《鼠疫》人物共现分析

数据采集

数据源来自 `https://luoxiadushu.com/shuyi/`，采集过程见 `crawler.py`，所有的章节都存入 `book` 目录下。
章节的 href 只在 `<div class=book-list clearfix><ul><li>` 下，因此从中提出所有章节链接，去掉译序和后记，只保留正文。每个章节分多个页面，标题和正文分别在 `<nr_title>` 和 `<nr1>` 为 id 的标签下。把原标题带上前缀编号方便文件有序呈现，清洗了正文中网站打广告的部分，分别存下，一共 39 个文件。

两种共现矩阵

共现分析方式

共现矩阵就是在一段文本中，记录任意两个角色同时出现次数的矩阵。这里的“一段文本”需要确定一个合适的粒度，如果是按照章的粒度，则太粗糙；按照段的粒度，又太细，很多时候人物对话是每句话自成一段。按照 NLP 常用的带着 overlap 切 chunk 的方式可能有损语义的完整，反而是网页原本的分片不长不短恰到好处，因此以 chunk 为单位计算，每一章的共现由这些 chunk 累加。
统计人物出现频次时，如果有 `"·"` 隔开的名字，需要把姓名分别提出统计并且累加，才能得到更准确的统计结果。例如，统计“贝尔纳·里厄”的出现次数，则需要把“贝尔纳·里厄”、“贝尔纳”、“里厄”的出现次数累加起来。

co_matrix

如果 chunk 中某两个人共现了，则对应矩阵中的值加一，定性体现了本章的共现与否。

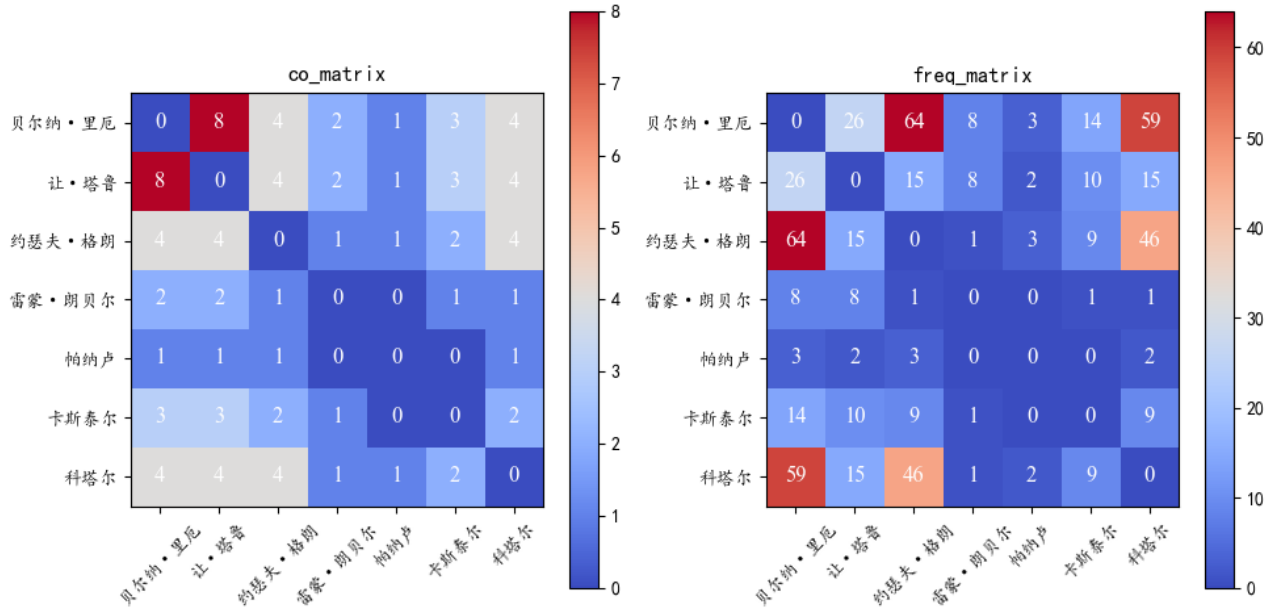
freq_matrix

如果 chunk 中某两个人共现了，则对应矩阵中的值加两个频率之中更小的值，定量体现了本章两人交流强度。

后者的值相对于前者越高，说明此人出现得更集中，大概率有大段对话；反之则是此人出现得更均匀，可能几乎每一节都露面一下，但是和别人没有很深的联系，大概率是话少神秘的个性，或者还未真正出场只是由别人提及。
共现矩阵信息存在 `output\co_occurence.json` 中，绘制的热力图存在 `figure` 中。

