可视化课程设计报告

——VAST Challenge 2014 MC1

Group 8: Bug 调不队

熊伟民 (1900011621)　张启哲 (1900011638)　董欣然 (1900013018)　谢悦 (1900013055)

2022 年 1 月 23 日

目录

1 数据详细描述

本次可视化数据来自 VAST Challenge 2014 MC1，主要故事背景为虚构，总部位于 Tethys 的 GASTech 公司一直在岛国 Kronos 经营一个天然气生产基地，并创造了巨大的利润。但是，却有人认为 GAStech 对 Kronos 的环境造成了破坏。2014 年 1 月 20 日，GAStech 公司的一次庆祝活动中，一些员工失踪，而同时一个被称为 Kronos 保护者（POK）的组织被怀疑与失踪有关。

关于案件所提供的数据主要包括 GAStech 公司员工信息记录及简历，公司内部员工邮件往来记录，国内外从过去到现在多篇相关新闻报道，以及当地的背景信息。我们主要处理并可视化了新闻、邮件及公司个人信息简历三方面数据。

1.1 新闻数据

共有 845 条国内外媒体从过去到现在多篇事件相关新闻报道。其中 580 条为事发日期 2014 年 1 月 20 日前，新闻时间分布从 1982 年 10 月 2 日到 2014 年 1 月 19 日。剩余 265 条为事发时间 2014 年 1 月 20-21 日当天新闻。对于事发当天的所有新闻，我们可以通过按时间排序梳理出失踪事件时间线。对于历史新闻，我们可以挖掘出其潜在信息，建立失踪事件的背景。

所有新闻来自 29 家媒体，通过文本比对和时间排序，我们发现并判断，所有新闻内容的关键信息源主要来自五家媒体，分别是 Homeland Illumination、Kronos Star、Abila Post、The World、International Times，而前三家媒体在事件发生当天发表了带具体时间标签的事件报道。其余媒体基本上都在这几家媒体提供的关键信息上转载或重新编辑发表新闻。所以我们的可视化重点在可视化这些关键信息，但同时其他媒体的转载重构也不容忽视，它们体现了所报道新闻的热度。

1.2 邮件数据

邮件数据为事发前两周从 2014 年 1 月 6 日到 2014 年 1 月 17 日 GAStech 公司内部员工所有邮件的收发人、收发时间及邮件标题。该时段主要邮件共 1170 封，来自公司 6 个部门 40 名员工。因为有大量的群发邮件，这些邮件掩盖了员工的私人邮件，所以我们把群发邮件清除了。我们可以结合公司员工信息，通过展示邮件收发时间和邮件内容信息，找到不寻常的邮件帮助案情分析和公司员工内部关系状态建立。

1.3 公司个人简历数据

有关于公司个人信息的数据为公司简历数据和公司员工基本信息表格。

简历数据共有 5 份公司 Executive 层人员的个人介绍和 30 份公司员工的简历，总计 35 份。从个人履历中我们可以找出与 POK 组织相关的员工信息，也可以挖掘有相似经历背景的人员，找出人员之间隐藏的联系。

公司员工基本信息表格详细记录了员工个人信息，主要包括出生年月、出生地、性别、国籍、护照相关、工作相关和参军相关。从该文档中，我们可以直观地看出各员工所属的部门还有邮箱账号，了解每一个员工的基本信息，找出可疑人员。

