可视语言与信息可视化

团队 ID: 19

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 8 日

目 录

摘 要

在 2014 年 1 月 23 日，阿比拉发生了多起事件。为了确定风险并更有效地缓解风险，系统要求进行回顾性分析。本文对两个主要来源的数据流——微博记录和紧急调度文字记录进行了分析，并对分析结果进行了可视化展示，并据此制定应对方案。具体来说，我们通过依次解决以下三个任务来完成本文的工作。

首先，我们对微博记录进行了分析，我们统计了微博记录中的 tag 出现情况并且借助词云对其进行了可视化展示。并且根据记录对消息进行了分类，借助散点图、柱状图等对消息进行了可视化展示。此外借助 tf-idf 算法对消息进行了关键词提取，对提取的后关键词以及作者信息借助 TreeMap 进行了可视化展示。通过多种可视化手段，我们识别了有意义的事件报告。

其次，我们在上一步的基础上，在消息出现的高峰期划分时间段，通过 QGIS 创建人群分布的地理地图。使用 D3 依据关键词随时间的变化绘制饼图，通过关键词所占比例，并参考提取数据中的 message 内容，评估了公众的风险水平在晚上如何演变。

最后我们依据事件绘制事件流图，根据整晚的事件信息，我们分析出急救人员应被派往火灾发生地附近，应及时救治在火灾、坍塌、爆炸事件中受伤人群。如果要求必须实时而不是回顾性地响应事件，我们通过关键词比例分布饼图，分析出急救人员将被派往警匪枪战对峙事件发生地附近。

综上，我们逐步推进分析，最终完成了确定风险以及缓解风险的目标。

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

作为一门问题和目标主要来自于现实世界的学科，数据可视化在很多领域获得了研究、应用和长足进步，在从研究范围（广度）、研究精细化（深度）不断拓展学科边界的过程中，逐渐收敛成为若干热点领域。本文以数据特征来划分，介绍其中的 3 类技术。

2.1 时间序列数据可视化

时间序列可视化随着时间的发展采集信息数据，运用可视化技术手段进行呈现，呈现出的可视化方式主要有 3 种。一是线形图，通过最开始的点展示不同时间段信息数据变化，在可视化过程中信息数据呈现较多时间维度，根据不同维度建立相应图标进行排列，观察数据的变化 [1]；二是堆积图，这类图主要对所有时间序列进行叠加，出现负数时，堆积图无法处理所有的时间序列，极大程度降低了可视化的呈现效果；三是地平线图，随着时间变化清楚地观察到信息数据的变化率，颜色的深浅表示正向、负向的变动效果 [2, 3]。

2.2 网络数据可视化

网络数据可视化技术手段核心是自动布局算法，将信息数据通过自动布局、计算，绘制成网状结构的图形。应用较广泛的有 3 类。力导向布局：借助力的概念，连接受力节点绘制网状图，由于互斥力的存在，可以减少节点间的重叠，适用于描述事物之间的关系，例如计算机网络关系、社交网络关系等各类关系网络情景 [4, 5]。圆形布局：将所有节点自定义排序，按照顺序在圆形上排列出来，快速分析出结果，受限于屏幕大小，节点数量较多时，圆形半径越来越大，难以直观显示全部节点，适用于查找较多关联关系的节点场景，例如在圆形布局图中可明显分辨出哪些节点关联关系较多。网格布局：采用网格设计方式绘制网格状信息数据网状图，适用于分层网络，利于观察整体层次 [6]。

2.3 层次信息数据可视化

层次结构常被用来描述具有明显层次结构的对象，包括图书馆标签、计算机层次系统或者面向对象程序类之间的继承关系等 [7, 8]。层次信息数据可视化用到的方法主要包括节点连接、空间填充、混合方法等。节点连接主要绘制不同形状节点表示信息数据内容，节点之间连线表示数据之间的关系。此类层次代表技术有空间树、圆锥树等。空间填充主要运用包围框表示层次结构信息数据，上层节点与下层节点之间包围关系表示信息数据间的结构关系 [9]。此类层次代表技术有树图、信息立方体等。混合方法结合多种可视化技术优点，使认知行为更加高效 [10]。此类方法代表技术有弹性层次层次网等。

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

(1) 借助 Python 处理数据，去除无关项与停顿词，提取关键词随时间变化的数据。

(2) 借助 Python 构建词向量，提取不同时间段内，与关键词相关的 message 数据。

(3) 借助任务一处理结果可以发现，大约在下午 6:50 和 7:50 左右，帖子数量处于高峰期，以此划分时间段，通过 QGIS 创建人群分布的地理地图。

(4) 使用 D3 依据关键词随时间的变化绘制饼图，通过关键词所占比例，并参考提取数据中的 message 内容，评估公众的风险水平在晚上如何演变。

4.3 任务三

(1) 提取不同时间段关键词，依据此关键词得出此时段事件，依据事件绘制事件流图。

(2) 借助 Python 处理数据，筛选出所有含有 latitude 和 longitude 的数据，再筛选出与特定事件即 dancing dolphin fire 有关的数据，据此绘制出 latitude-longitude 坐标图。

