可视语言与信息可视化

团队 ID:

刘京宗 3019213043

杨朝涵 3020244160

张雪雅 3020244317

天津大学, 智能与计算学部

2023 年 1 月 6 日

目 录

摘 要

大作业要求大家按照论文短文的格式进行书写，参考文献 [1, 2]。

1 引言

Figure 1: 背景

2014 年 1 月 23 日，阿比拉发生了多起事件。系统已要求您根据发生的有限信息进行回顾性分析。您的目标是确定风险以及如何更有效地缓解风险。

您可以访问包含两个主要来源的单个数据流：

由自动筛选器识别为与正在进行的事件潜在相关的微博记录

阿比拉，克罗诺斯岛地方警察和消防部门的紧急调度文字记录。

根据这些数据，您可以评估公众不断变化的风险水平并提出建议的措施吗？

您还可以访问阿比拉地图和背景文件。（注意：这些是“迷你挑战 1”和“迷你挑战 2”中提供的相同材料）

使用视觉分析来分析可用数据并制定对要提供的问题的响应。此外，准备一段视频，展示您如何使用视觉分析来解决这一难题。

2 相关工作

调研相关论文发表，搜索 CNKI 或者 Google scholar 等学术引擎，了解该领域研究现状。参考文献格式为 [1]。

3 问题描述和需求分析

使用视觉分析来表征数据集中不同类型的内容。什么是有意义的事件报告与典型的垃圾邮件或垃圾邮件区别开来？请将答案限制为 8 张图片和 500 个单词。

使用可视化分析来表示和评估在晚上的过程中对公众的风险程度如何演变。考虑这种情况的潜在后果以及可能受到影响的人数。请将答案限制为 10 张图片和 1000 字。

如果您能够将一组急救人员发送到任何地方，它将在哪里？提供您的理由。如果您必须实时而不是回顾性地响应事件，那么您的响应会有什么不同？请将答案限制为 8 张图片和 500 个单词。

