公众讨论下的 DeepSeek—基于 BERTopic 的短视频文本主题分析

摘要： 杭州深度求索人工智能基础技术研究有限公司（DeepSeek）于 2025 年初推出了低成本、高性能推理模型 DeepSeek-R1，引发了科技圈震动和国内抖音等短视频平台的广泛讨论。尽管了解公共对人工智能态度愈发重要，DeepSeek 成功背后国内公众在短视频平台对其观点尚需分析；基于 BERTopic 的主题建模又为相关分析提供了一项成熟优质方案。因此，本研究使用 BERTopic 主题建模方法，分析了国内短视频平台公众对 DeepSeek 的观点看法。研究发现，DeepSeek 公众讨论可以总结为“模型股市影响”、“模型中英命名”、“模型使用反馈”、“国家民族情感”、“中外科技竞争”、“模型相关背景”、“模型技术原理”、“用户信息障碍”八类观点。基于上述发现，研究得以讨论 DeepSeek、人工智能的公众态度，和 BERTopic 在该类文本的任务实践、调整空间。

关键词： DeepSeek；短视频文本；BERTopic；主题建模

1. 引言

2025 年初，杭州深度求索人工智能基础技术研究有限公司（DeepSeek）发布了推理模型 DeepSeek-R1。该模型以低训练成本，与彼时最强推理模型 OpenAI o1 性能相当（张慧敏, 2025），在算力赛道下走出了创新性的技术路径。该事件引发了科技圈震动，也在国内抖音等社媒短视频平台激发了广泛的公众讨论。尽管了解公共对人工智能态度愈发重要（Kim、Lee, 2024），DeepSeek 成功背后国内公众在短视频平台对其观点尚需分析。对于相关分析，基于 BERTopic 的主题建模又提供了一条成熟优质的方案。因此，本研究旨在结合 BERTopic 主题建模，了解国内短视频平台公众对 DeepSeek 的看法。该探索一方面可通过 DeepSeek 短视频讨论的窗口丰富国内人工智能公众态度相关认知，一方面又是 BERTopic 在该类文本主题分析的实践运用。具体研究问题是：

BERTopic 主题建模下，抖音短视频中 DeepSeek 相关公众讨论反映了哪些关注焦点？

2. 文献综述

2.1. DeepSeek 与公众短视频讨论

近年来，全球人工智能竞赛将巨参数、大数据、强算力作为大模型研发的基础。OpenAI、微软、谷歌等科技公司囤积大量数据算力，主导相关行业。该背景下，杭州深度求索人工智能基础技术研究有限公司（DeepSeek）仅使用两千余块性能较低的 NVIDIA H800 GPU，在两个月内完成了六千余亿参数的混合专家模型训练，在 2025 年 1 月发布了推理模型 DeepSeek-R1——其性能与当时最强推理模型 OpenAI o1 相当，而训练成本仅为后者 3%-5%（张慧敏, 2025, p. 1）。可以说，DeepSeek-R1 通过架构算法、训练方式、工程优化等创新找到了差异化的技术路径。因此，该模型一经推出便引发了全球科技圈、技术市场的震动。同时，随着人工智能逐渐嵌入日常生活，DeepSeek 的成功也引发了国内社会广泛关注。其中，抖音等短视频平台作为当前国内领先的社交媒体平台，在传播相关消息后激发了大量的公众讨论。在相关人工智能快速发展、遍布生活的当下，调查公众对其的态度愈发重要（Kim、Lee, 2024, p. 9909）；同样，国内群众对于取得成功后 DeepSeek 的观点也尚待调查。因此，结合短视频的用户讨论度，有必要分析其中公众对 DeepSeek 的了解与看法。

2.2. BERTopic 主题建模

在文本分析上，主题建模是一项成熟的信息检索与自然语言处理方法。该技术基于数据降维和特征抽取，可以“扫描一组文档并检测其中的单词和短语模式，将文档集合中的词语规约到主题维度，从而达到高维数据降维的目的”，“同时主题中也包含了文档及其词语的潜在语义信息”来实现语义表达（逯万辉, 2024, p. 23）。经典的主题建模方式包括 LSA（潜在语义分析）、pLSA（概率潜在语义分析）和 LDA（隐含狄利克雷分布）；而在近年深度学习发展下，新生了 BERTopic（Grootendorst, 2022）等基于预训练词嵌入算法的主题建模技术。其中 BERTopic “采用基于 BERT 的深度学习预训练模型，结合 Sentence-Transformers 等嵌入模型和 c-TF-IDF 算法对句子进行编码与计算”（2024, p. 24），通过上下文更强的语义学习能力表现出更好的主题识别效果。具体来看，BERTopic 提供了六步可调整的算法框架：①文本嵌入；②嵌入向量降维；③降维向量聚类；④文档分词；

