Exploring Major Events and Event Connections between 2019 – 2020 Using Topic Modeling via Non-negative Matrix Factorization

Yian Zhou

1 Introduction

Topic modeling has emerged as a powerful tool to discover the hidden themes that pervade a corpus. By treating documents chronologically, we can use topic modeling to analyze the evolution of focus over time. In this project, I am particularly interested in applying topic modeling via Non-negative Matrix Factorization on a corpus of 1.5*k* Chinese language news articles originally published on multiple news outlets between 07-20-2019 to 04-20-2020 and retrieved from Nei.st. I identify the topics and examine topic prevalence and inter-topic relationship over time, which reveals major political, economic, and social events and the connections between events.

2 Data Collection

2.1 Data Extraction

On April 21st, 2020, I extracted all news articles on Nei.st[1] by web scraping using Beautiful Soup[2]. Robots.txt file of Nei.st has been checked beforehand to ensure that the owners of the website allow me to do scraping on this website. Then these articles are stored in a single text file and formatted so that one article appears on each line.

The final corpus consists of 1519 unique news in Chinese language, much fewer that I expected[3]. I take some time to observe the corpus and find there are 9 news articles published in 2018 and the rest published in the time period 2019-07-20 to 2020-04-20. In order to examine the prevalence of topics over time, I decide to remove the texts from 2018.

2.2 Data Pre-processing

Data pre-processing is arguably one of the key components in the text mining process and crucial for generating a useful topic model. There are some prerequisites in this step i.e. I install jieba[4] and download the stopwords list by the BaiduGuide.

It is noticeable that the raw texts include elements that might add noise to my analysis. So, first, I remove spaces and non-Chinese characters like numerals, English letters, punctuation marks and other symbols. Furthermore, tokenization of raw texts is a necessary standard pre-processing step. Chinese, standardly written without obvious delimiter or marker (like spaces in English) between

[1] Nei.st is a news aggregator website (https://nei.st/) that specialize in fetching Chinese-language news articles. Nei.st daily updates the newly published articles it fetches from various influential and credible news sources based in different regions (for example, Chinese editions of The Wall Street Journal, The Economist, New York Times, and Caixin 财新, Initium 端传媒, Southern Weekly 南方周末). It provides convenience in web scraping by its nature. Its selection of news sources secures news quality and accuracy to a great extent and alleviates potential media bias caused by government censorship, propaganda, and political affiliation of the sources. Those factors would hugely influence my research result. Overall, the benefits of working with the corpus retrieved from Nei.st outweighs drawbacks.

[2] Beautiful Soup is a Python library that transforms the markup into a parse tree that can be easily navigated and searched by specifying tag names. It greatly simplifies the process of online data extraction.

[3] This problem will be discussed later in Discussion.

[4] Jieba is a Python Chinese word segmentation module that can be used in different segmentation modes.

words, requires a more complicated way of segmenting words. I use jieba in its "accurate mode" to cut the sentences into the most accurate segmentations.

Then stopwords, the words that contain no significant information to the document, need to be removed from the token list. I amend the stopwords list provided by Baidu Guide[5] to create a custom stop words list and use it to filter out the stopwords before processing of texts.

3 Methodology

3.1 Construct Document-term Matrix

The result of data pre-processing is a list of texts tokenized into words that can be fed into a vectorizer to construct a document-term matrix **A**. Rows of **A** represent n documents and columns of **A** represent m unique terms present across all articles (i.e., the corpus vocabulary). Although CountVectorizer from Scikit-learn is an option, I apply Term Weighting with term frequency-inverse document frequency (TF-IDF) using TfidfVectorizer to generate matrix **A**. It effectively differentiates rarely and commonly occurring words and gives more weight to the rare terms that characterizes a certain group of documents, improving the performance of topic modeling. Once I have the document-term matrix, I can apply topic modeling algorithms to explore the data.

3.2 Topic Modeling

Topic modeling aims to automatically discover the hidden thematic structure in a large corpus of otherwise unorganized documents. While it often involves the use of LDA, NMF can also be applied and the results have been proved successful[6]. Specifically, applying a log-based TF-IDF weighting factor to the data prior to topic modeling has shown to be advantageous in producing diverse but semantically coherent topics[7]. This makes NMF suitable when the task is to identify both broad, high-level groups of documents, and niche topics with specialized vocabularies. This is particularly desirable in my research, as it can distinguish more focused discussions on major political, economic, and social events from general ones and identify their significance as "topics" over time.

Applying NMF

Applying NMF to the document-term matrix results in a low-rank approximation in the form of the product of two non-negative factors $\mathbf{A} \approx \mathbf{WH}$, where the objective is to minimize the reconstruction error between **A** and **WH**.

The rows of matrix **H** can be interpreted as k topics, defined by non-negative weights for each of the m terms in the corpus vocabulary. Ordering each row of **H** would provide us a topic descriptor, in the form of a ranking of the terms relative to the corresponding topic. The columns of **W** provide membership weights for all n articles with respect to each of the k topics. They can be used to associate individual articles with the topics they are related to, and when we know the publication date of articles, we can thus measure significance of a given topic in a certain time period.

---

[5] It is found that the Baidu stopwords list outperforms ones made by Harbin Institute of Technology and the Machine Learning Laboratory of Sichuan University on improving the result of text clustering of Chinese texts especially news reports (Qin, Deng and Wang, "Chinese Stopwords for Text Clustering: A Comparative Study," 78). I create my own stopwords list based on the Baidu one and will discuss the impact of stopwords list selection later.

[6] Wang et al., "Group matrix factorization for scalable topic modeling."

[7] O'Callaghan et al., "An analysis of the coherence of descriptors in topic modeling."

In practice, I use a fast implementation of NMF provided by Scikit-learn. One key parameter selection decision in topic modeling via NMF pertains to the number of topics *k*.