2 项目设计及最终系统解释

下图为项目设计可视化系统总图。

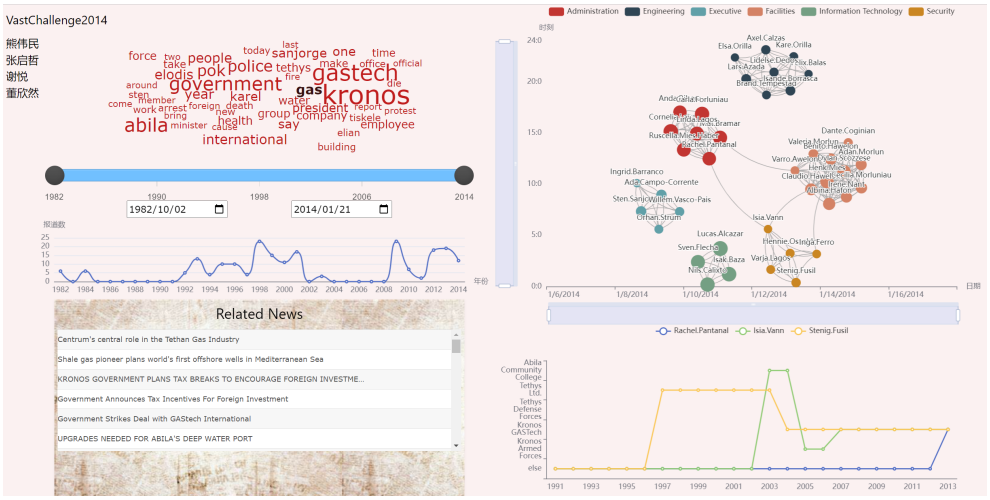

图 1: 可视化系统

系统主要分为左右两部分，共 6 个组件，左半部分为新闻数据的可视化，有 4 个组件，分别是时间轴组件，关键词词云、新闻列表和关键词折线图，点击新闻列表，可以弹出有新闻详细信息的提示框。右半部分为公司人员数据的可视化，有 2 个组件，分别为公司内部人员邮件收发关系力导向图和人物履历轨迹图，分别点击力导向图的点和边，可以弹出有人物履历和邮件收发详细信息的提示框。下面分步介绍这些组件的设计想法和使用方法。

2.1 新闻数据可视化

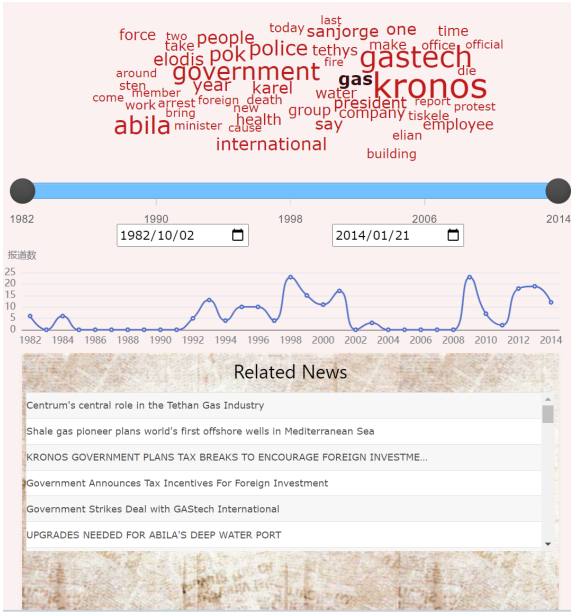

图 2: 可视化系统左侧

如图2，可视化系统左侧为针对新闻数据处理可视化，从上往下分别为词云、时间轴、折线图和新闻列表。时间轴定位新闻时间范围，词云展现新闻的关键信息，折线图展现关键词按时间

分布热度，新闻列表列举详细新闻信息。在交互中，时间轴控制全部数据的范围，对其他三者均有影响，通过词云选关键词定义折线图和新闻列表中新闻内容。

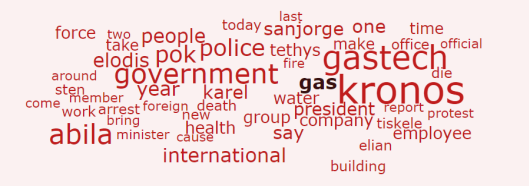

图 3: 关键词词云

图 5: 新闻列表

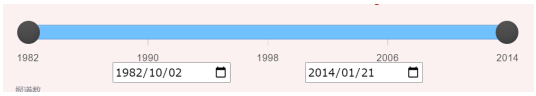

图 4: 时间轴组件

图 6: 关键词折线图

2.1.1 时间轴组件

如图4，为时间轴组件。为了按时间筛选新闻信息，我们设计了时间轴组件。组件分粗准对位功能和精准对位功能，粗准是一个可拖拽时间轴，可以粗略选定想要的时间范围；精准是时间轴下方的两个始末日期设定，可以具体到日。根据选定的时间范围，词云和新闻列表相应变化。若选择起始时间 2014/1/20，终止日期 2014/1/22，新闻列表会按时间顺序展示失踪事件发生那两天的所有时事新闻，列出时间线。

