footnote-2)可见，《新闻联播》作为最重要、收视率最高、影响力最大的综合电视新闻节目该节目从改革开放以来一直保留至今，而且至今仍然是各大地方电视台甚至是广播节目需要同时并机直播和转播的节目，可见其重要性。[[3]](#footnote-3)而在其内容上也呈现出“国内时政+要闻 + 各地形势 + 国际新闻”的固定形式。

《新闻联播》是中央电视台综合频道推出的晚间新闻节目，节目中涵盖了包括政治、经济、科技、社会、军事、外交、文化、体育、农业、交通等各个方面的新闻，拥有十分重要的社会影响力，节目宗旨为“宣传党和政府的声音，传播天下大事”。

《新闻联播》从开播以来历经了数十年的时间，是中国收视率最高、影响力最大的电视新闻栏目，以客观、生动、丰富的纪实手段记录了中华大地每一天发生的变化，是一部最真实的历史教科书。《新闻联播》的文本数据在某种程度上也可以被称为大数据，可以使用现有的大数据技术对其进行分析，从中挖掘出有意义的信息。

#### 探索性数据分析

《新闻联播》的每日播报的新闻条目平均为15.3条，每条新闻标题平均长度为18.7个字，日播报内容长度为平均每天7046字。以下三项指标每日波动情况如下图 1 每日新闻数量、标题平均长度、播报内容长度所示。虽然每日新闻总数波动较大，但是播报总长度波动幅度较小，因此每日详略安排可以一定程度上体现新闻事件的重要性。

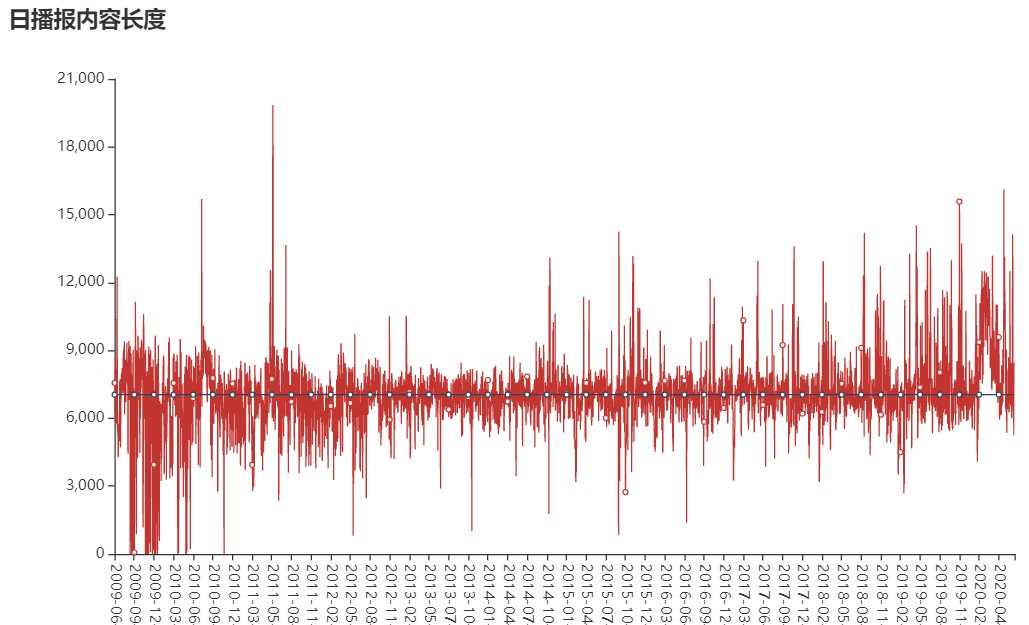


图 1 每日新闻数量、标题平均长度、播报内容长度

而对于播报内容最长的5篇新闻的标题和内容分别如图 2所示，其主题为地区经济振兴和会议公报。其中最长的一条新闻长达4816字，体现了国家在政治传播中对经济建设的重视。