(3) 借助 Python 处理数据，筛选出所有含有 location 的数据，再筛选出含有关键字 fire 的数据，据此绘制词云。

(4) 提取指定时间段关键词，选择出现次数多且有意义的关键词数据，据此绘制关键词饼图。

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。

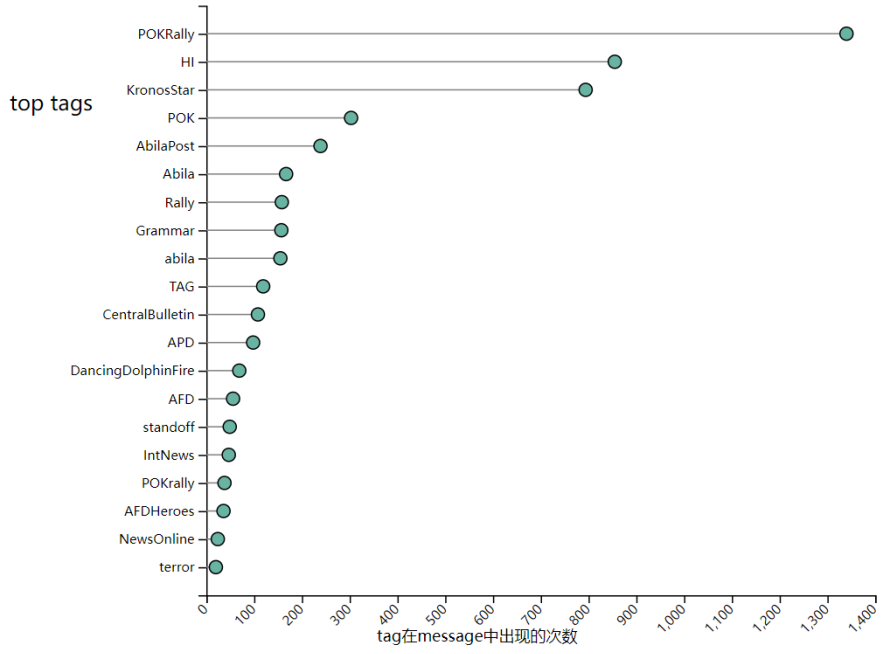

Figure 2: 不同 tag 在 message 中出现的次数

Figure 3: tag 词云

Hashtag(#) 用于将消息与特定主题相关联。例如 #POK 的标签显然与 POK 组织有关，#AbilaPost 的标签则表明该消息当地的媒体报告有关。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.