⑤主题词加权；⑥结果微调（可选）（Grootendorst, 2024）。首先，BERTopic 默认使用 Sentence-Transformers 对原始文档进行文本嵌入，生成以文档为行向量、特征为列向量的矩阵；接下来，BERTopic 默认使用 UMAP 对所得矩阵行向量（文档）在特征上降维；基于降维后矩阵，BERTopic 默认运用 HDBSCAN 对行向量（文档）聚类，生成文档的主题分组；为呈现每个主题的关键词，BERTopic 默认提前用 CountVectorizer 对原始文档分词；模型拟合后，BERTopic 默认使用 c-TF-IDF 加权计算并呈现每个主题的关键词；最后，用户可选择对结果呈现进行一定微调。上述框架中，BERTopic 在每一步均为使用者预留了模型和算法的修改空间，展现出良好的可调整性。可以说，该主题模型以较高性能、灵活架构为文本分析提供了较优路径。

2.3. 当前研究

综上所述，在 DeepSeek 的成功引发众多关注下，国内短视频平台上公众对其观点仍需调查；基于 BERTopic 的主题建模方法又能为相关文本分析提供较优方案。因此，本研究结合 BERTopic 主题建模，分析了国内短视频平台公众对 DeepSeek 的观点看法。具体研究方法见下。

3. 研究方法

3.1. 数据准备

参考公众群体、DeepSeek 话题讨论度，研究文本来自抖音平台相关短视频评论。本研究使用爬虫抓取了 2025 年 1 月 27 日至 1 月 29 日“中国大模型‘搅动’硅谷…”（凤凰卫视）、“…DeepSeek 超越 ChatGPT 登顶苹果 App 下载榜首…”（凤凰卫视）、 “火爆全网，国产 AI 震动美国科技界…”（央视新闻）、“3 分钟看懂为什么 DeepSeek 能震惊世界…”（澎湃新闻）视频评论，总计 3000 条。针对所得数据，使用正则表达式清除了“@用户名”、“表情”等无意义评论与其他评论内部的“@用户名”、“表情”部分。清理后总计评论 2870 条，以文本形式保存。

3.2. 数据处理与分析

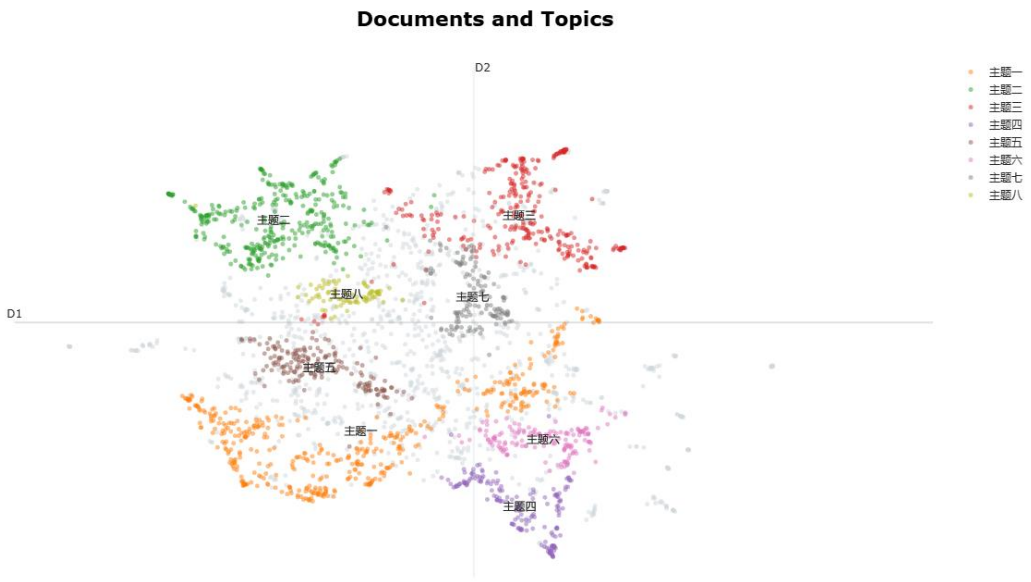
基于所得数据，本研究使用 BERTopic 进行主题分析并进行了适当模型、超参数调整。具体可分为文本嵌入、降维、聚类、分词、主题词加权、主题分析六