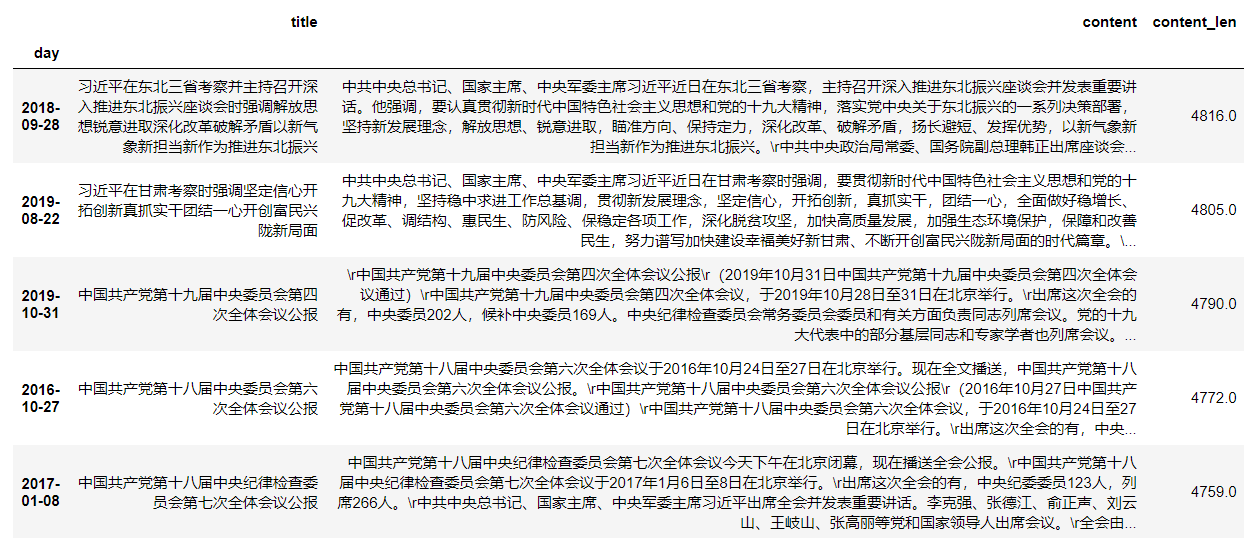


图 2 内容字数最多的新闻条目

#### 词频分析

词频，是字词在一个语料库的出现次数，用来评估一个词在语料库中的重复度，体现了这个词在语料库的重要程度。

首先，先进行词性分析。如词性分析结果所示，较高频率的地名，强调宣传客体的地区多样性。《新闻联播》注重实词且修饰词较少的语言风格，能够客观、精炼地传达信息，能够最大程度上确保信息传播的准确性，体现了新闻语言的严肃性和规范性。

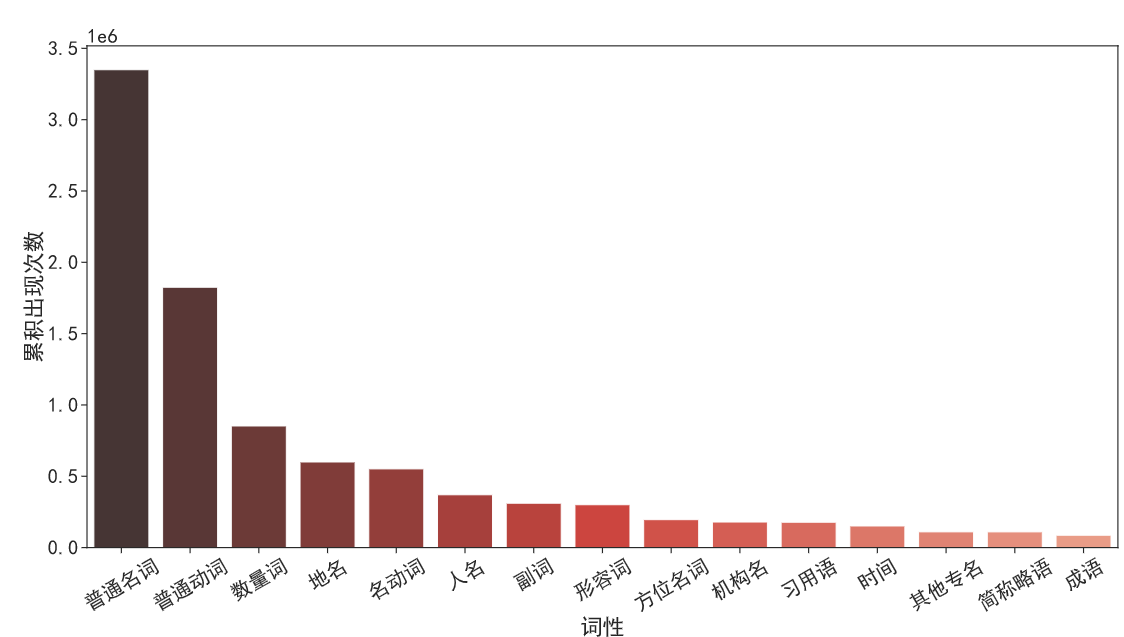


图 3 词性分析

对于词频分析，本文利用Python中jieba库按照TextRank 抽取特定词性的实词关键词，并依照百度停用词表，删去停用词后，得到每条新闻中的关键词。将以上关键词统计后，得到图 4中的词云并统计了词频最高的20个关键词。