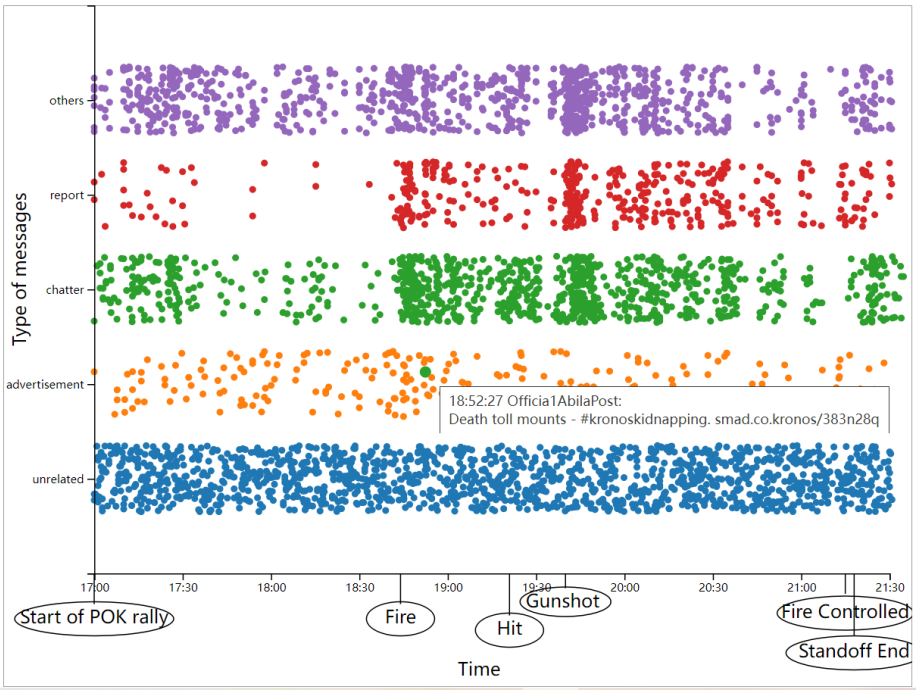

Figure 4: 散点图

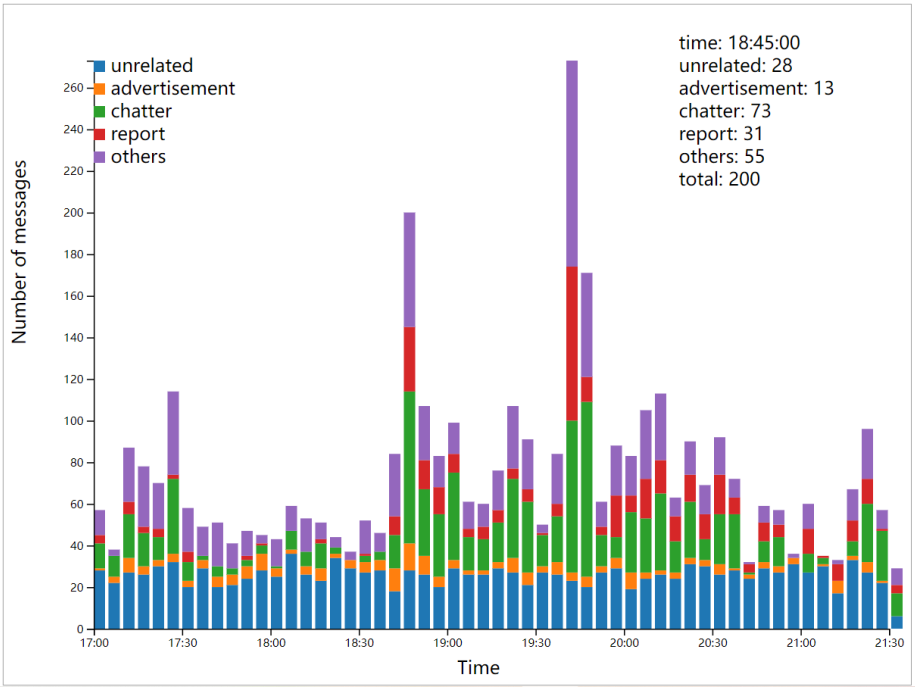

Figure 5: 柱状图

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。

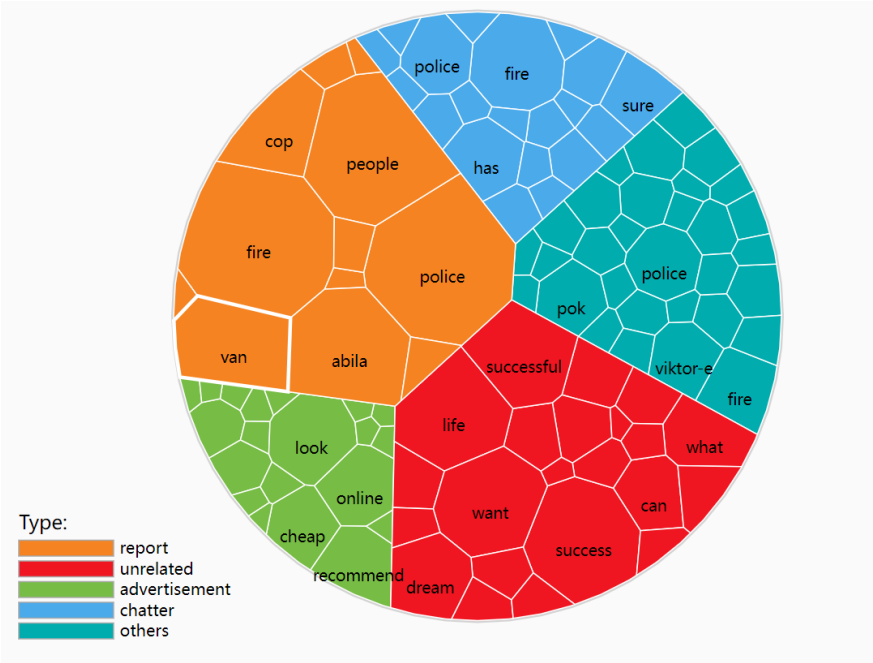

Figure 6: 关键词 TreeMap 可视化

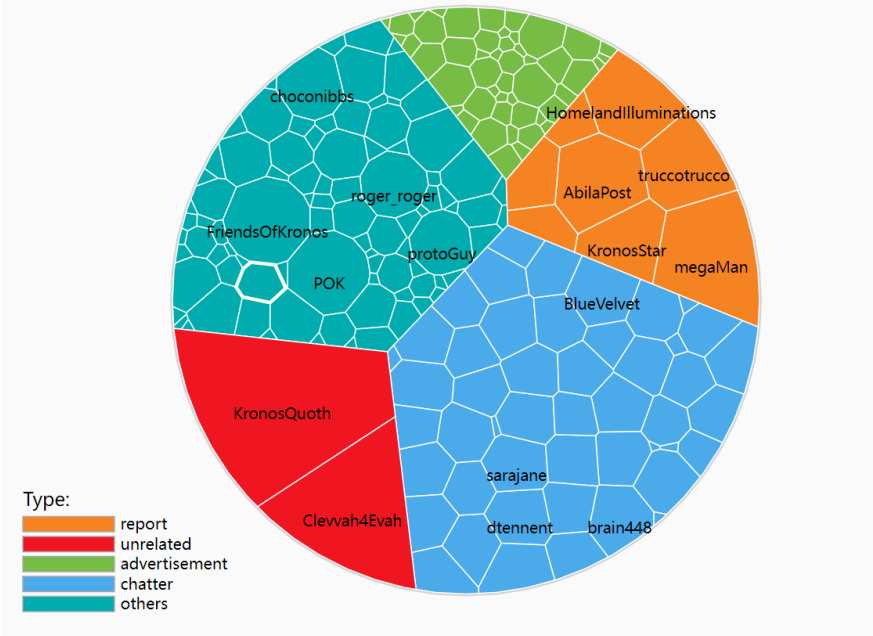

Figure 7: 作者 TreeMap 可视化

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。从中，我们不难获得一些有用的信息，例如我们最为关注的 report 类的信息中，“police”，“fire”，“van”等词概括性地表明了当晚的事件，且事件的报道的主要媒体是 @AbilaPost, @megaMan 等。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

根据任务一得到的结果可以发现，大约在下午 06：30 到晚上 07：00 和晚上 07：30 到晚上 08：00 的时间帧之间，帖子的数量处于高峰期，说明在这些时间段很明显发生了有意义的事件。据此，我们主要对这些重要的时间节点来进行分析。

（1）下午 5:00-5:30 时间段

据图 8 所示，“POK 拉力赛”、“暴力”、“警察”等关键词在饼图中有突出的位置。而从提取的 message 数据（图 9 ）中我们了解到，阿比拉城市公园的 POK 集会在该时间段开始举行。可以推测在这段时间里，POK 集市上可能发生了暴力事件，警察因此关闭了集会附近的街道，暗示这段时间里公众有轻度至中度的安全风险。

