# Python数据挖掘与可视化分析：东航坠机事件-微博话题舆情分析

## 一 、引言

随着经济社会的发展和科技的进步，互联网已深入 人们的日常生活，截至 2021 年 12 月，互联网普及率达 73%，较 2020 年 12 月提升 2.6%［1］。微博集娱乐、社 交、营销于一身，以其无门槛、传播快等特点深受大众 喜爱，其每天都会产生大量的数据，成为影响力日益增 强的网络舆论传播中心［2］。因此，搜集这些信息进行 分析，了解网民的立场和情感态度，对把握网络舆情有 一定的参考价值。 本文以情感倾向性分析为基调，引入情感维度分析不同情感状态的维度状况，从不同角度上细致地反应舆情情感状况，为网络舆情形势研判、预警处理和公关决策提供切实的依据。

## 二、相关研究

网络多媒体的广泛性和多平台的互通性，实现并满足了 众多网络用户的情感抒发诉求，而用户群体之间存在文化层 次不齐、思维方式多样等特征，使得网络舆情情感动态呈现 以下特征：①舆情主体需求存在同质化与差异化的统一。同 质化是指网络主体存在共同的需求和关注，包括个人、他人、 社会的认同和关注；差异化是指各个主体有自身的需求和不 同的关注对象。②网络舆情存在各类情绪的相互转化。而 各类情绪主要分为两种，即正面情绪与负面情绪，这两种情绪易受舆情的刺激，产生正负情绪的相互转化，尤其是正面 情绪转为负面情绪或是负面情绪极端化，都会造成舆情情感 的巨大波动，导致舆情信息的风向转变，产生恶劣的负面影 响。③网络舆情情感复杂化。不仅指情感类型的多样化，也 指情感的展现方式有显性和隐性之分。大多数情感能够展 现出直接的情绪、态度和意见，但有的须从舆情相关要素反 映出来，如舆情事件的热度和关注度等要素侧面表达出情感 偏向。④舆情情感的焦点和指向的快速扩散性和聚焦性。 充斥舆情的环境能快速传递不同情感，从起初的关注个人到 关注社会的焦点与指向变化，使得情感扩散化、极端化，舆情 随时产生演化，应及时提高预警反应。⑤网络舆情情感过程 遵循情感生命周期。一般情感的生命周期都呈抛物线状，是 情感复杂化和正负情绪转化直接导致的结果，生命周期根据 舆情事件的事发、事中和事后的发展，产生由较低程度快速 上升到情感巅峰爆发状态，并持续一段时间，随后转入较低 程度的情绪状态，最终情感会伴随舆情事件的淡化而平息。

## 三 、数据挖掘

1. 数据采集及处理

依托 Python 自定义编程，有目标地搜集微博话题 “东航坠机事件”的相关评论， 爬取 2022 年 3 月 21 日— 3 月 28 日之间的评论数据，共计 21054 条，进行数据清洗，最终得到有效数据 12666 条，保存到项目目录data/东航坠机.csv下

（二）数据分析

对有效数据进行高频词统计分析，如表 1 所示，除却东航坠机本身事件的相关讨论，“飞机”“搜救”“人员”“黑匣子”等词语也是网民讨论的热点。

表 1 高频词

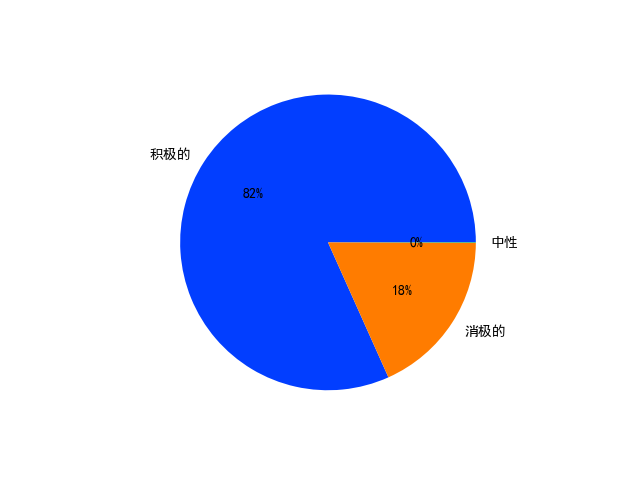
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排序 | 词汇 | 频次 | 排序 | 词汇 | 频次 |
| 1 | 现场 | 19432 | 11 | 东航 | 39300 |
| 2 | 了 | 16916 | 12 | 人员 | 10376 |
| 3 | 搜寻 | 9395 | 13 | 日 | 10351 |
| 4 | 的 | 77362 | 14 | 坠机 | 39729 |
| 5 | 和 | 11499 | 15 | 黑匣子 | 10830 |
| 6 | 微博 | 9959 | 16 | 工作 | 9245 |
| 7 | 在 | 25387 | 17 | 收起 | 10620 |
| 8 | 事故 | 24407 | 18 | 搜救 | 12712 |
| 9 | 飞机 | 15045 | 19 | 视频 | 12445 |
| 10 | 救援 | 15276 | 20 | 是 | 11583 |