2.1.2 关键词词云

如图3，为新闻关键词词云，居可视化系统左上方。为了展现所选时间范围内新闻所含重点信息量，滤去新闻文本中大量冗余文字信息，我们设计了一个词云组件。

组件以所选时间段内所有新闻为数据，关键词的大小和位置与该词在新闻中出现次数相关。关键词越大，说明这个关键词在该时间段内出现频率越高。用户可以通过选择多个词汇，和其他图表进行交互。点击词云中的关键词，新闻列表和折线图里的内容将会被所选中关键词限制。新闻列表数据为时间轴组件所给范围内，包括所选关键词的新闻。折线图中显示出所选关键词在各个年份的新闻报道中分布数量曲线。

2.1.3 新闻列表

如图5，为新闻列表，居可视化系统左下方。为了展示所有新闻，方便用户查阅，我们设计了新闻列表。列表将新闻按时间从古至今排序，显示新闻标题。点击新闻条目，可以如图7显示该新闻详细信息，包括新闻标题、所属媒体、具体发布时间和详细内容。新闻列表受时间轴组件和词云组件影响。

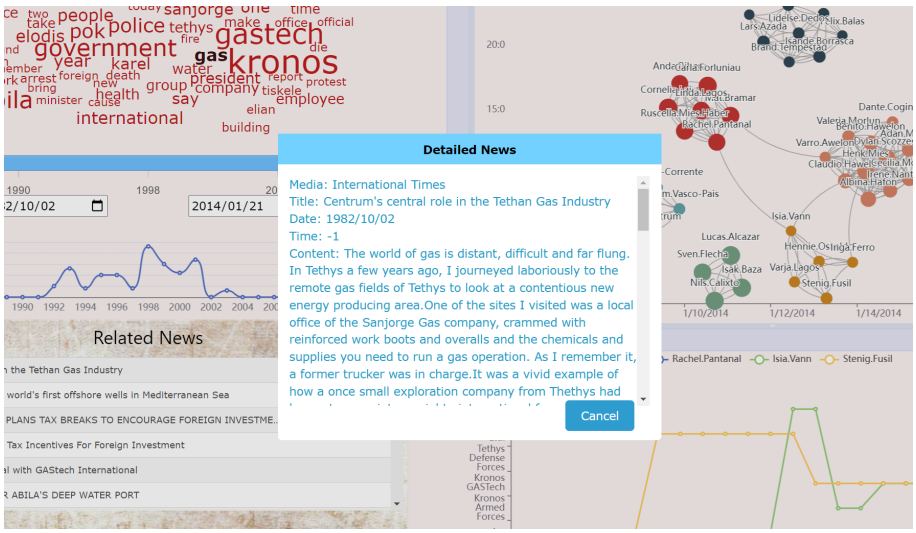

图 7: 新闻弹窗

2.1.4 关键词折线图

如图6，为关键词折线图，居可视化系统左侧中部。为了展示按时间分布关键词的热度，我们设计了折线图，将新闻逐年统计，横轴是新闻的所属年份，纵轴是该年份下新闻的数量。按横轴时间显示各个时间段中特定关键词出现的频率，来检测是否有新闻中关键词在某一时间段频繁出现。折线图受时间轴组件和词云组件影响。

2.2 公司人员数据可视化

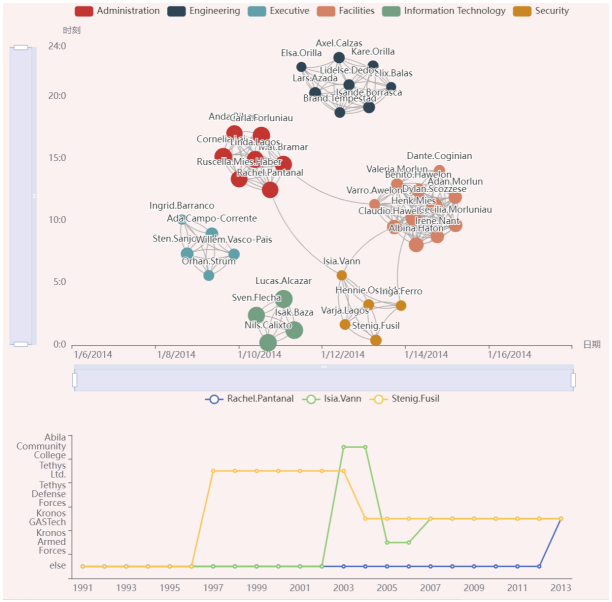

图 8: 可视化系统右侧

如图8，可视化系统右侧为针对公司人员数据处理可视化，从上往下分别是邮件收发关系力导向图和人物履历轨迹折线图。邮件力导向图通过邮件收发展示了公司人员的关系，人物履历轨迹图展现了公司人员入公司前的大致履历轨迹。在交互中，力导向图的时间轴和 legend 控制力导向图的数据内容，点击力导向图中的人物，控制履历轨迹图中的折线数量。