4 解决方案

4.1 任务一

(1) 借助 Python 处理数据，提取 message 中的 tag，并且统计每个 tag 出现的次数，据此绘制词云。

(2) 借助 Python 处理数据，对 message 进行分类，借助散点图和柱状图表征数据集中不同类型的内容。

(3) 使用 tf-idf 算法，对不同类型的 message 提取关键词，并且用 TreeMap 展示。

(4) 统计不同类型的 message 的作者信息，同样用 TreeMap 进行展示。

4.2 任务二

4.3 任务三

5 实验结果和案例分析

5.1 任务一

统计不同 tag 在 message 中出现的次数，如图 2 所示。

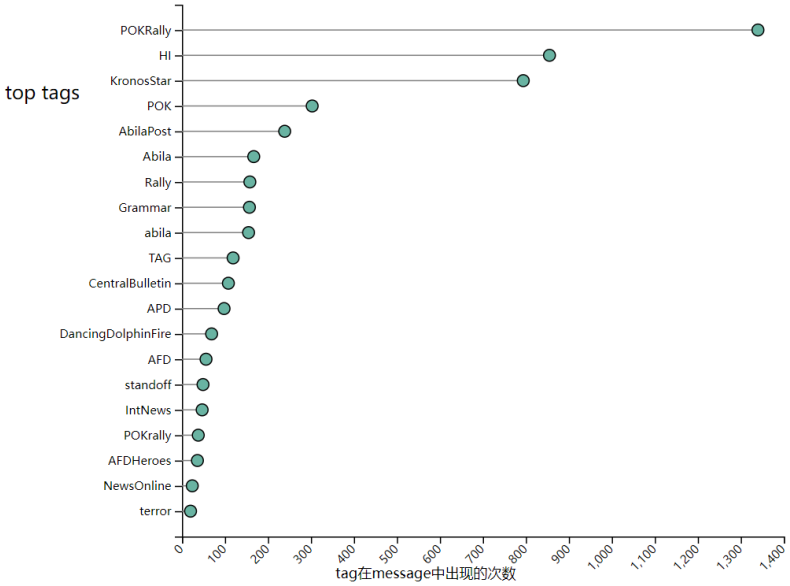

Figure 2: 不同 tag 在 message 中出现的次数

Figure 3: tag 词云

通过这些 tag，我们可以对 message 的主题有一个把握，并且词云的形象化展示，让我们对数据有了更直观的认识。

借助 Python 处理数据，对 message 进行分类，具体如表 1 所示。

表 1: 不同类别的 message 分类说明

种类	说明
unrelated	与报道完全无关的消息。主要由用户 @KronosQuoth 和 @Clevvah4Evah 发出。二人的消息主要是一些“心灵鸡汤”、励志格言，与报道内容完全无关，共 1418 条。
advertisement	广告。其内容中往往包含了网页链接，如“I recommend this site #abila dates.kronos/clickhere”借助正则表达式'.*.*/' 进行筛选，共 227 条。
chatter	闲聊。这部分内容主要为用户的闲聊，主要特征为通常以“RT @”开头，共 1006 条
report	报道。主要为当地媒体对新闻的报导，如关于火灾，枪击的报道。主要由 @AbilaPost, @megaMan 等用户发送，共 424 条。
others	其他。不能被以上四类所典型概括的消息，统一归为其他，共 988 条。

在对 message 进行分类后，我们采用散点图和柱状图表征数据集中不同类型 message 的内容.