## 四、实验分析

（一）情感分析

情感分数值在 0 ～ 1 之间，当结果大于 0.5，情感 较为积极，当结果小于 0.5，情感较为消极。对两个部 分中的每条评论进行情感打分，得到最终平均值，其中 关于东航坠机的情感平均值为 0.748。绘制情感分数波动图，有关 东航坠机 事件（见图 1）的评论大都偏积极。

图1



1. 词云

在对微博评论信息进行去停用词等处理后，得到 语料库，统计出现频率前 20 位的关键词得到表 1。可以看出网民评论最多的词语主要有“飞机”“搜救”“人员”“黑匣子”等等，充分体现了网民 伴随搜救进程网民最真实的心理反应状态，对搜救人员的尊敬、对遇难乘客的惋惜和痛心，这对心理学上研究和掌控社会舆情发展朝向积极乐观的态势发展，有重要的意义。 运用sytlecloud将得到的语料库进行高频词汇展 示，词云展示出现频率最高的 150 个词语，如图 2 所示

图2

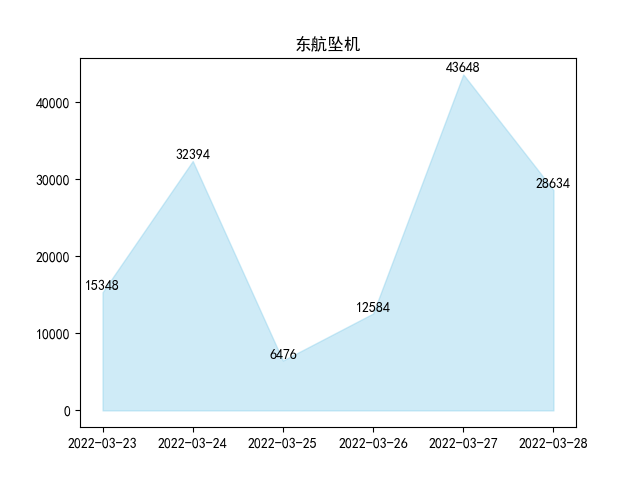
文本

描述已自动生成

1. 微博内容和评论的时间变化分析

考虑到每一天中深夜时间段人们相对在网络上不活跃，所以在既保证数据有效，又能得出时间与坠机发生关系基础上，结合活在爆发在网络上开始发布的 3 月 21 日 14 时，统计每一个时间段的微博话题“东航坠机“发布数量，得出图 3。

图3



## 五、结束语

本文针对“东航坠机”这一微博热点话题引起的舆情情况进行分析，利用 weiboSuperSpider 爬虫对话题数据进行爬取，运用 SnowNLP 对情感色彩和情绪态度进行分析，利用 Matlabplot 进行时间舆情分析，在使用 Stylecloud对高频词汇进行统计和展示，得出本次舆情发展过程中的发展情况，本论文的研究意义主要有以下几个方面：

（1）沿着时间线进行分析，本次舆情发展分为三个时期都有重要的现实意义：第一个时期是热点爆发期，这一阶段关注度陡升，迅速引起社会影响；第二个阶段是信息下降期，这一阶段的网民数量已经达到顶峰，政府可以进行网络上的信息安抚和虚假信息辟谣；第三阶段就是热度衰退后的淡化期，这一阶段伴随着搜救工作的基本完成，网民情绪等都变得缓和。

（2）对于东航坠机舆情的情感分析和关键词分析是很好的掌握事态发展的重要特征，体现了东航坠机爆发的过程中网民面对突发灾难时最真实的情感反应和对遇难同胞的同情与惋惜。这对于政府及时掌握舆情发展态势、安抚情绪、掌控大局有重要意义，并且为应对以后发生的相关事件提更了宝贵经验。

（3）此文的研究意义还体现在对待社会舆情热点信息处理和分析上提供了一种新的方法和解决方案，证明分布式爬虫对于热点话题爬取效率较高，SnowNLP也很好的反映了网民情绪是积极还是消极的，在对之后发生热点事件时，可以更有效、更客观、更直接的解释舆情的发展情况。

## 参考文献：

［1］ 中国互联网络信息中心 . 第49次中国互联网络发展状况统计

报告［EB/OL］. 中国互联网络信息中心，2022-02-25.

［2］ 纪雪梅 . 特定事件情境下中文微博用户情感挖掘与传播研究

［D］. 天津：南开大学，2014.

［3］ 邢云菲,王晰巍,韦雅楠,等.新媒体环境下网络舆情用户情感 演化模型研究——基于情感极性及情感强度理论[J].情报 科学,2018,36(08):142-148.［4］ 安璐，吴林 . 融合主题与情感特征的突发事件微博舆情演化分

析［J］. 图书情报工作，2017，61（15）：120-129.

［5］ 唐晓波，兰玉婷 . 基于特征本体的微博产品评论情感分析［J］.

图书情报工作，2016，60（16）：121-136.