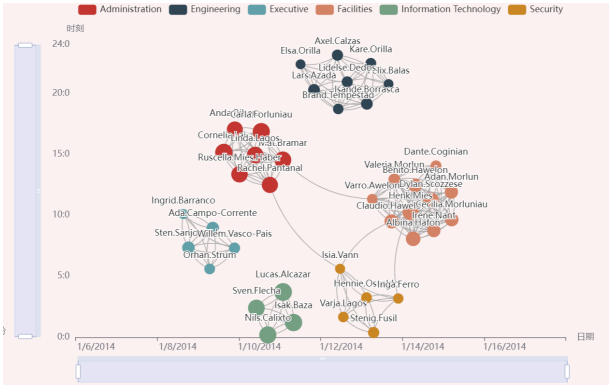

图 9: 邮件收发关系力导向图

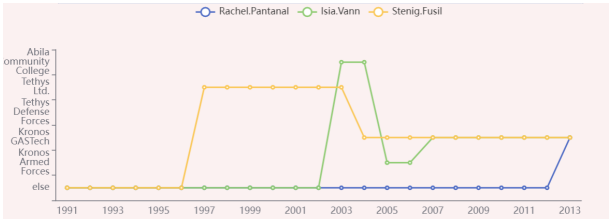

图 10: 人物履历轨迹图

2.2.1 邮件收发关系力导向图及人物信息弹窗

如图9，为邮件收发关系力导向图，居可视化系统右上。邮件表现了公司群体中两两之间交往的状态，某种程度上可以展现公司内部人员交往情况，发现潜在可疑线索。力导向图可以完成很好的聚类，方便用户看出公司人员间的亲疏关系。于是我们根据公司内部邮件往来设计了力导向图，依据邮件的数量设置引力和斥力。

力导向图下方是日期时间轴，左侧是一天内具体时间时间轴。由于邮件是从 2014/01/06 到 2014/01/17，通过选择起始日期限定邮件的日期范围，可以查看非工作日和工作日的邮件收发区别；通过选择一天 24 小时内不同的时间段，可以查看不正常时间段邮件收发情况。力导向图上方是不同部门的 legend，公司员工分为 Administration、Engineering、Executive、Facilities、Information Technology、Security 六个不同的部门，单击 legend 可以选定部门，对邮件进行筛选。

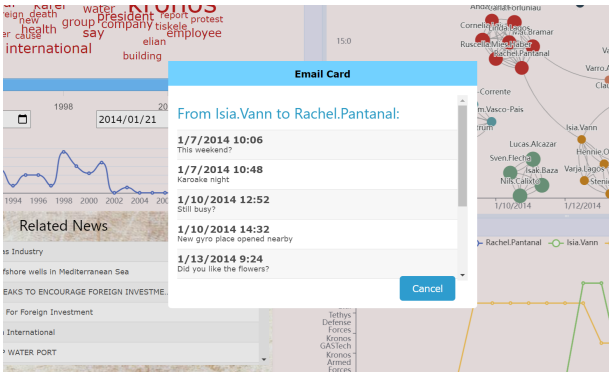

图 11: 邮件详细信息弹窗

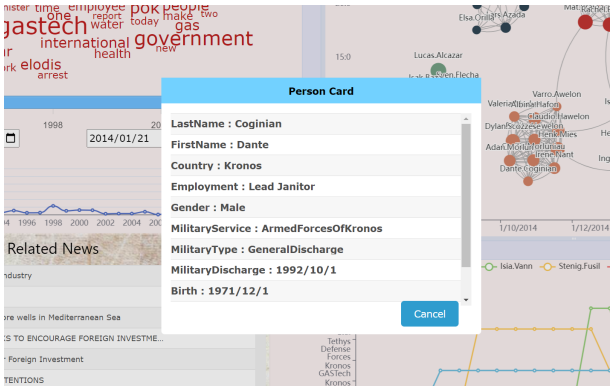

图 12: 邮件详细信息弹窗

如图11，为了展示邮件收发的详细信息，点击两个人之间的连线，弹窗弹出双方邮件往来详情，即邮件标题。此外，如图12，为了展示人物基本信息，点击每个人的圆点，弹窗弹出人的基本详情，包含出生、性别、所属国别、入职年月等等。