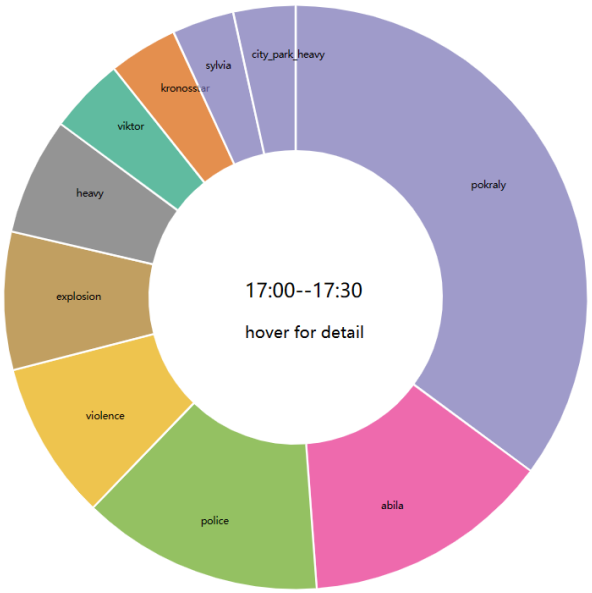

Figure 8: 5:00-5:30 主要关键词分布图

Message

17:33:00 message: RT @AbilaPost #APD has closed Parla St from Egeou St North to Alm St due to size of POK rally
17: 33: 45 message: RT @CentralBulletin police have closed streets surrounding POK rally in Abila #Abila

Figure 9: 5:00-5:30 部分 message 数据

对比 5:40 左右的饼图发现关键词没有太大变化，说明公众风险基本保持不变，这里不再过多赘述。

（2）下午 6:00-6:40 时间段

由图 3 可知出现了新的关键词“environmental”、“newman”、“education”等，推测似乎有一个集群在讨论环境活动，并且有几位演讲者，例如 Newman 博士。通过图 11 可以发现，相关参加者的活动范围非常集中且长时间没有改变，所以推测该活动的规模较小，不会对公共安全构成风险。

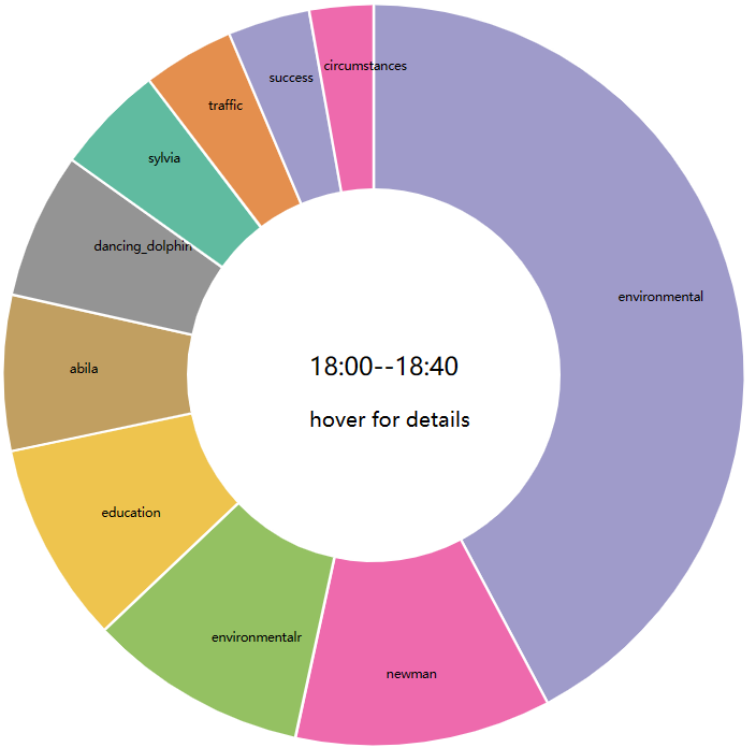

Figure 10: 6:00-6:40 主要关键词分布图

Figure 11: 6:00-6:40 人群分布地理地图

（3）下午 6:40-7:00 时间段

由图 12 得出“dancing_dolphin、“fire”、“police”等主要关键词，据图 13 可以推测在下午 6:40 左右，跳舞的海豚位置发生了火灾，火势还可能蔓延到了建筑物中。对比之前的地理地图可以发现，在图 14 中人群的分布范围开始扩大。据此可以推断，本次火灾的事件规模很大，对公众安全造成了中度至重度威胁，并且如果火灾在商业区蔓延，很有可能由于无法短时间内疏散人群而造成数百人伤亡。

