一、语义网络体现关键词的逻辑关系。

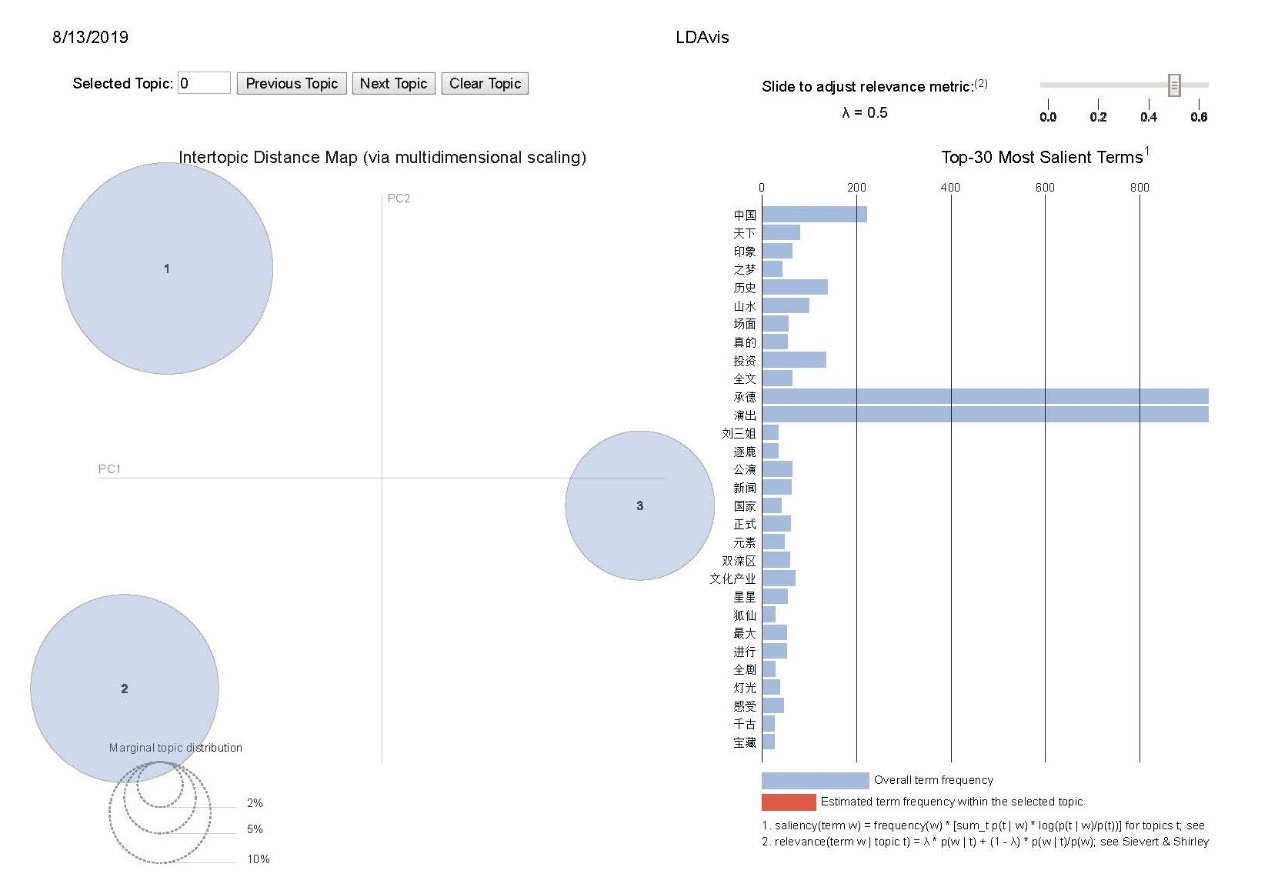


图 1 非营利性社交媒体网络平台评论LDA聚类图

Fig.1 LDA Clustering Map of Non-profit Social Media Platform Websites’ Comments

如上图所示，从非营利性社交媒体网络平台(新浪微博)上爬取有关“鼎盛王朝”话题的文本数据，经LDA算法处理后共聚成3类，各类之间并无交叉。对于整个文本来说，“承德”和“演出”的重要性（Saliency）最大，但其对文本话主题的贡献（Relevance）较小；相反“中国”的重要性（Saliency）较小，但其对文本话主题的贡献（Relevance）最大。