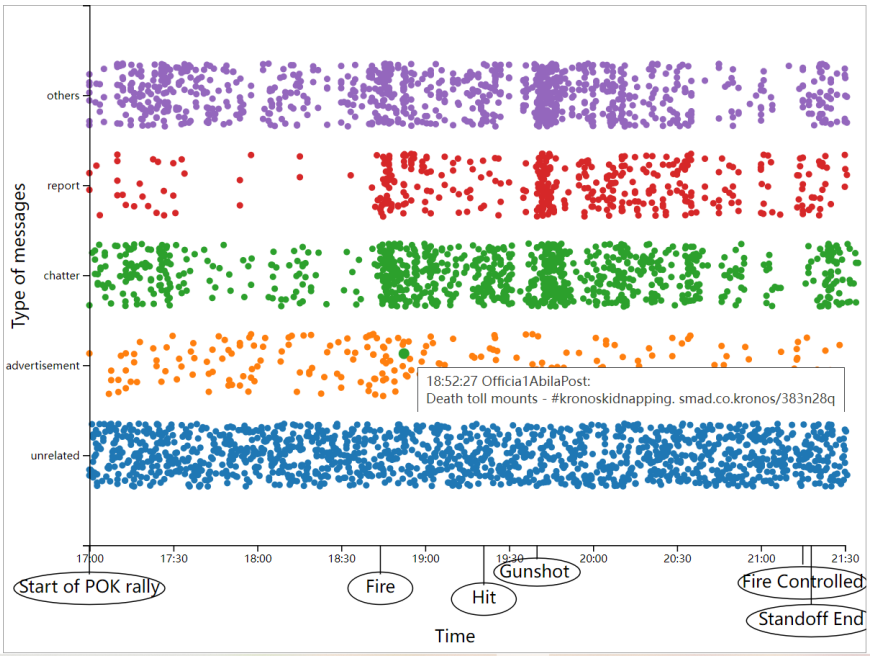

Figure 4: 散点图

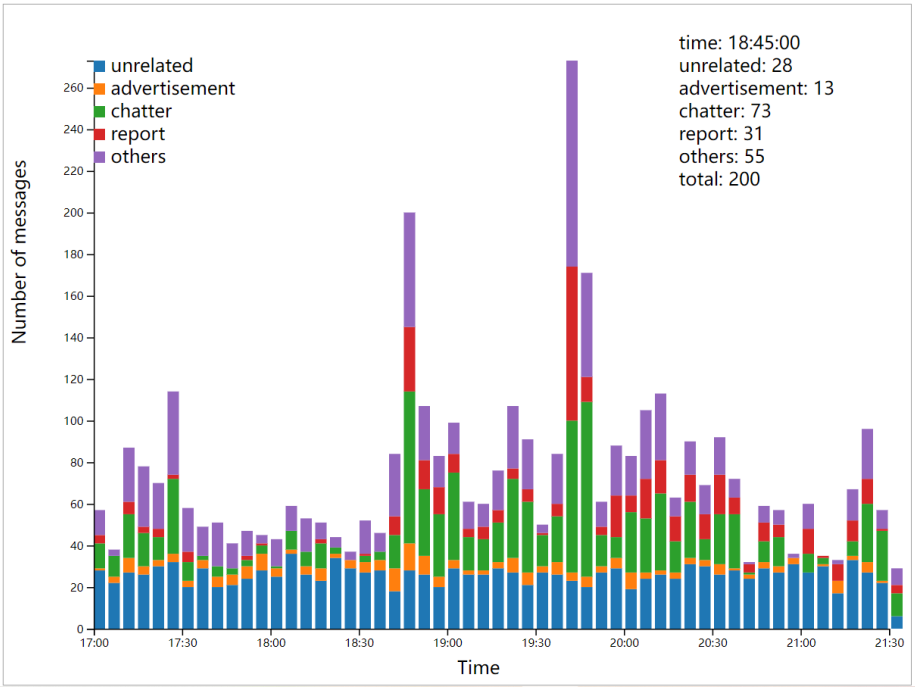

Figure 5: 柱状图

通过散点图和柱状图，我们可以对不同类型的 message 随时间的变化有一个直观的认识。且我们能直观感受到有重大事件（如图 4 横轴所示）发生时，chatter 和 report 类型的消息数量会有一个明显增加。且我们为这两幅图增加了一定的交互性，当鼠标悬停在散点图中的点上时，会有颜色变化，同时显示出该点对应的 message 信息，包括时间与作者信息。当鼠标悬停在柱状图中的柱子上时，会显示出该柱子对应的时间段以及该时间段内的各种 message 的数量。

之后，我们使用 tf-idf 算法对 message 进行关键字提取，使用 TreeMap 对这些关键字进行可视化，通过关键词简洁明了地展示不同类别的 message 的内容。类似的，我们统计了不同类别的 message 的作者信息。同样选择使用 TreeMap 对这些信息进行可视化。具体如 6 和 7 所示。