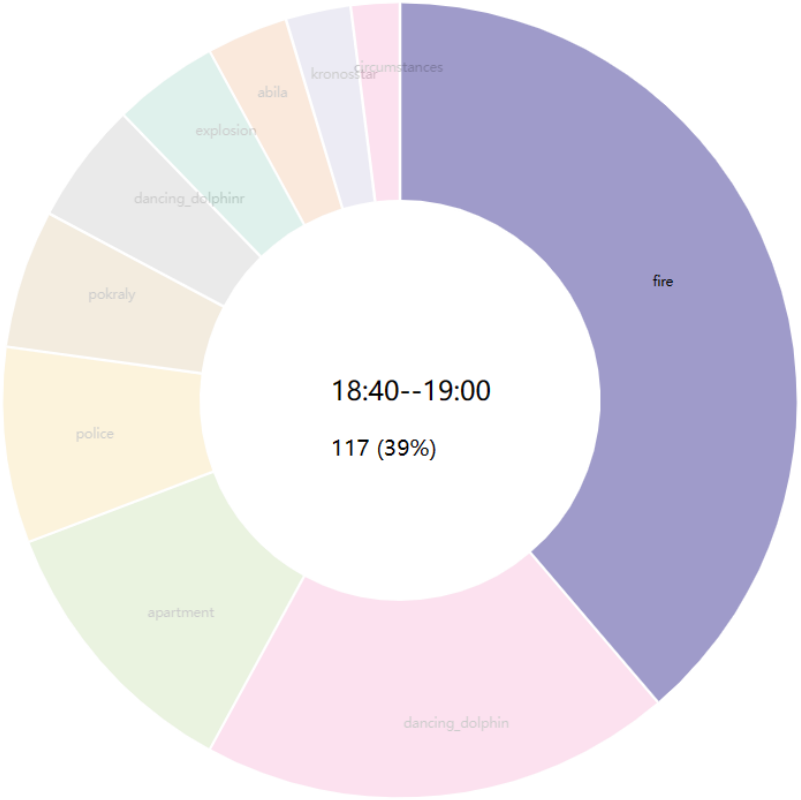

Figure 12: 6:40-7:00 主要关键词分布图

Message

18:33:24 message: Just found some local hot peppers at the Abila market. Feelin' the burn! #abilafire
18:42:10 message: Abila Fire Department has dispatched 2 units to a possible fire at the Dancing Dolphin apartment complex. #KronosStar

Figure 13: 6:40-7:00 部分 message 数据

Figure 14: 6:40-7:00 人群分布地理地图

在 7:30 左右，“fire”关键词的频率显著减小，表明火灾的严重程度已经得到控制，可以推测此时火灾仅对公共安全构成中等风险。

（4）下午 7:30-8:00 时间段

由图 15 可知新关键词“hit”被引入并迅速增长，且出现了“van”、“killing”等关键词，结合 message 数据（图 16 ），可以推测出当晚有一名警察在犯罪事件的追逐过程中被枪杀，由此可以推断出事件现场已经对公共安全构成严重而紧迫的威胁，在这个时间点，平民似乎可能会陷入这一危险事件的交火中。

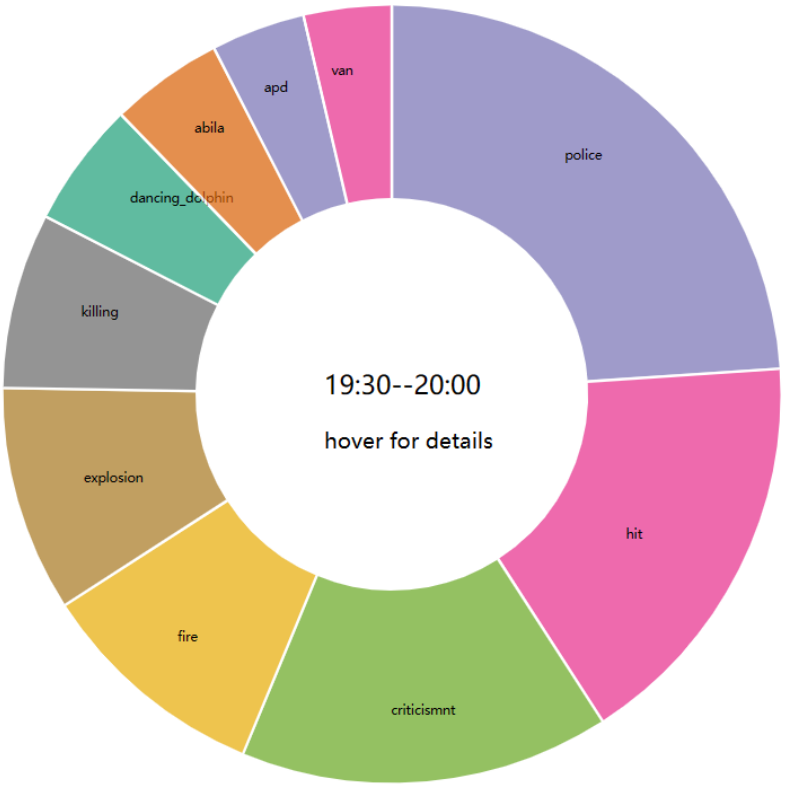

Figure 15: 7:30-8:00 主要关键词分布图

Message

19:41:13 message: Nooooo! a cop just got killed - the shooter is still in the gelatogalore parking lot
18:42:45 message: Shots fired at Gelato Galore. Police officer is down. #KronosStar

Figure 16: 7:30-8:00 部分 message 数据

（5）下午 8:00-8:30 时间段

据图 17 可知“hospital”、“criticismnt”、“police”、“standoff”关键词占较大比例，说明枪击事件已经发展到对峙阶段，事件的严重程度和危险性升级，特别是“hospital”关键词表明有人受伤，推测这种情况可能会继续对公共安全构成严重威胁。