2.2.2 人物履历轨迹图

如图10，为人物履历轨迹图，居可视化系统右下。公司内部人员简历上有其之前待过的地方，有不少人在相同时段待的地方是一样的，为了探索人物潜在关系信息，我们将他们的时空轨迹做成横轴时间，纵轴地点的轨迹图，时空轨迹图的折线展现了该员工的工作轨迹随年份的变

化。同时，时空轨迹图和邮件关系图关联，点击邮件图上的人、关系线，将在时空图中呈现该人的轨迹。这样可以看出像某两个人之间可能在进入公司之前就可能认识这样的信息。

3 发现的结果

本可视化系统旨在方便警察分析浩如烟海的文本数据。我们使用我们的可视化工具，对相关案情进行了分析。在已知案情发生时间为 2014 年 1 月 20 日到 2014 年 1 月 21 日的境况下，我们在新闻检索模块中选定该时间区间，同时，本次案件发生在 kronos，是关于 Gastech 员工的失踪事件，故选定关键字 kronos，Gastech，employee，miss。在根据选定的关键词筛选出的新闻列表中，我们可以发现一些关键信息（图13）：14 名员工被绑架；POK 组织有重要嫌疑；Edward Vann 有重大嫌疑。

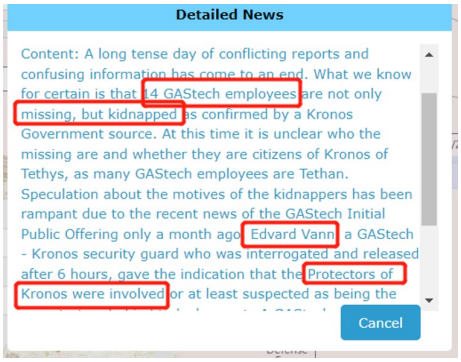

图 13

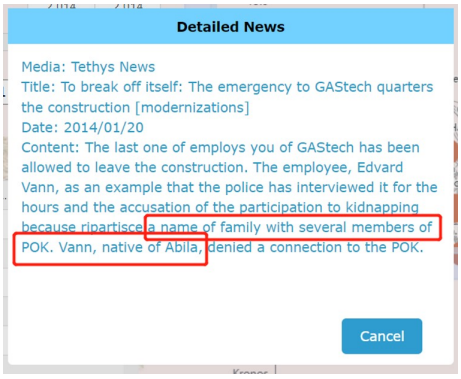

图 14

通过选定关键词“绑架”，查看新闻，我们可以初步认定，POK 策划了此次失踪案件。在词云图中选定“POK”，” Vann “，看到一条有价值的线索：Edvard Vann 有着和 POK 成员相同的姓氏。有就是说 POK 大多成员的姓氏是 Vann，是 Abila 本地人（图14）。

观察邮件 csv 数据后，我们可以发现大量的群发邮件无法提取有效信息，因此对邮件进行了筛选。筛选后可以发现除了群发邮件外，不同部门之间只有少量的邮件往来。进一步筛选邮件时间后，我们可以发现一些可疑的邮件往来，经过点击后可以查看邮件详情（图15）。

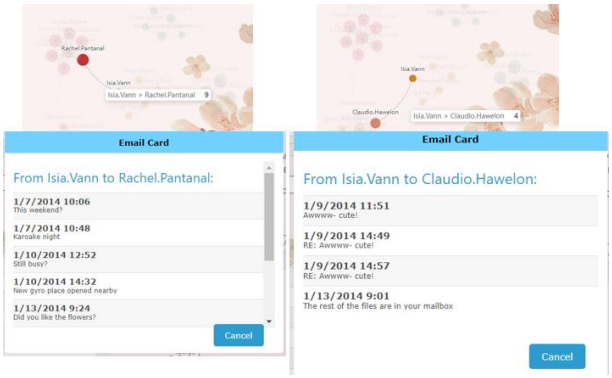

图 15

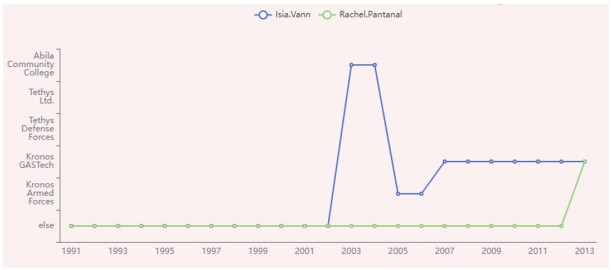

图 16

结合新闻中的线索我们可以得知，Isia Vann 的姓氏和 POK 成员相似，从时空轨迹图中我们可以得知，Isia Vann 在 Abila 上大学，在 Kronos 当地服役，最后进入 GesTech 公司，因此我们得出初步的判断：Isia Vann 可能是 POK 成员（图16）。由新闻报道可知：政府断定 POK