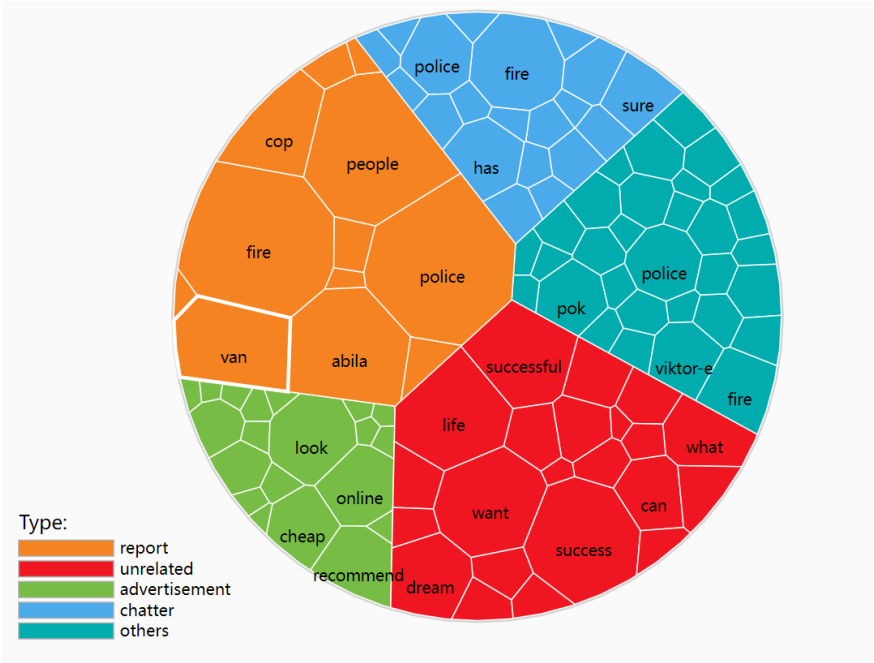

Figure 6: 关键词 TreeMap 可视化

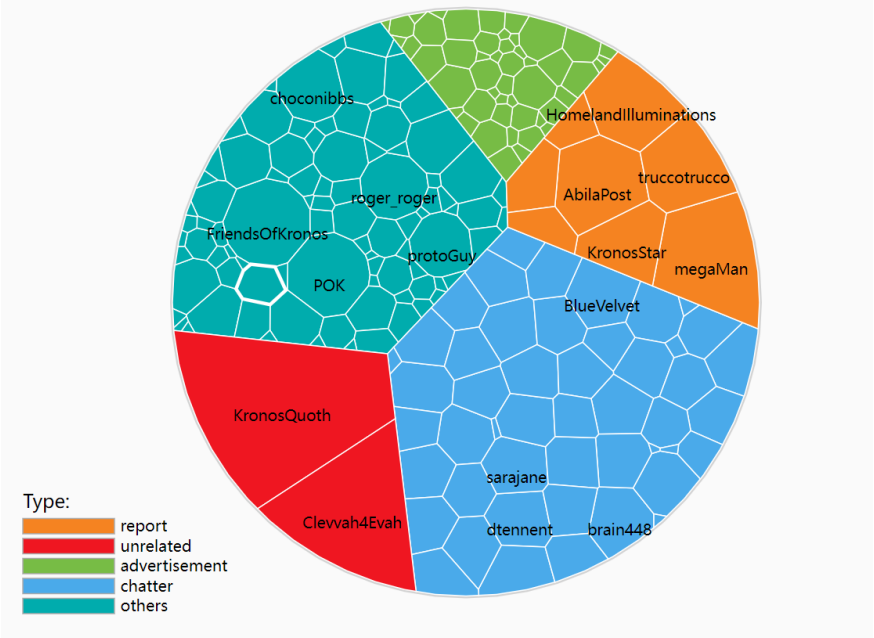

Figure 7: 作者 TreeMap 可视化

在上述两幅图中，我们选择展示面积占比较大的关键词和作者，以便于我们对数据集中的内容有一个直观的认识。对于占比较小的关键词和作者，只有鼠标悬浮在其上时才会显示出其具体信息。

5.2 任务二

5.3 任务三

6 总结

刘京宗：

在本次实验中，我主要负责任务一的完成。对于我这部分内容来说，在进行可视化之前，需要对数据集进行预处理，这部分工作花费了我较多的时间。但这部分工作一方面让我熟悉了 Python 的一些语法和相关库的使用，也是后续可视化工作的基础。在进行可视化工作时，我还是对 D3.js 的接口不够熟悉，这也是我在本次实验中遇到的最大的困难。但在查阅相关资料后，我对 D3.js 有了一定的了解，也完成了本次实验的任务一。最后通过撰写文档，我对 Latex 的使用也有了一定的了解。总的来说，本次实验让我收益匪浅。

参考文献

[1] R. G. Bayrak, N. Hoang, C. B. Hansen, C. Chang, and M. Berger. Pragma: Interactively constructing functional brain parcellations. *arXiv preprint arXiv:2009.01697*, 2020.

[2] P. Govyadinov, T. Womack, J. Eriksen, D. Mayerich, and G. Chen. Graph-assisted visualization of microvascular networks. In *2019 IEEE Visualization Conference (VIS)*, pages 1–5. IEEE, 2019.