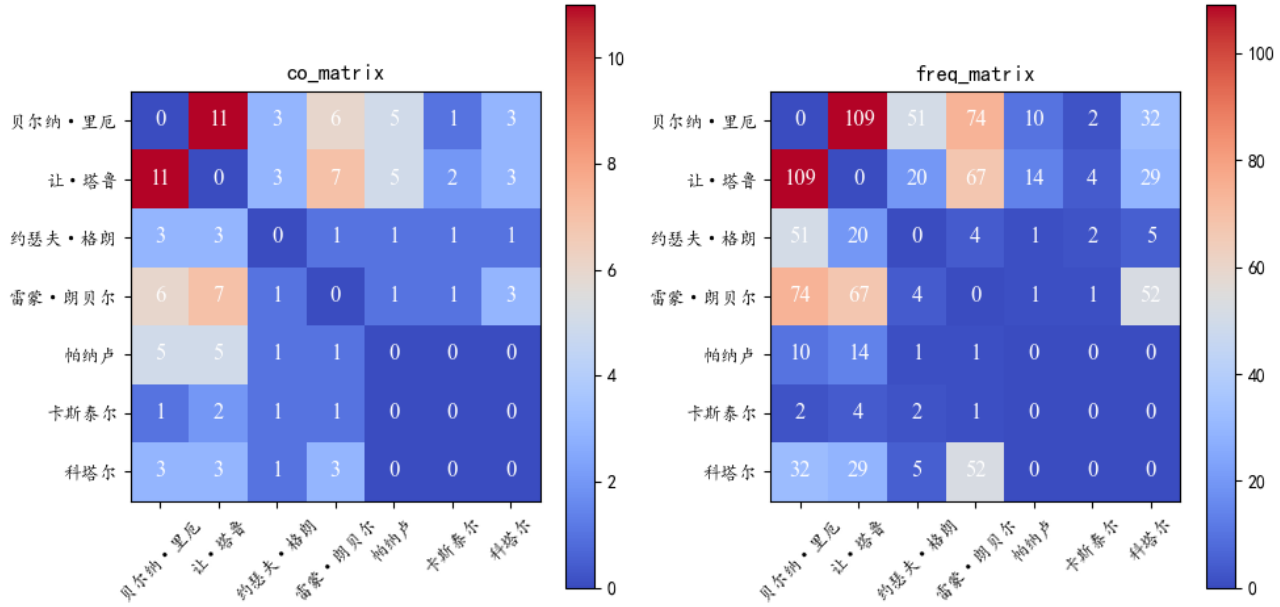
热力图分析

第一部



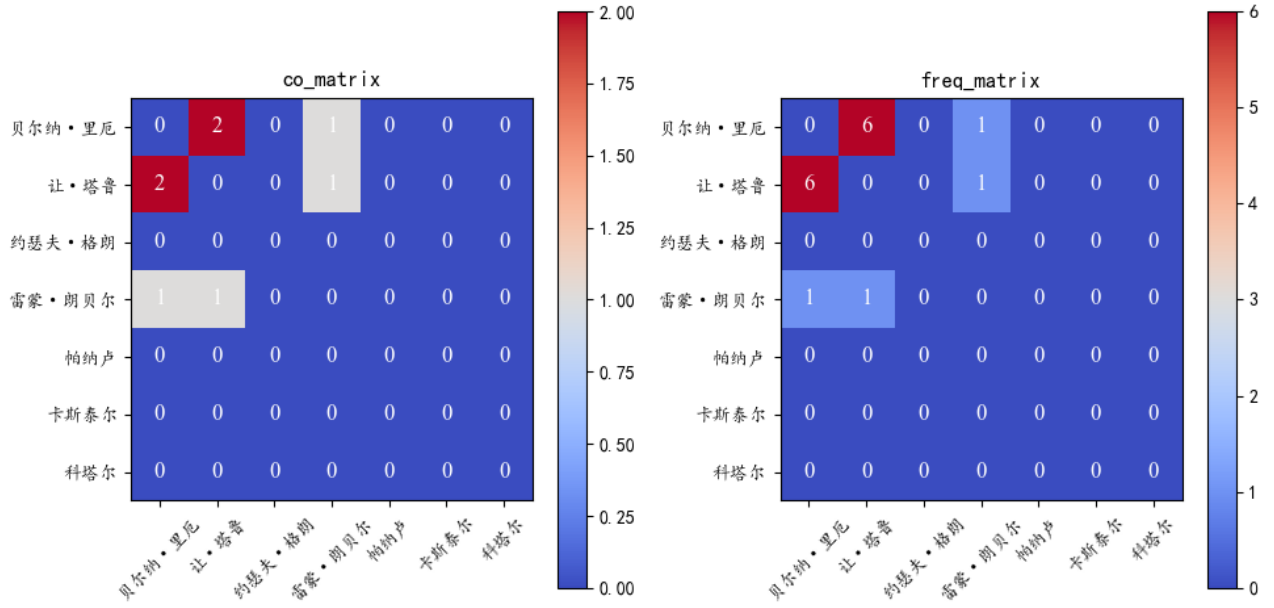
第一部主要人物均已出场，塔鲁出现较为均匀，且和里厄的关联最大；格朗和科塔尔出现较为集中，且里厄、格朗、科塔尔三人组之间大概率是进行了密集的谈话。
这和情节是相符的：第一部鼠疫正在初期，塔鲁还是一个游走旁观的姿态，里厄作为大夫一直是中心人物，格朗为了邻居科塔尔自杀未遂、疑似染病的事情来找里厄，里厄去找科塔尔了解情况。卡斯泰尔老大夫和里厄简短通过电话，帕纳卢神甫只是被提及到，朗贝尔记者也只是在走访过程中简短和三人组交谈过。