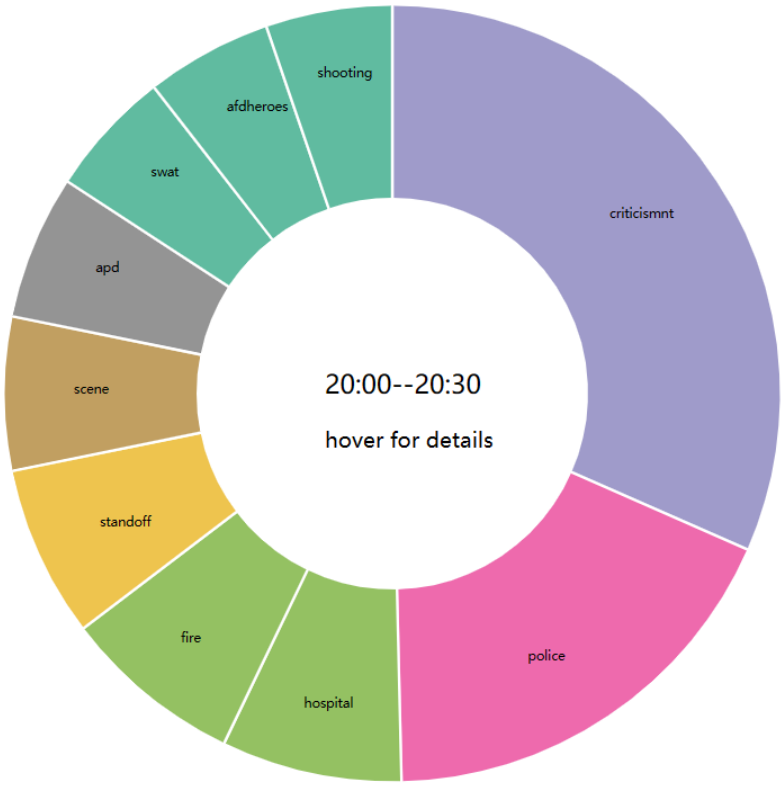

Figure 17: 8:00-8:30 主要关键词分布图

（6）下午 8:30-9:00 时间段

在这段时间里,“fire”关键词的比例上升,同时出现了“afdheroes”、“building”、“avoid”关键词，表明之前的火灾已经死灰复燃，考虑到火灾发生半个小时以内就已经得到控制，并且群众以及开始疏散，所以推测本次事件对公共安全仅构成轻度至中度风险。在 9:30 左右一些动词和形容词的占比开始增加，推测公众应该已经开始主动避开事发现场。但犯罪、警察等关键词出现的频率依旧比较频繁，推测 pok 集会和枪击事件的情况仍然紧迫，多重事件仍然对公共安全构成中度和严重威胁。

概括来讲，在晚上 5:30 左右，集会似乎煽动了公众情绪，有大量警察在场。6:30 左右发生的大火也很危险，后期甚至卷土重来。7:30 左右发生了枪击警察事件，这有可能对公众造成致命后果，从而对公共安全构成严重威胁。而在 9:30 左右，公众开始撤离，但暴力事件情况依然紧迫，对公共安全有潜在严重威胁。

5.3 任务三

Question1: 如果您能够将一组急救人员发送到任何地方，它将在哪里？

Answer1: 急救人员将被派往火灾发生地附近，其经度为 36.059，纬度为 24.894，位置为 N. Achilleos St / N. Madeg St。