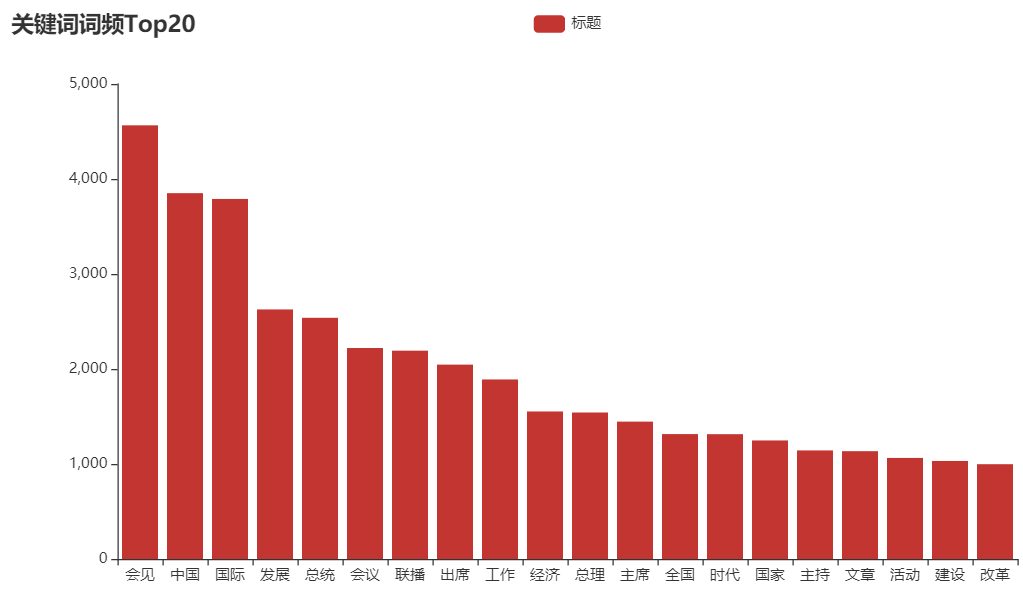


图 4 关键词词云和词频

根据以上结果可以很清晰地抓住《新闻联播》的重点内容。其中，“中国、全国、国家”等词的高频出现，体现了《新闻联播》作为官方宣传的主体性，表达了对国家主体性的认知。在这一现象中，艾红红(2008: 94) 认为这是一种“自觉地为党和政府‘代言’的身份， 以一种不可质疑的全知全能姿态， 向受众描述它认为应该了解的过去所发生的事件以及这些事件背后的意义，同时还有其关于世界、关于政治、经济以及人性的一整套理念”[[4]](#footnote-4)。而“发展、经济”这类表明经济发展的词频很高，体现了党和国家以经济建设为中心的基本路线。

进一步分析新闻内容中含有“经济”的条目，总计15646条，占全部新闻条目数的25.4%。可见经济主题已经渗透在相当一部分的播报新闻中了。进一步按年统计汇总后，除了第一年和最后一年中不是全年数据，其余年份中经济相关报道的总数较为稳定，经济发展始终维持在重要地位。

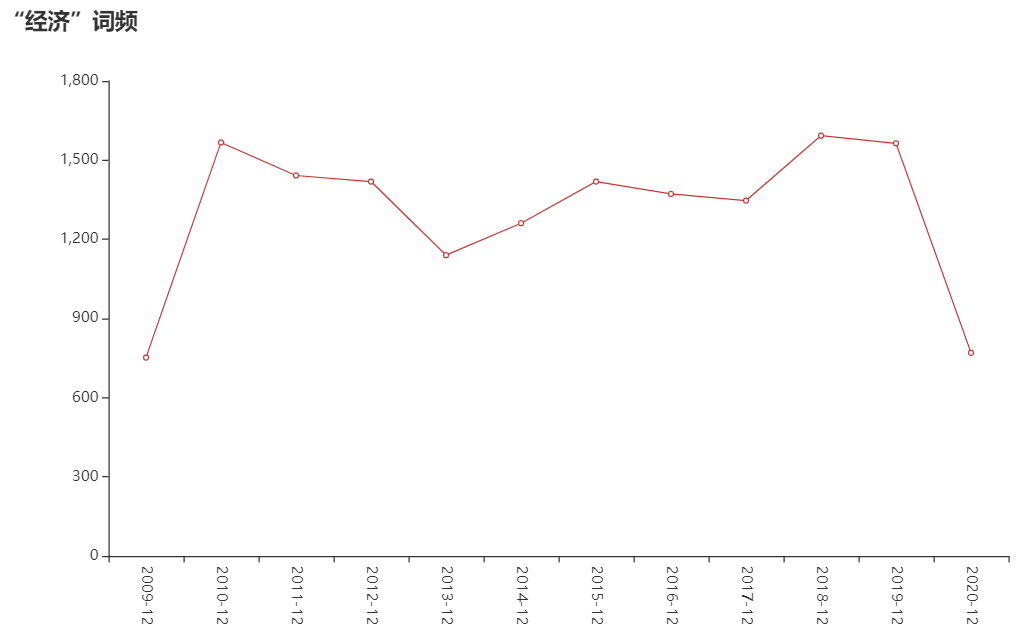


图 5 包含“经济”的新闻条目数按年汇总

#### 地名作为新闻报道客体的地理分布分析

在上文的客体分析和词性分析的结果显示，《新闻联播》中地名的出现次数在所有词性中位居第4位，地区作为报道的客体具有非常高的出现频率，彰显了其在语料库中的重要性。而《新闻联播》作为全国收视、各地方台都需转播的全国性新闻类节目，其报道地区也覆盖了全国各地，因此对地名进行地理分布分析，能够具有较高的数据挖掘价值。

图 6 中国各省份被报道次数分布图显示，北京、上海被报道的次数最多，因其作为中国最大的直辖市和政治经济中心，经济社会运转迅速，发生事件的数量多，因此被报道的次数最多不足为奇。此外，曝光次数次之的则是新疆、西藏等西部地区，其被多次报道这一现象，可以体现中央对西部地区的平安稳定和繁荣发展的高度重视和关心。除去这几处地区，《新闻联播》对于不同省份的报道并没有太大差异，各省的报道频率基本相近。因此，对于不同省份的报道，呈现出“均衡关注，重点突出西部地区”的特点。