步，在 Python 3.11.9（Python Software Foundation, 2024）上运行。

首先，本研究使用基于 BERT 的哈工大模型“hfl/chinese-bert-wwm”（Cui 等, 2019）进行了文本嵌入。经对比，该模型对现有语料的主题识别比默认基于 SBERT 的“paraphrase-multilingual-MiniLM-L12-v2”模型更清晰：其一可能在于哈工大模型对中文语料进行了更好的预训练；其二，短视频评论较短，语义常集中于句内命名实体、关键表达——相较基于句嵌入的默认模型，基于词嵌入、采用全词掩码的哈工大模型可能更适合抓取文本细节信息。研究使用支持 512 个词嵌入单位的哈工大模型对原始语料每条评论截长补短，将每个词嵌入单位（多为单个汉字）转化为 512 维向量，并以包含全部嵌入单位信息的 512 维 CLS 标记向量作为每条评论的整体语义向量。堆叠全部评论语义向量后，得到形状为 2870*512（评论数*特征）的结果矩阵。

接下来，研究采用 UMAP 模型将所得矩阵行向量（评论）特征从 512 维降至 5 维。同时，在参数邻近范围为 15、最小距离为 0.0、距离度量用余弦相似度、随机种子为 42 时，研究在后续离群值控制和主题归类上效果较好。最终得到形状为 2870*5 的矩阵用以聚类。

得到降维矩阵后，研究运用 HDBSCAN 模型对行向量（评论）聚类，得到不同主题簇。在模型参数上，当最小簇大小为 50、最小样本数为 20、距离度量为欧氏距离时，结果离群值较少、主题归类效果较好。该步得到八个主题。针对聚类效果，研究使用 UMAP 将文本嵌入向量降至二维后，观察了评论文档散点的主题分布（见图一）。由图可见，文档散点可较清晰地分类到八个主题簇。



图一：评论文档主题聚类散点图

为用关键词表示各主题，研究使用 jieba 0.42.1（Sun, 2020）对原始语料分词。分词前，由于哈工大模型最大支持 512 个词嵌入单位，每条评论保留前 512 个字符。同时，研究还将“深度求索、幻化量方”等词加入预定义词典防止切分，清除了文内标点，用哈工大停用词表（哈尔滨工业大学, 2024）过滤了“的、是、太”等高频低意义表达，以保证一定关键词质量。

为选取每个主题的核心关键词，研究选用了 c-TF-IDF 算法为关键词加权。该算法以每个主题簇为单位，得到每个主题簇中独特并区别于其他主题簇的关键词。最终按 c-TF-IDF 值，获得每个主题的前十关键词。

基于上述结果，研究还保存了各主题所含具体评论、将主题关键词结果可视化，来解读分析各个主题。

4. 研究结果

结果共得到八个主题与各主题代表性关键词，如表一所示。针对研究问题，各主题反映了抖音短视频中 DeepSeek 的公众讨论分别围绕“模型股市影响”、“模型中英命名”、“模型使用反馈”、“国家民族情感”、“中外科技竞争”、“模型相关背景”、“模型技术原理”、“用户信息障碍”展开。