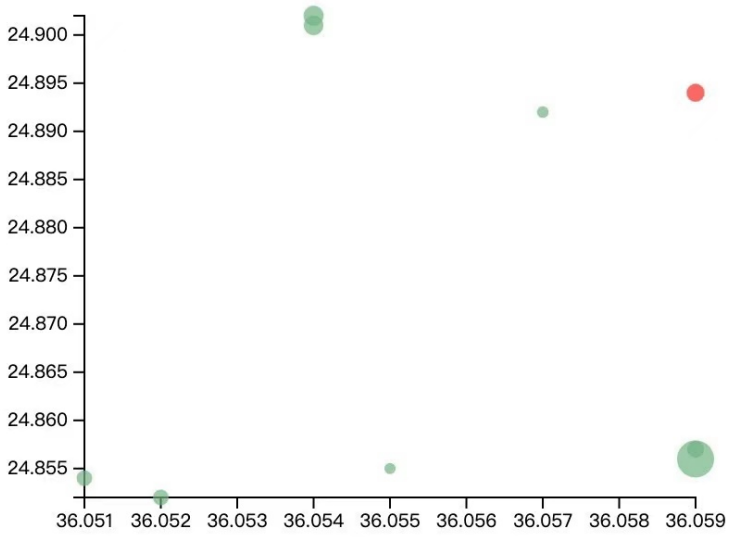

Figure 18: latitude-longitude-1

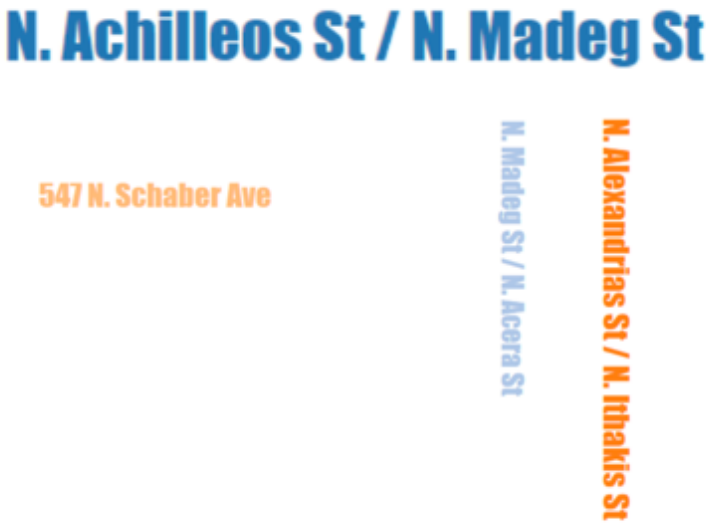

Figure 19: location

Cause1：依据得出的各个时间段的关键词信息，可以得出事件流如图所示。

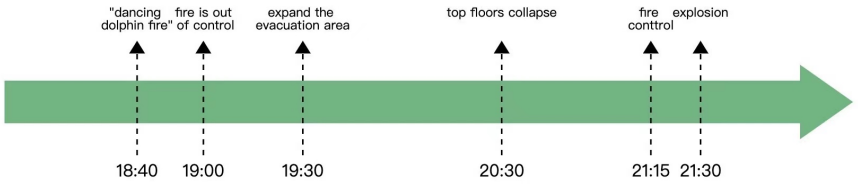

Figure 20: Dancing Dolphin fire

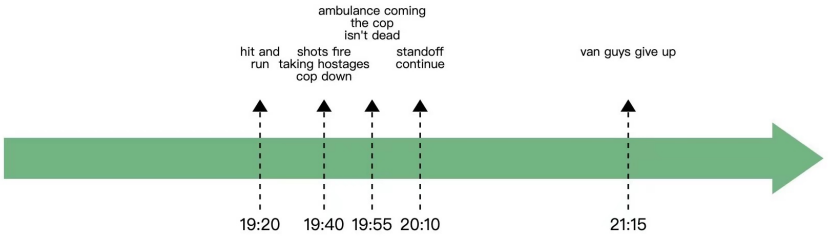

Figure 21: standoff

可以直观看出火灾在 18：40 发生；19：00 时火势失控，同时开始扩大疏散区域；后续火势持续失控，在 20：30 时顶层发生坍塌；21：30 时发生爆炸且有人员受伤。在火灾事件中，自火灾发生，火势近三小时内一直未得良好控制，属于高风险度事件；同时，发生了坍塌和爆炸此类重大风险事件，爆炸影响范围超出人群疏散范围，严重危害到群众的生命财产安全。

反观发生在 19：40 的警匪对峙事件为高风险度事件，起因为 19：20 发生的肇事逃逸，被撞人员并未受伤。19：40 时发生枪战，货车司机绑持人质，一名警员在枪战中受伤倒地，评估为高风险度事件；19：55 救护车救治受伤警员，大约 15min 后发布报道称警员处于危险但稳定的状态下；后续，对峙持续进行，警察发布公告远离此区域并保持静止；最终，在 21：15，货车司机放弃抵抗，人质获救。对峙事件和平结束，无消息表明除警员外有其他人受伤，受伤警员已被救治且无生命危险。

综上，警匪对峙事件与火灾事件相较，影响范围小，受伤人群少，风险系数较低。在枪战事件中，货车司机向警察开枪，除人质外并未袭击普通群众，同时，普通群众有能力寻找掩蔽体以保护自身安全，最后，可以确定除警员外无人员受伤；而在火灾，尤其是后续的爆炸事件中，群众无法估计或错误估计其影响范围，在爆炸发生时无法实施自救，同时已明确有人员受伤，在建筑内可能也存在未被解救人员，更为危险，更需要急救人员在场。所以，回顾事件，急救人员应被派往火灾发生地附近，应及时救治在火灾、坍塌、爆炸事件中受伤人群。

Question2：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？

Answer2：若实时性响应事件，急救人员将被派往警匪枪战对峙事件发生地附近，其经度为 36.059，纬度为 24.856。

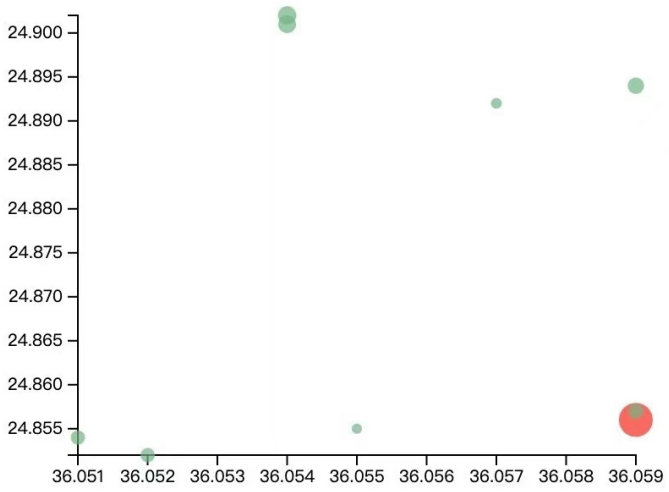

Figure 22: latitude-longitude-2

Cause2: 考虑到火灾事件 18：40 时发生，19：30 时警察扩大人群疏散范围，提取了自 19：30 至 19：32 时的信息关键字数据；考虑到对峙事件 19：40 左右发生，提取了自 19：41 至 19：43 时的信息关键字数据。选取两份数据中出现次数最多的且有意义的关键字，可视化其比例分布饼图，如图所示。