谋划了此次失踪事件，所以 Isia Vann 是员工失踪案的犯罪嫌疑人。

4 关于本工作的讨论

本小组成员利用学期中每周日晚对本课程设计工作进行了详细的讨论，从最初的数据分析到采取什么样的可视化手段对数据可视化。因为我们的可视化系统最重要的职能是帮助警方尽快掌握所有数据线索，看清案情正向，所以我们结合数据本身，思考最能展示出数据对案情分析帮助的可视化手段。

4.1 思考

新闻数据记录的是故事，是这个地区有关于 POK 组织和 GAStech 公司从古至今演变发生的事情，是一个多角度多时空信息量浩如烟海的数据体。我们要从这样一个数据体中找到潜在的信息是一件很复杂的事情，利用文本分析手段，先文本比对，断定文本信息关系，找到并剔除信息冗余，再对较为精炼的文本进行可视化。新闻数据的时间属性非常重要，有时间上的筛选和展示很重要。词云手段可以突出关键信息，而关键信息和时间二者之间的联系可以使用简单的二到三维可视化手段。课程项目也引发了我们对大数据文本处理手段的更多探索渴望。

邮件数据背后隐藏着人物之间的关系，收发邮件的密切程度从某个侧面展现公司人员的关系，这就可以使用图可视化手段。邮件遍布案发前的两周，可以对标题文字查找是否有提前与此次案发事件有关的邮件。时间也是邮件信息的一个关键属性，可以筛选非正常时间的信息来找到隐藏线索和关系。公司的不同部门之间的人有联系是什么原因？是否和此次事件相关？我们需要具体到公司员工个人的详细信息上，去了解探索出来的事件核心或者可疑人物的详细背景信息，去了解其作案动机、事件发生的原因，嗅出事件发生的各个细节，并帮助警察通过合理推理还原案情最真实的样子。

4.2 展望

我们的可视化系统还可以进一步优化：

1. 左侧新闻时间轴的比例可以根据新闻时间分布特殊性将事件发生前的新闻时间段长度和事件发生时的新闻时间段长度调整统一，我们需要探索更好的实现方式；

2. 我们希望可以进一步归纳出人物潜在关系，可以考虑添加更多的可交互部分，如新闻与邮件详情中文本的可点击交互，简化信息的提取过程。我们希望交互系统可以更多层更深入探寻数据内在信息，

3. 我们希望更加充分地利用历史新闻，提取构建出 GAStech 公司人员结构的变化关系，以期从中找出 GAStech 与 POK 组织成员间的联系，甚至找到 POK 组织成员并构建出 POK 组织成员关系图。

4. 对于庞大的文本数据体，我们后续可以借助 NLP 相关知识对其进行分析。

5. 邮件的标题较短，但足以透露出两人之间交谈的总体内容，我们希望之后可以通过对标题定性，用更合理的手段结合标题更加深入的展示公司人员的关系。

5 小组成员分工

本可视化系统设计的创意由本小组所有成员共同讨论完成，前期设计报告由大家轮流完成，后期代码修整由大家共同添砖加瓦。

熊伟民：新闻数据处理及可视化，报告撰写。

张启哲：新闻数据处理及可视化，代码整合，海报设计。

董欣然：邮件及简历数据处理及可视化，代码整合，demo 录制。

谢悦：新闻数据处理，关键词折线图绘制，报告撰写。

6 收获与致谢

通过本项目设计和一学期的课程学习，本小组成员收获了丰富多彩的可视化知识，学会在特定情况下使用何种可视化手段帮助数据展示，也熟练掌握了 javascript 编程并获得了绝佳的团队合作经验，我们获益匪浅。故在此：

特别感谢袁晓如老师在本学期精彩的授课与悉心的指导！

感谢课程助教李思航与谭绍聪辛勤的付出与热心的帮助！

感谢其他小组同学带来的优秀的项目展示与中肯的建议！

感谢本小组所有同学在项目中充分的投入与默契的合作！