表 1 官方推文LDA 模型下的关键词重要性(Saliency)表

Tab.1 Saliency Table Based on LDA Model of Official Tweets

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类一 | | 类二 | | 类三 | |
| 鼎盛 | 1650 | 康熙 | 1648 | 天下 | 56 |
| 大典 | 1400 | 鼎盛 | 1648 | 印象 | 45 |
| 演出 | 1260 | 王朝 | 1647 | 王朝 | 1648 |
| 王朝 | 1648 | 大典 | 1560 | 之梦 | 34 |
| 康熙 | 1648 | 承德 | 1123 | 鼎盛 | 1645 |
| 承德 | 1060 | 演出 | 1369 | 全文 | 33 |
| 实景 | 700 | 实景 | 710 | 刘三姐 | 24 |
| 大型 | 400 | 震撼 | 36 | 逐鹿 | 22 |
| 文化 | 360 | 场面 | 25 | 国家 | 23 |
| 旅游 | 250 | 真的 | 25 | 山水 | 41 |
| 我们 | 236 | 元素 | 22 | 狐仙 | 16 |
| 投资 | 180 | 演绎 | 24 | 历史 | 76 |
| 中国 | 287 | 传奇 | 23 | 全剧 | 15 |
| 皇家 | 289 | 灯光 | 18 | 宝藏 | 14 |
| 文化产业 | 56 | 户外 | 18 | 千古 | 14 |
| 公演 | 57 | 电话 | 18 | 天门 | 13 |
| 新闻 | 23 | 舞台 | 20 | 冰雕 | 13 |
| 正式 | 22 | 大帝 | 28 | 中国 | 114 |
| 元宝山 | 22 | 盛世 | 24 | 西安 | 12 |
| 双栾区 | 21 | 一生 | 23 | 梅帅元 | 11 |
| 星星 | 18 | 效果 | 19 | 情怀 | 11 |
| 最大 | 16 | 感觉 | 18 | 禅宗 | 11 |
| 进行 | 16 | 值得一看 | 18 | 少林 | 12 |
| 河北 | 20 | 非常 | 13 | 表现 | 10 |
| 演出季 | 19 | 看到 | 14 | 艺术节 | 12 |
| 游学 | 14 | 晚上 | 22 | 丽江 | 12 |
| 观看 | 25 | 大型 | 48 | 感受 | 18 |
| 承德市 | 23 | 这个 | 26 | 康熙 | 1645 |
| 正在 | 18 | 好看 | 15 | 音乐 | 13 |
| 第三季 | 17 | 宏大 | 14 | 天命 | 8 |

表注释：类一：文化旅游；类二：演出实景；类三：国家历史。

如上表所示，类一中，具有属性意义的名词中，“文化”和“旅游”的重要性较强，且其与类一的相关性也较强。同比，类二重要性及相关性较强的词分别为“演出”和“实景”；类三重要性及相关性较强的词分别为“国家”和“历史”。