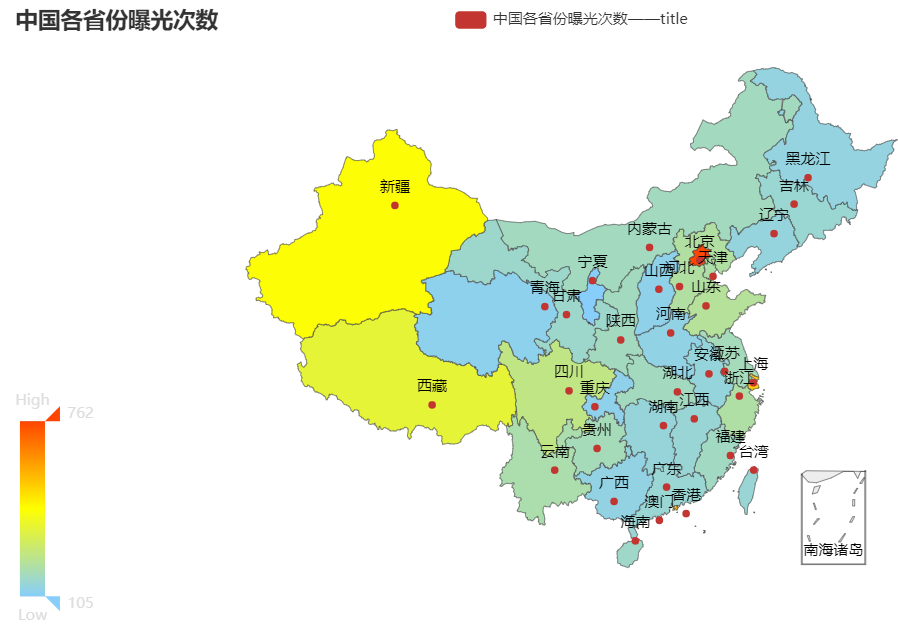


图 6 中国各省份被报道次数分布图

对于下图图 7 中国各城市被报道次数热力图所示的结果，就与上图的省份分布有较大差异 ，报道的城市主要分布在珠三角、长三角和京津冀三处经济圈，差异的原因应当在于新闻标题较为简洁，对于出现具体城市名，相比于出现省份名，其报道内容通常为小范围高强度的事件，因此报道内容会主要集中在人口较为密集、经济社会活动较为频繁的东部沿海城市圈。

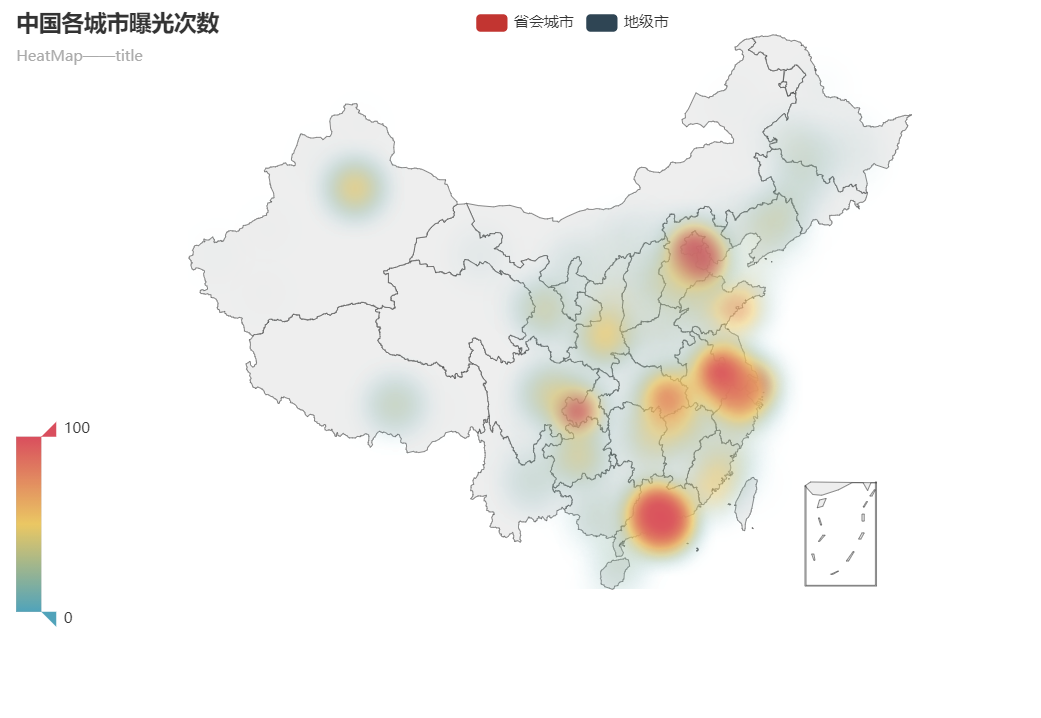
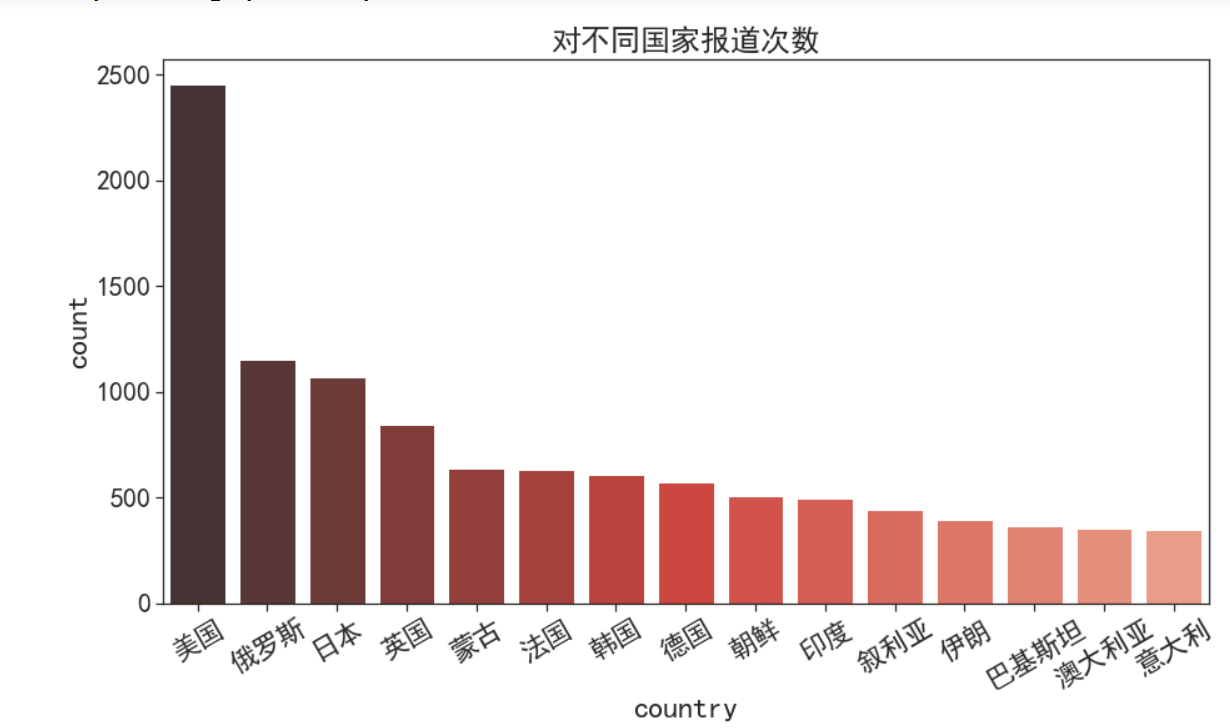