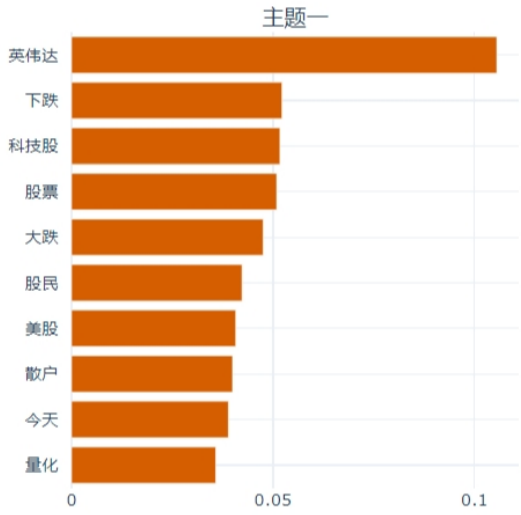
主题	评论数	关键词*
一	453	英伟达, 下跌, 科技股, 股票, 大跌, 股民, 美股, 散户, 今天, 量化
二	358	中文, 名字, 下载, 国内, 中国, ai, 英文, 是不是, 英文名, 英文名字
三	306	下载, 不了, 好用, 登录, 注册, 感觉, 软件, 一下, 豆包, 服务器
四	172	加油, 中国, 祖国, 科技, 骄傲, 世界, 智慧, 伟大, 除夕, 繁荣昌盛
五	168	中国, 美国, 打压, 大家, 没有, 国家, 不是, 西方, 真的, 现在
六	150	中国, 收割机, 量化, 模型, ai, 深度, 梁文峰, 公司, 真绿, 666
七	132	数据, 硬件, 模型, 显卡, 算力, 算法, 需要, ai, 先进, 芯片
八	80	看不懂, 听不懂, 听懂, 知道, 这是, 明白, 干嘛, 到底, 没看, 分钟
离群值	1051**	

*关键词使用 c-TF-IDF 算法计算。

**离群值占比 36.5%：HDBSCAN 聚类关注所得主题清晰度，该离群值可接受。

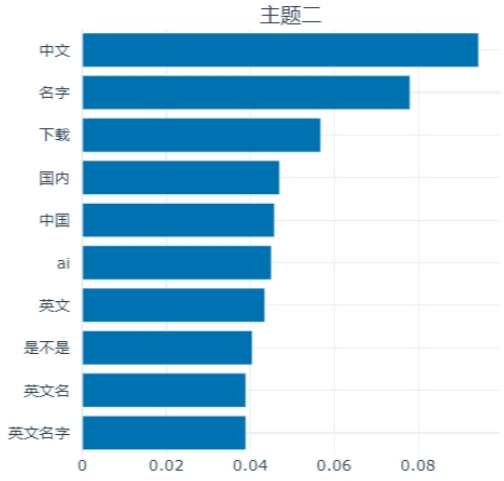
表一：主题建模结果

具体来看，主题一包括了“英伟达”、“下跌”、“科技股”、“股票”、“美股”等关键词（见图二）。这些关键词对应着 DeepSeek 受到全球关注后，其低训练成本技术路线冲击算力规则所引发的相关股市、以英伟达为代表的公司市值下跌。例如第 737 条评论“*英伟达夜盘现在已经大跌 8 个点*”和第 810 条评论“*科技跌疯了，劝大家不要炒股*”。这些评论反映了公众对 DeepSeek 的关注较多集中在股票这一生活密切相关领域。



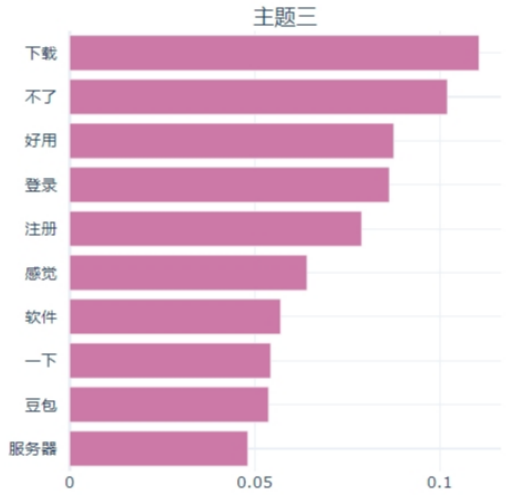
图二：主题一关键词

主题二则包括“中文”、“名字”、“英文”、“英文名”等关键词（见图三）。这些词实际反映的是 DeepSeek 宣传中英文命名引发的问题。第 198 条评论“*我只想问问中文名字叫什么？*”和第 172 条评论“*是中国的应该取中文名，这样才彰显中文的强大，现在的人做软件怎么老想取洋文！*”便反映了 DeepSeek 宣传时中文名称的普及度问题，与国内公众在其命名上的民族情感需求。这些评论体现了公众对 DeepSeek 的名称形象关注。