第二部



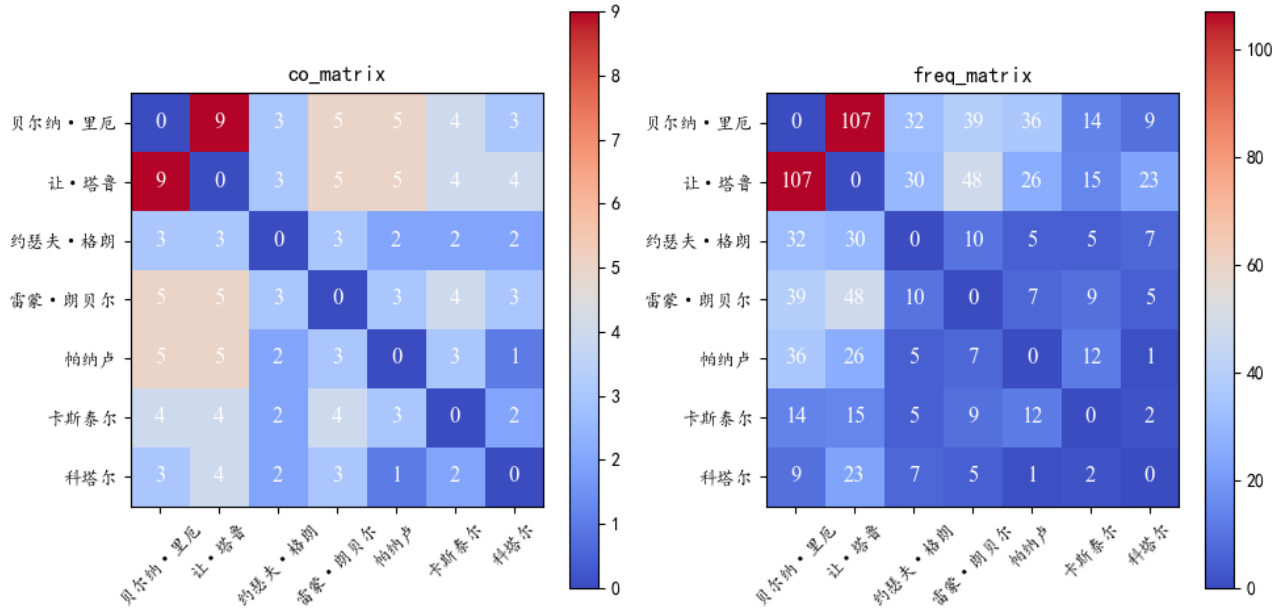
塔鲁和里厄的交谈激增，里厄、塔鲁、朗贝尔形成了新的三角，朗贝尔和科塔尔之间也有一些个人联系，格朗与里厄、塔鲁的关联较多。
第二部鼠疫进一步爆发，情节进一步展现了塔鲁的责任意识，他向里厄提议创立防疫志愿组。朗贝尔作为外地的记者被逐渐严格的封城政策困在城内，且与塔鲁住在同一家旅馆；他开始焦急，因为爱人在城外，并且向里厄寻求出城的方法。科塔尔是一个灰色地带的人物，一个神秘的酒类代理商，也做一些走私的业务，因此当朗贝尔最终发现官方无法批准，转而求助科塔尔是否有办法让他出城。格朗作为一个政府小职员，也担起责任加入里厄和塔鲁的防疫志愿组。