图 7 中国各城市被报道次数热力图

《新闻联播》的国际新闻模块，作为一个固定的节目类型，一直长期稳定地存在于节目的最后部分。国际新闻从内容上分为五类：地区热点、突发事件、国际关系、会议报道、后续报道。

根据图 8 除中国外世界各国报道次数柱状图及分布地图所示，《新闻联播》中报道频率较高的主要是以美国、英国和法国未首的西方国家，其次是如俄罗斯、日本、韩国和朝鲜等周边邻国，再次为叙利亚、伊朗和巴基斯坦等中东战乱国家。在这些国际新闻中，涉及具体国家时，大多是战争、政变、恐怖事件、天灾人祸以及部分国家的政党轮替等新闻，在报道国家组织的会议时，又时常会突出强调国际霸权主义等。对于国外，特别是对部分西方国家的负面报道，一定程度上有效地防止了受众认识上的片面和偏激，有利于增强社会凝聚力，维护国内的安定团结。从这一点上看，《新闻联播》爱憎分明，坚定、昂扬的话语表征，是符合“坚持团结稳定鼓劲，牢牢把握正确的舆论导向”的方针的。[[5]](#footnote-5)



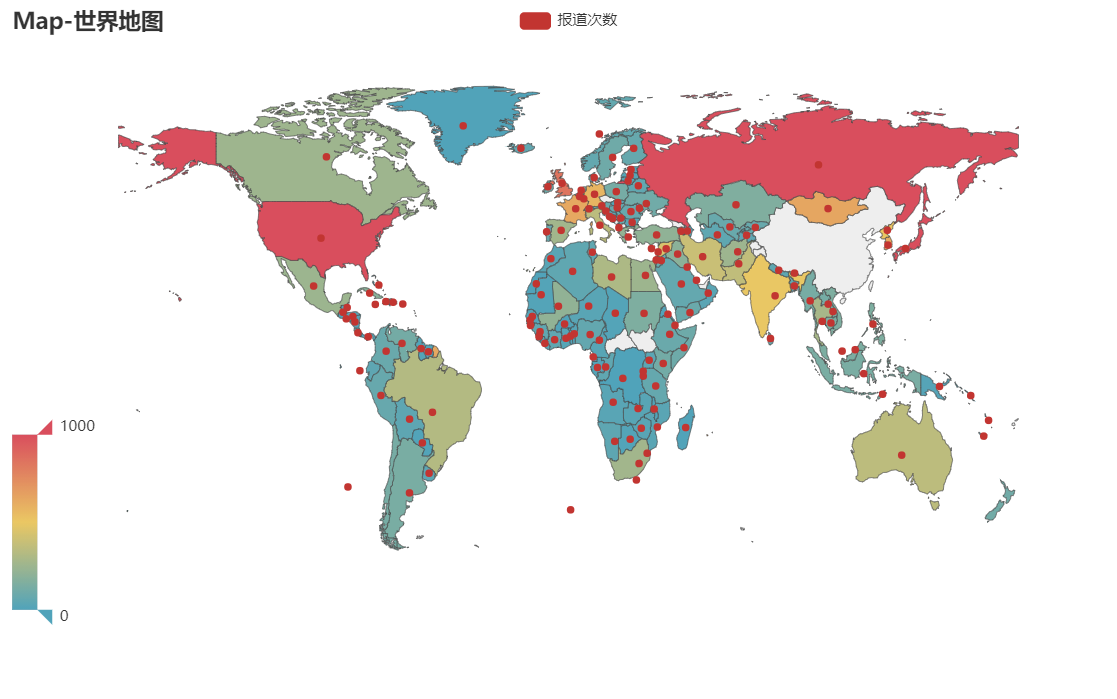


图 8 除中国外世界各国报道次数柱状图及分布地图

#### 内容挖掘：主题挖掘和聚类分析

##### **基于****潜在狄利克雷分配（LDA）的内容主题挖掘**

潜在狄利克雷分配（Latent Dirichlet Allocation，LDA）是一种非监督机器学习算法模型，用来识别语料库中潜藏的主题信息。它采用了词袋的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。

本文利用Python中的gensim库实现LDA模型。gensim用于主题建模、文档索引和大型语料库的相似性检索。目标受众是自然语言处理(NLP)和信息检索(IR)社区。[[6]](#footnote-6)