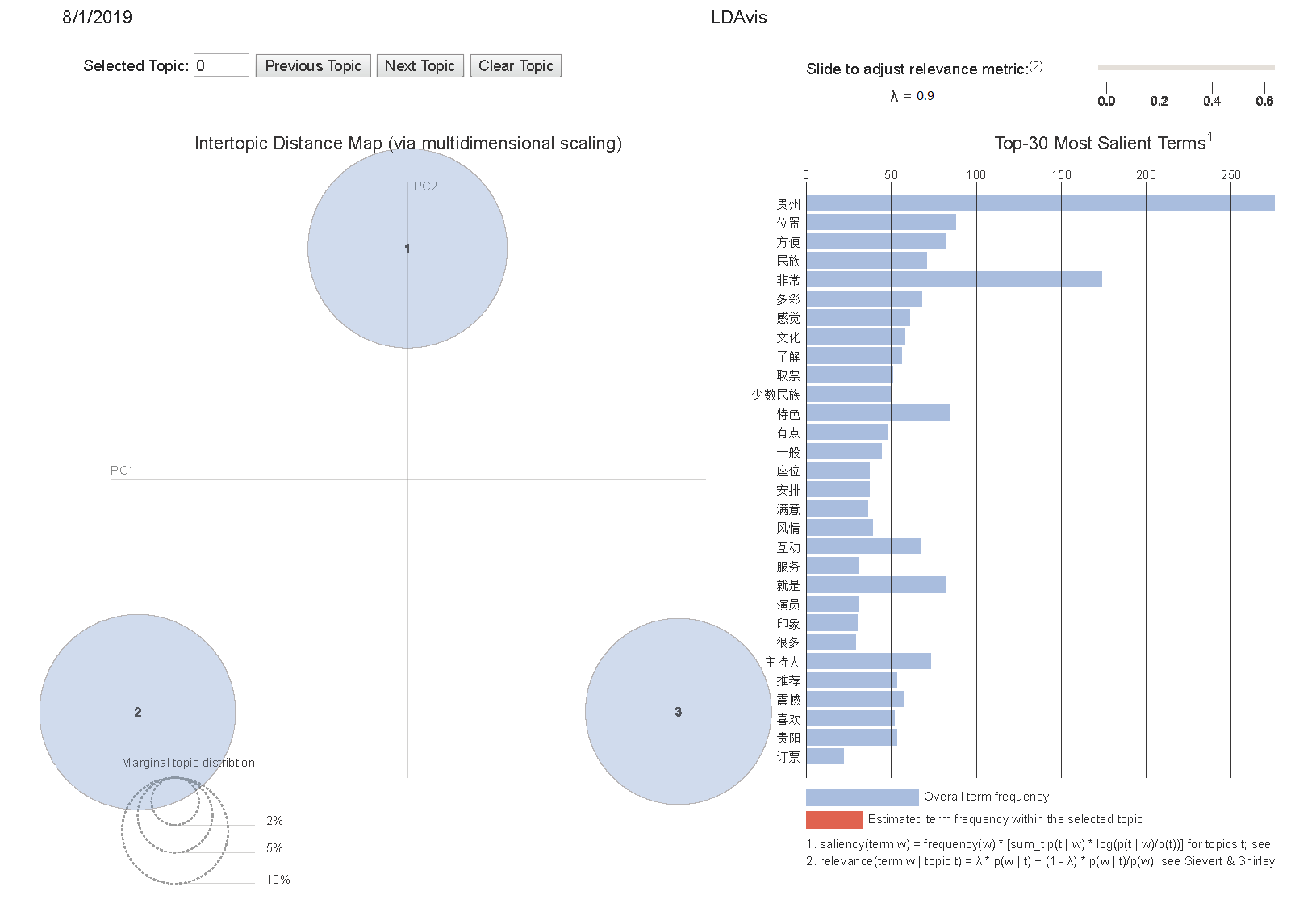


图 2 营利性社交媒体网络平台评论LDA聚类图

Fig.2 LDA Clustering Map of For-profit Social Media Platform Websites’ Comments

如上图所示，从营利性社交媒体网络平台上爬取有关“鼎盛王朝”话题的文本数据，经LDA算法处理后共聚成3类，各类之间并无交叉。对于整个文本来说，“不错”的重要性（Saliency）最大，且其对文本主题的贡献（Relevance）也最大。

表 2 旅游网站评论LDA 模型下的关键词重要性(Saliency)表

Tab.2 Saliency Table Based on LDA Model of Tourism Websites’ Comments

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 类一 | | 类二 | | 类三 | |
| 不错 | 261 | 康熙 | 188 | 晚上 | 89 |
| 非常 | 233 | 历史 | 65 | 知道 | 45 |
| 震撼 | 480 | 舞台 | 63 | 酒店 | 45 |
| 就是 | 130 | 演出 | 489 | 自己 | 44 |
| 演出 | 480 | 灯光 | 92 | 一定 | 68 |
| 感觉 | 108 | 大型 | 54 | 坐垫 | 41 |
| 还是 | 78 | 震撼 | 488 | 衣服 | 41 |
| 好看 | 78 | 实景 | 112 | 露天 | 39 |
| 值得 | 143 | 王朝 | 68 | 工作人员 | 38 |
| 有点 | 79 | 一生 | 45 | 建议 | 56 |
| 值得一看 | 136 | 江南 | 34 | 超级 | 29 |
| 真的 | 77 | 大帝 | 33 | 不要 | 26 |
| 方便 | 76 | 文化 | 26 | 主要 | 26 |
| 比较 | 75 | 故事 | 24 | 外套 | 25 |
| 很多 | 75 | 鼎盛 | 72 | 已经 | 27 |
| 取票 | 74 | 场面 | 187 | 承德 | 152 |
| 表演 | 132 | 大典 | 68 | 门口 | 27 |
| 没有 | 84 | 演员 | 88 | 赞赞 | 23 |
| 场面 | 184 | 讲述 | 44 | 毛毯 | 22 |
| 壮观 | 102 | 背景 | 45 | 这里 | 23 |
| 位置 | 68 | 宏大 | 78 | 还行 | 22 |
| 可以 | 108 | 剧情 | 44 | 携程 | 21 |
| 觉得 | 66 | 演绎 | 43 | 一场 | 26 |
| 还有 | 65 | 体验 | 44 | 提前 | 26 |
| 现场 | 66 | 看过 | 56 | 可以 | 106 |
| 节目 | 68 | 皇家 | 40 | 演出 | 396 |
| 但是 | 69 | 声光电 | 39 | 因为 | 41 |
| 时间 | 56 | 奔腾 | 38 | 山庄 | 26 |
| 这个 | 89 | 设计 | 41 | 军大衣 | 27 |
| 场景 | 91 | 音乐 | 35 | 网上 | 26 |