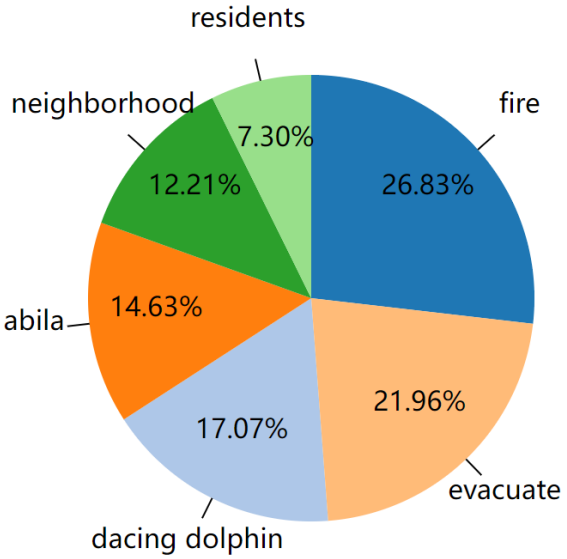

Figure 23: 19:30-19:32 keywords

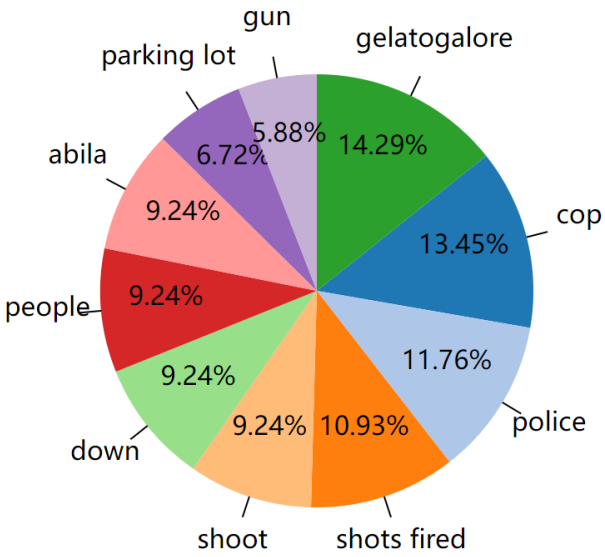

Figure 24: 19:41-19:43 keywords

对比两张比例饼图，可以直观看到：

- 19：30 左右”fire” 占比最大，其次是”evacuate”。结合所有关键词我们可以得出结论：19：30 左右以发生在 dancing dolphin 的火灾事件为主，附近居民有序撤离，那么，也就意味着群众的安全在撤离到安全区域后得以保证。

- 19:40 左右地点关键字”gelatogalore” 占比最大,其次为”cop” 和”police” 此两个在此同义的词，次之为”shots fire” 和”shoot” 此两个在此近义的词。结合所有关键词可以得出结论：19：40 左右以发生在 gelatogalore 的停车场的枪战事件为主，有警员受伤倒地，可能危害到群众安全。

综上，警察已对火灾事件在 19：30 左右给出相应应对措施，即扩大人群疏散区域，以此来保证人民群众安全。截止此时，距火灾事件已过近一个小时，相关部门给出了良好处理方案并实施，且有报道显示无严重受伤情况发生。而在之后的十分钟后，19：40 发生的严重枪战事件，一名警员受伤倒地，其余警员继续与匪徒对峙，可能危及周围群众。此时，枪战事件刚刚发生，该事件还未得到良好的处理方案，目前急需的是对受伤警员进行救治，同时确认周围群众的健康状况。所以，对事件进行实时性响应，应将急救人员派往对峙事件发生地。

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。

杨朝涵：

在本次实验中，我主要负责任务二内容的实现。关于这部分内容，我认为处理得到关键词随时间变化的数据是任务二实现的基础，分析预处理后的重要关键词与 message 的关系是分析公众风险水平的关键。此外，通过分析 message 数据，我学会了如何充分利用已有文本类数据提取关键信息，并结合可视化推测在不同时间段发生的事件。通过制作带有动画效果的可交互的饼图和地理地图，我对 D3 动画与交互有了更加生动的认识，也能更加熟练地使用 QGIS 软件实现地理位置的可视化。虽然在实验过程中，由

于我对 D3 知识掌握的不够熟练，出现了有些元素渲染不到等问题，关于用 Python 处理数据也耗费了较长时间，但通过查阅相关资料和求助同学，我最终完成了本次实验，并且收获良多。

张雪雅：

在本次大作业中，我主要负责任务三的部分。在此部分，我认为最重要的是在对数据分析后进行的思考，也耗时最久。因为本任务为开放性问题，要首先通过数据分析确定出发角度，沿此思路完成后续对数据信息的可视化。在数据分析阶段，沿用了任务一中的思路，利用 python 进行数据提取，因有良好基础，这部分并不复杂；比较困难的是对有用数据进行筛选，但通过资料的查阅也得到良好解决。在数据可视化阶段，这是本次大作业的重中之重，在数据的基础上考虑可视化样式，因刚接触 d3 不久，对此方式可视化数据不是非常熟悉，在这阶段耗时较久，通过回顾理清 html 和 js 语言，最终完成了本次绘制。在报告撰写阶段，因在之前小学期阶段，已使用过 latex 撰写报告，这部分并不困难，在完成报告的过程中，再次梳理任务完成思路，回顾历程。总之，本次大作业从方方面面都带给我很大的收获！

参考文献

[1] Bingkun Chen, Hong Zhou, and Xiaojun Chen. E-embed: A time series visualization framework based on earth mover’s distance. *Journal of Visual Languages & Computing*, 48:110–122, 2018.

[2] Mohammed Ali, Mark W Jones, Xianghua Xie, and Mark Williams. Timecluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer*, 35(6):1013–1026, 2019.

[3] Ashley Suh, Mustafa Hajij, Bei Wang, Carlos Scheidegger, and Paul Rosen. Persistent homology guided force-directed graph layouts. *IEEE transactions on visualization and computer graphics*, 26(1):697–707, 2019.

[4] 韩刘. 多元异构网络复杂多维数据可视化方法. 计算机仿真, 37(11):299G303, 2020.

[5] 王悦. 可视化指导下的图像领域深度学习模型优化方案设计. PhD thesis, 合肥: 中国科学技术大学, 2020.

[6] 田丰 and 喻纯. 自然人机交互新进展专题前言. 软件学报, 30(10):2925–2926, 2019.

[7] 田宇荃, 赵迪, and 陈虎. 海绵城市建设前后的可视化效果分析———以常德市为例. 智能建筑与智慧城市, 6, 2020.

[8] 陶芳. 网络舆论数据可视化技术研究. PhD thesis, 北京: 北京邮电大学, 2019.

[9] 张甜甜, 李妮, 龚光红, and 卢元杰. 一种基于数独分组的拉丁超立方试验设计方法. 系统仿真学报, 32(11):2185, 2020.

[10] Lidong Zhang, B Vinodhini, and T Maragatham. Interactive iot data visualization for decision making in business intelligence. *Arabian Journal for Science and Engineering*, pages 1–11, 2021.