利用上述模型，输出结果如图 9所示，新闻主题可划分为行政治理、经济增长、改革创新、社会民生、疫情防控、脱贫攻坚、核心价值、精神文明、强军战略、外事活动、国际合作、全球事件类的内容。可以看出《新闻联播》作为官方新闻，内容紧紧围绕党和国家的工作重心————以经济建设为中心，并且除了像其他新闻对于突发事件的报道，还承担着巩固意识形态、公开政治决策、宣传国际形象的重要功能。

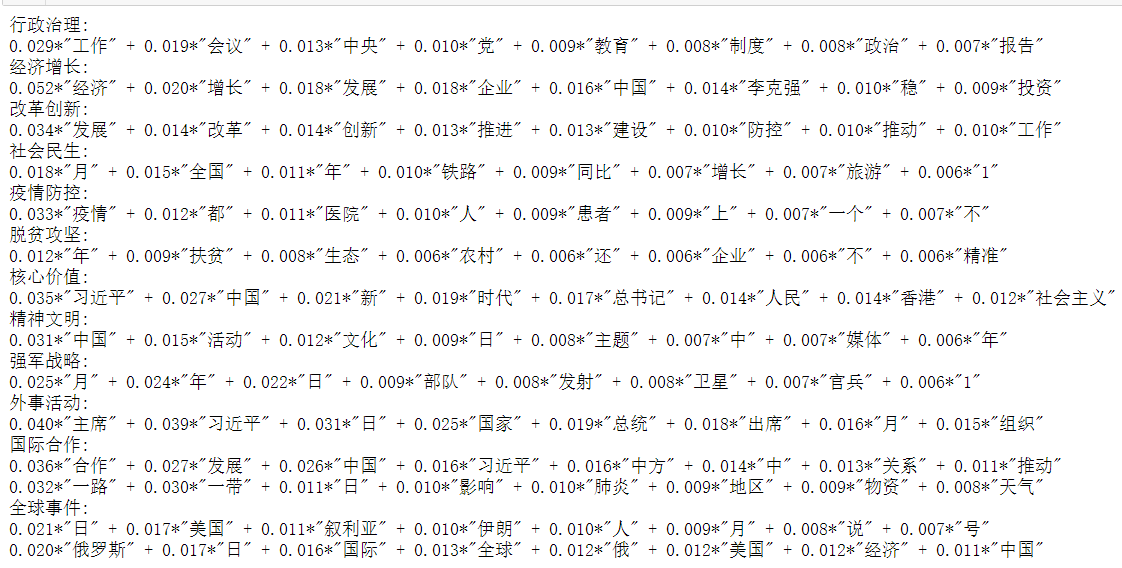


图 9 基于潜在狄利克雷分配（LDA）模型提取主题信息

##### 基于词频-逆文本频率指数(tf-idf)的K均值（K-means）聚类

首先，从sklearn.feature\_extraction.text库调用TfidfVectorizer功能，利用词频-逆文档频率指数（term frequency–inverse document frequency，tf-idf），将原始文档集合转换为tf-idf特性的矩阵，反映词项在文档间的分布，提供了基于频次的特征选择。tf-idf是一种用于信息检索与数据挖掘的常用加权技术，得分较高的词对于文章的主题的能力强，因此赋予更高的预测权重。例如以下图 10是是《新闻联播》语料库中tf-idf得分最低的十个词，是该语料库去掉停用词后仍然中出现频率较高的词，对判断一则特定新闻内容有帮助不大。而图 11是《新闻联播》语料库中tf-idf得分最高的十个词，主要为人名地名和专有名词，具有较强的辨识性，对判断一则新闻主题内容的重要性很高，判断主题过程中应当给予更高权重。

图 10 tf-idf得分最低的十个词

图 11 tf-idf得分最高的十个词

完成将原始文档集合转换为tf-idf特性的矩阵的向量化过程后，利用K均值聚类（K-means），聚类原始结果按照字数汇总后如图 12所示，其中因为第2类和第4类内容较少，故删掉这两类后，参考主题分析结果为每一类命名，重新汇总每一类新闻被报道的总字数和条目数，统计结果和可视化分别如以下图 13和图 14所示。可见，社会民生和改革发展为《新闻联播》最关注的两类内容，而根据图 15 聚类被分为“改革发展”的5条新闻内容，体现了《新闻联播》作为官方媒体，关注的客体主要为人民群众的生活福祉和社会的发展问题，彰显了其作为受众最广泛的新闻媒体对于人民生活朴素的关怀。