第三部



第三部很特殊，人物几乎没有出场，只有里厄、塔鲁、朗贝尔之间很淡的联系。因为这一部篇幅很短，且作为一个间章，没有继续人物之间的叙事，而是描述了鼠疫之下这个城市的整体状况，从人物故事记叙为主转为整体记叙、议论、抒情为主。

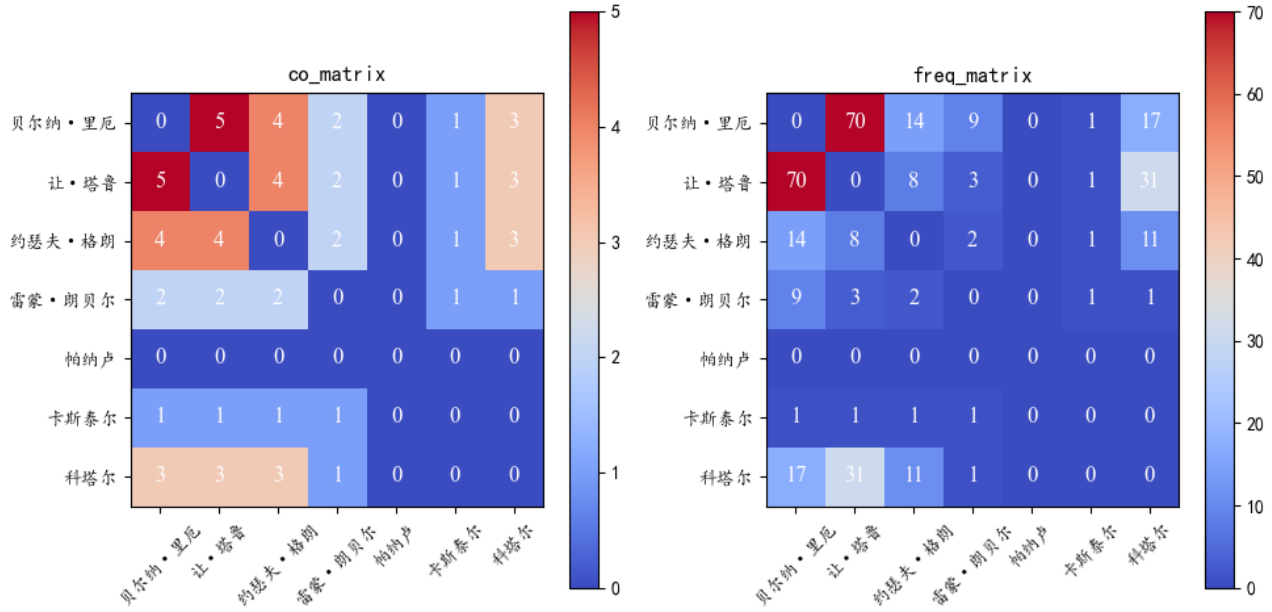
第四部



第四部中里厄和塔鲁的关系更加紧密，帕纳卢神甫的出现也从这一部开始多了起来，整体看来人物之间的联系较为广泛。

由于到这里就是我还未读到的部分，无法提供具体情节对应的分析。

第五部



第五部里厄和塔鲁的关系依然紧密，不过人物的出场就不像第四章那么全，又转为了围绕两个主角的叙事。

整体的脉络中，里厄是一以贯之的主角，而塔鲁是逐渐出现在读者视野中的。他的出场较为神秘，甚至一开始我未能意识到这会是之后的重要人物，但是自从他主动承担社会责任开始，就一直和里厄保持着高频的联系。