表注释：类一：演出节目；类二：历史文化；类三：场地设施。

如上表所示，类一中，具有属性意义的名词中，“演出”和“节目”的重要性较强，且其与类一的相关性也较强。同比，类二重要性及相关性较强的词分别为“历史”和“文化”；类三重要性及相关性较强的词分别为“露天”和“坐垫”等。**2 相关研究综述**

**2.1 LDA**

LDA模型是基于Dirichlet先验分布无监督机器学习模型。对于文本数据的主题挖掘，每个单词或词组都相当于基于以单个文档为单位，语序逻辑顺序排列的多项式矩阵，基于LDA模型的研究方法用以挖掘文本数据中隐含中心话题及主题词相关词，该模型已成为目前文本挖掘及主题提取的主要方法[1]。LDA模型联合概率核心公式表现为公式（1）[2]：

(1)

设定提取主题目标个数，文档总数，第个文档中的词个数，和分别表示每个文档下主题的多项分布的Dirichlet先验分布和每个主题下词的多项分布的Dirichlet先验分布，和分别表示为第个文档的第个主题以及第个文档的第个词[2][3]。

Putri和Kusumaningrum（2017）使用LDA二分类模型对有效旅游评论信息进行正、负向情感分类，并通过对每类的关键词提取，实现印度尼西亚顾客满意度调查研究[4]。

**3 实验设计**

**3.1 LDA 聚类及主题提取**

本研究使用Python及SQL编程语言，结合scrapy结构模型、Jiaba、Gensim、PyLDAvis和matplotlib等模块，进行网络数据的爬取、预处理、LDA机器学习模型的建立及结果的可视化。LDA主题提取步骤如下：

使用Jiaba实现中文分词，基于爬取到的评论数据建立语料库。

去掉出现次数多于10%、评论条数少于5的词及其他停用词。

使用Gensim建立LDA模型并输入主题数目，迭代次数等参数。

使用PyLDAvis实现聚类结果可视化，提取每类重要性前30的关键词，并根据相关性排序。

重要性计算公式如公式（2）[5]所示，相关性计算公式如公式（3）[6]所示。

（2）

（3）

其中，表示该词在本类中的重要性，表示该词在本类中出现的频率；的条件概率以及对隐藏主题的边际概率分别为和。表示词与隐藏主题的相关性,为权重参数（）。和分别与重要性及相关性呈正相关。

本实验中，选取聚类数目3个，。

**Reference**

[1] Naili, M., Chaibi, A. H., & Ghézala, H. B. (2016). Empirical study of LDA for arabic topic identification. In *Proceedings of CARI* (p. 139).

[2] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, *42*(1-2), 177-196.

[3] Naili, M., Chaibi, A. H., & Ghézala, H. B. (2016). Empirical study of LDA for arabic topic identification. In *Proceedings of CARI* (p. 139).

[4] Putri, I. R., & Kusumaningrum, R. (2017, January). Latent Dirichlet allocation (LDA) for sentiment analysis toward tourism review in Indonesia. In *Journal of Physics: Conference Series*(Vol. 801, No. 1, p. 012073). IOP Publishing.

[5] Chuang, J., Manning, C. D., Heer, J. Termite: Visualization techniques for assessing textual topic models[M]. In Proceedings of the international working conference on advanced visual interfaces,2012: 74-77.

[6] Sievert, C., Shirley, K. LDAvis: A method for visualizing and interpreting topics[M]. In Proceedings of the workshop on interactive language learning, visualization, and interfaces, 2014:63-70.