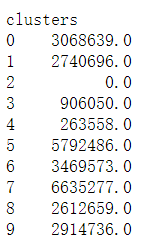
 

图 12聚类原始结果字数汇总统计

图 13 聚类结果命名并统计字数和条目数

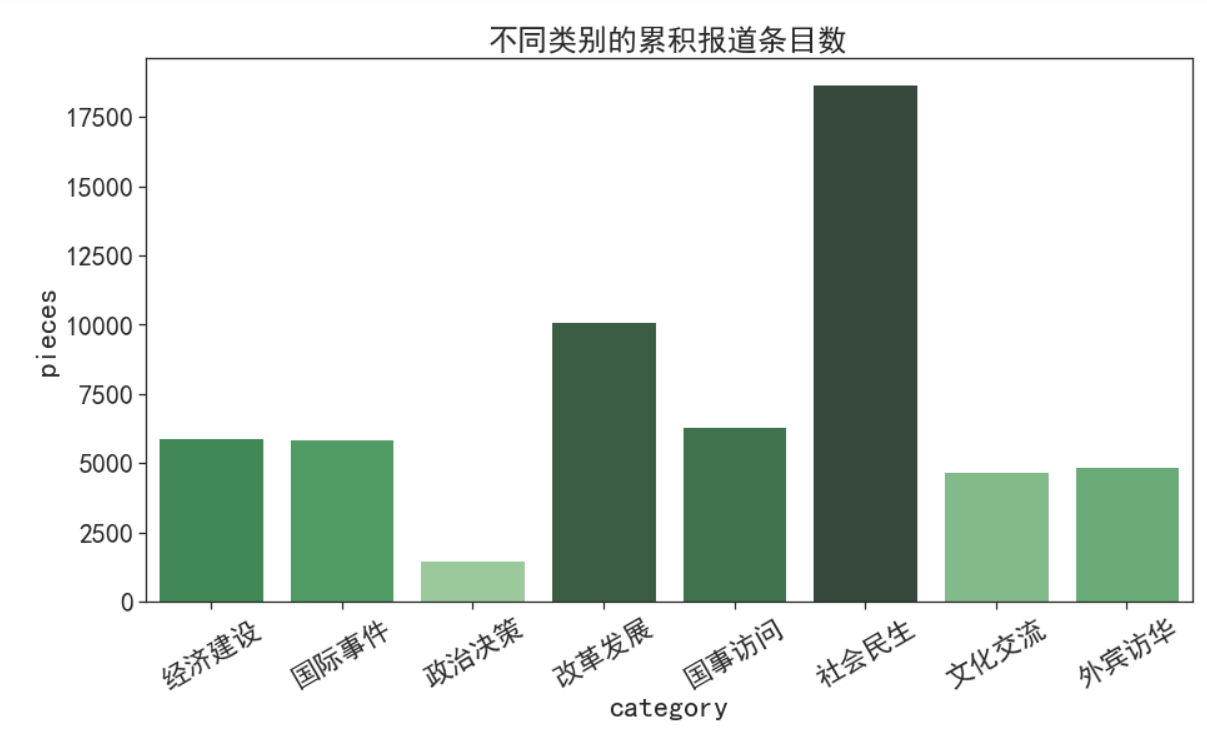
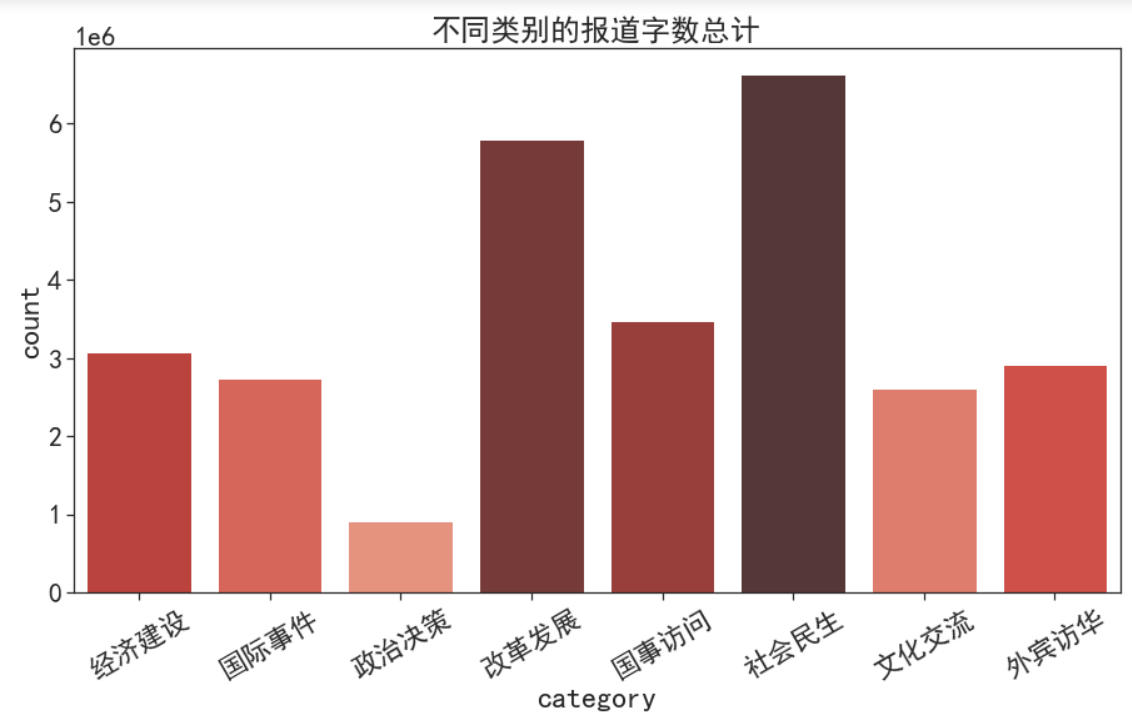


图 14 不同类别新闻节目累积报道字数和条目数可视化分析



图 15 聚类被分为“改革发展”的5条新闻内容

#### 情感倾向分析

本文调用了利用PaddleHub获取了百度“飞桨”（PaddlePaddle）生态下的自然语言处理部中面向工业应用的中文NLP开源工具——情感倾向分析（Sentiment Classification，简称Senta）模型。其中构建了双向长短期记忆神经网络（Bidirectional Long Short Term Memory，BI-LSTM），该模型为序列模型，采用双向LSTM结构，能够更好地捕获句子中的语义特征。[[7]](#footnote-7)