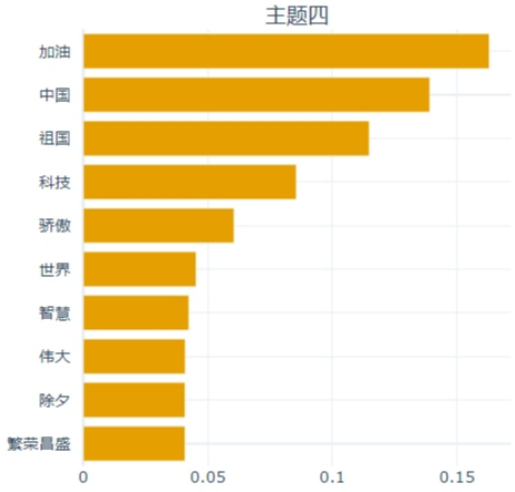
图三：主题二关键词

主题三关键词则包括了“下载”、“好用”、“登录”、“注册”等（见图四）。这些词都与用户使用 DeepSeek 的各类反馈评价相关。比如第 2076 条评论“*下载了，为什么登录不了，一直失败*”和第 688 条评论“*这个写论文还是不错的，非常好*”。这些评论既包括了用户遇到的具体使用问题，也涉及了他们在软件应用后的好评。



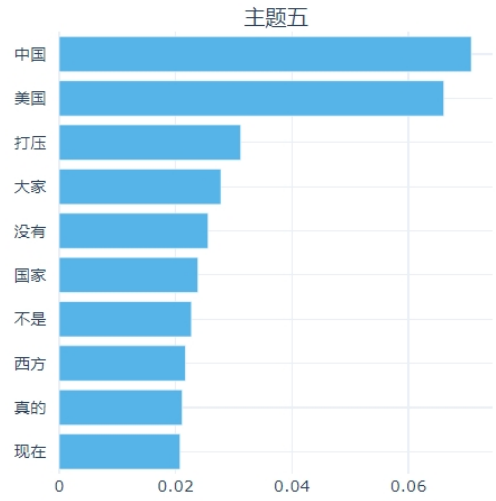
图四：主题三关键词

主题四则含有“加油”、“中国”、“骄傲”、“伟大”等关键词（见图五）。这些关键词反映了 DeepSeek 成功后公众对国家科技发展的自豪感。例如第 45 条评论“*前来留名，加油厉害了我的国！*”和第 721 条评论“*真的不错啊，科技为国争光*”。这些评论传达了公众的激动心情和为 DeepSeek 成功赋予的国家民族情感。



图五：主题四关键词

主题五围绕“中国”、“美国”、“打压”、“西方”等关键词展开（见图六）。这实际对应着当下存在的中外科技竞争。第 878 条评论“*这下老美更怕了。中国要是用最先进的芯片能搞出更恐怖的 ai*”和第 2558 条评论“*怎么打压都打压不住呢又是白菜价了美国又要喝西北风啦*”便能体现，DeepSeek 的成功让公众看到了当前竞争中国家科技的立足点和后续发展希望。



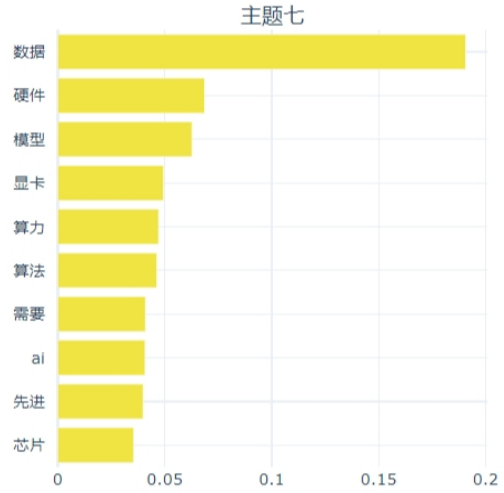
图六：主题五关键词

主题六与“收割机”、“量化”、“梁文峰（锋）”、“公司”等关键词密切相关（见图七）。这些关键词实则能联系到 DeepSeek 模型与公司的相关背景。比如第 1096 条评论“*最近震惊国内外大模型，来自大 A 收割机，量化大模型*”和第 2106 条评论“*广东小伙湛江人梁文峰（锋）中华好男儿*”。这些评论分别涉及了 DeepSeek 母公司幻方量化的公众评价和相关创始人梁文锋的个人信息，反映相关背景讨论。