本文利用上述模型，对每条新闻的情感倾向进行预测分析。情感得分分布在0至1之间，得分越高证明情感越积极。《新闻联播》语料库的平均感情倾向得分0.728，即有73.8%的报道为积极情感。在图 16 新闻的情感得分分布情况箱线图和分布图中，可以看出情感得分分布呈双峰分布，集中在两端，表明绝大多数报道的感情倾向很明确。符合新闻报道清晰明确的特点，尤其是《新闻联播》作为官方媒体其鲜明的立场和观点。

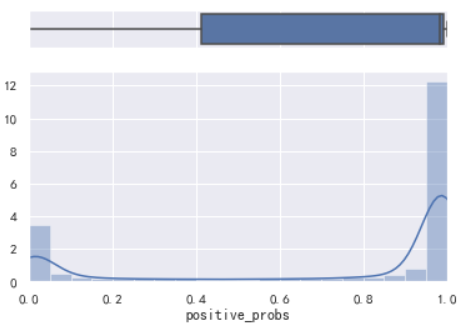


图 16 新闻的情感得分分布情况箱线图和分布图

#### 小结

本文整理了自2009年6月26日至2020年7月8日《新闻联播》11年全文本结构化数据，基于Python搭建机器学习算法和模型，进行大数据分析和挖掘。

在探索性数据分析部分，采用了pandas库对新闻数量、标题长度和每日播报总字数进行统计，以此为特征反映每条新闻的重要程度。在词频分析中，对《新闻联播》的词性进行分析，发现地名的出现频率较高；在对词频较高的关键词的语义分析中，本项目通过发现文本中对“中国、全国、国家”等词高频的使用，显示出其官方主体性，在对被报道的地区频次分析中，报道地名作为客体的分布情况，并利用pyecharts库中的内置地图进行可视化绘制，以《新闻联播》对不同地区的播报次数，反映其对不同地区发展的重视程度和对其他不同类型国家的外交态度。在主题内容挖掘中，搭建基于潜在狄利克雷分配（LDA）模型提取主题、计算词频-逆文本频率指数(tf-idf)并利用指K均值聚类算法(K-Means)进行文本聚类，进而将不同新闻分类研究，剖析其主题构成情况，细化了对《新闻联播》主题的解构。在情感倾向分析中，本项目调用了百度飞桨生态下的中文NLP开源工具Senta构建双向长短期记忆网络（Bidirectional LSTM）模型进行情感得分预测，可以对新闻报道中的感情色彩进行评分，发现《新闻联播》作为官方媒体其鲜明的立场和阶级。

综上，本文利用以上角度切入《新闻联播》全文本数据，解构了《新闻联播》作为国内影响力最大、受众面最广的新闻类节目的语言特点，挖掘出了其作为官方宣传媒体在经济、政治、文化、外交等多领域的隐含的宏观信息。

**参考文献**

* 1. 艾红红,2008,《〈新闻联播〉研究》, 北京: 中国广播电视出版社
  2. 金海媚. 中美贸易战新闻语篇的批评话语对比分析[D].广东外语外贸大学,2020.
  3. 孟天广,郭凤林.大数据政治学:新信息时代的政治现象及其探析路径[J].国外理论动态,2015(01):46-56.
  4. 李小龙. 《新闻联播》关于突发事件报道的新闻话语与新闻观念探究（2003-2013）[D].广西大学,2014.
  5. 邵梓捷,戴遥遥,徐湘林.国际政治传播中的议程设置——基于2003-2015年《新闻联播》国际新闻文本分析[J].国家行政学院学报,2016(04):29-34.
  6. 邵梓捷,张小劲,孟天广.政治传播视角下《新闻联播》的宣传模式分析[J].清华大学学报(哲学社会科学版),2015,30(03):30-42
  7. 夏静华. 论《新闻联播》的公共危机报道（2003-2008）[D].苏州大学,2009.
  8. 朱辰雨.《新闻联播》新冠肺炎疫情报道中的国家认同建构[J].新闻研究导刊,2020,11(16):61-62.
  9. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

1. 中国社会科学院大学商学院，北京市房山区，邮编102488； [↑](#footnote-ref-1)
2. 央视广告经营管理中心，2020年初新闻类节目收视盘点：《新闻联播》收视率高达15%”，央视网，<http://1118.cctv.com/2020/03/05/ARTIXAYL6ihKJFxqI2RMYvC5200305.shtml>，2020-03-05 [↑](#footnote-ref-2)
3. 邵梓捷,张小劲,孟天广.政治传播视角下《新闻联播》的宣传模式分析[J].清华大学学报(哲学社会科学版),2015,30(03):30-42. [↑](#footnote-ref-3)
4. 艾红红,2008,《〈新闻联播〉研究》, 北京: 中国广播电视出版社 [↑](#footnote-ref-4)
5. 邵梓捷,张小劲,孟天广.政治传播视角下《新闻联播》的宣传模式分析[J].清华大学学报(哲学社会科学版),2015,30(03):30-42. [↑](#footnote-ref-5)
6. [Scikit-learn: Machine Learning in Python](http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), Pedregosa et al., JMLR 12, pp. 2825-2830, 2011. [↑](#footnote-ref-6)
7. 百度自然语言处理部-PaddleNLP, 文本情感分类使用文档<https://nlp.baidu.com/homepage/nlptools/document?f=%E6%96%87%E6%9C%AC%E6%83%85%E6%84%9F%E5%88%86%E7%B1%BB>，最后访问时间2020-11-10。 [↑](#footnote-ref-7)