图七：主题六关键词

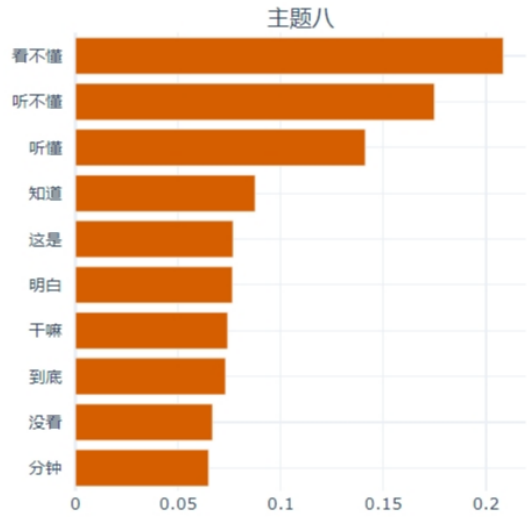
主题七关键词囊括了“数据”、“硬件”、“模型”、“算力”、“算法”等（见图八）。这些关键词则与 DeepSeek 技术原理内容相关。第 639 条评论“*就好比打十分靶，gpt 用一百支枪同时打(靠硬件堆积起来)，ds 只用一支枪命中(只用算法)，非常精简节约*”提及了 DeepSeek 的算法工程优势，而第 743 条评论“*硬件还是刚需，软件弥补只是暂时的，等别的 ai 也和你用同样的算法时，还是要拼算力*”

则进一步探讨了其技术路线的发展考量。该类评论体现了公众的技术思考。



图八：主题七关键词

主题八关键词则有“看不懂”、“听不懂”、“干嘛”等（见图九）。这些词表明 DeepSeek 在公众中仍有信息差。例如第 147 条评论“*根本看不懂谁能用大白话说这是干嘛的*”和第 248 条评论“*普通人还是看不懂啥意思*”。这些评论直接反映公众在了解和接触 DeepSeek 上仍有信息障碍。



图九：主题八关键词

5. 结论

综上，本研究使用 BERTopic 进行主题建模，挖掘了抖音代表下短视频中公众对 DeepSeek 的了解与看法。研究发现，DeepSeek 公众讨论可以总结为“模型股市影响”、“模型中英命名”、“模型使用反馈”、“国家民族情感”、“中外科技竞争”、“模型相关背景”、“模型技术原理”、“用户信息障碍”八类观点。从这些观

点可以发现，短视频平台 DeepSeek、人工智能的公众态度多围绕个体生活、民族情感层面展开，且前者在公众认知中仍有产品形象宣传、产品使用障碍、公众信息差等问题。另一方面，上述发现也说明了 BERTopic 在此类主题分析任务的适用性、有效性。当然，本研究在算法框架、实现路径上仍有较大可调整、提升空间，为后续研究提供了参考价值。

参考文献

[1] 哈尔滨工业大学. 哈工大停用词表[DS/OL]. 2024-1-25[2025-02-11]. https://github.com/goto456/stopwords

[2] 逯万辉. 科学文献主题建模方法及其效果评估[J]. 现代情报, 2024, 44(4): 22-31.

[3] 张慧敏. DeepSeek-R1 是怎样炼成的？[J]. 深证大学学报理工版, 2025, 1-7.

[4] Cui Y, Che W, Liu T, Qin B, Yang Z, Wang S, Hu G. Pre-Training with Whole Word Masking for Chinese BERT[J]. arXiv, 2019. https://huggingface.co/hfl/chinese-bert-wwm

[5] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure[J]. arXiv, 2022.

[6] Grootendorst M. BERTopic: The Algorithm[EB/OL]. 2024[2025-02-11]. https://maartengr.github.io/BERTopic/algorithm/algorithm.html

[7] Kim S, Lee Y. Investigation into the Infuence of Socio-Cultural Factors on Attitudes toward Artifcial Intelligence[J]. Education and Information Technologies, 2024, 29: 9907-9935.

[8] Python Software Foundation. Python 3.11.9[CP/OL]. 2024-04-02[2025-02-11]. https://www.python.org/downloads/release/python-3119/

[9] Sun J. jieba 0.42.1[CP/OL]. 2020-01-20[2025-02-11]. https://pypi.org/project/jieba/