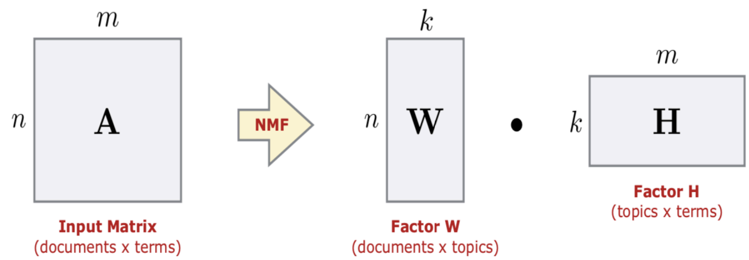

Figure 1. An illustration of NMF.

Parameter Selection

When using topic modeling, I need to specify the number of topics *k*. Choosing too few topics will produce results that are overly broad, while choosing too many will lead to many small, highly overlapped topics. One general strategy has been to compare the topic coherence of topic models generated for different values of *k*. I use a recently proposed measure Topic Coherence via Word2Vec (TC-W2V), which evaluates the relatedness of a set of top terms describing a topic[8].

I process the corpus to build a Skipgram Word2Vec model[9] using Gensim, calculate the individual topic coherence score and the mean of them using the model, and derive the mean coherence score of a topic model. An appropriate k value can be identified by examining a plot of the mean TC-W2V coherence scores for range of *k* and selecting a value corresponding to the maximum mean coherence. As shown in Figure 2, I achieve the highest coherence score = 0.4394 when the corresponding number of topics = 48.

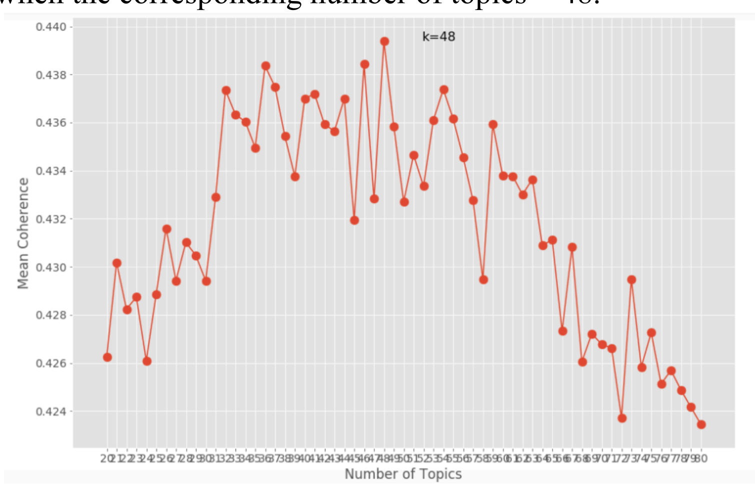

Figure 2.

4 Results and Discussion

With the produced matrices **H** and **W**, we can easily look at the topic descriptors for the 29 topics with lists of top-ranked terms in each and also the snippets for top-ranked documents for each topic (results shown in Figure 3 and 4). They give us a rough sense of the content of the collection. However, visualization is an important and indispensable step to better summarize and

---

[8] Greene and Cross, "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach," 81.

[9] Word2vec model involves computing a set of vector representations for all the terms in the corpus.

interpret the topic model and to effectively communicate and demonstrate the result to readers. Without visualization, topic models would still remain a black box given their complexity.

```
Topic 01: 银行, 贷款, 金融机构, 金融, 监管, 存款, 财新, 银行业, 信用, 央行, 还款, 商业银行, 风险, 客户, 监管部门
Topic 02: 医院, 病人, 医生, 患者, 医护人员, 发热, 床位, 门诊, 收治, 救治, 武汉市, 治疗, 确诊, 重症, 肺炎
Topic 03: 世纪, 历史, 人类, 资本主义, 作者, 年代, 世界, 社会, 时代, 读者, 出版, 战争, 自由主义, 自由, 精英
Topic 04: 供应链, 工厂, 生产, 制造, 供应商, 关税, 工人, 制造业, 跨国公司, 零部件, 贸易战, 贸易, 越南, 库存, 制造商
Topic 05: 计算机, 机器, 人工智能, 算法, 机器人, 技术, 学习, 人类, 传感器, 软件, 智能, 识别, 预测, 数据, 生成
Topic 06: 人口, 年龄, 劳动力, 富裕, 收入, 养老, 养老金, 老年人, 健康, 老龄化, 退休, 家庭, 财富, 寿命, 生活
Topic 07: 海鲜, 冠状病毒, 肺炎, 蝙蝠, 疾控中心, 病毒, 华南, 武汉, 病原体, 样本, 宿主, 检测, 传染病, 动物, 序列
Topic 08: 口罩, 复工, 物资, 疫情, 武汉, 湖北, 防疫, 医用, 防控, 防护服, 春节, 订单, 肺炎, 医疗, 复产
Topic 09: 关税, 协议, 贸易, 特朗普, 农产品, 加征, 第一阶段, 谈判, 贸易战, 达成, 贸易谈判, 莱特, 商品, 中方, 希泽
Topic 10: 朋友, 父亲, 家里, 孩子, 感觉, 生活, 喜欢, 儿子, 家人, 母亲, 父母, 离开, 出门, 晚上, 回来
Topic 11: 电影, 导演, 作品, 观众, 影片, 好莱坞, 奖项, 故事, 演员, 编剧, 剧本, 奥斯卡, 创作, 最佳, 上映
Topic 12: 香港, 抗议者, 抗议, 示威者, 活动, 示威, 香港政府, 警方, 法案, 郑月, 警察, 暴力, 引渡, 内地, 骚乱
Topic 13: 香港, 社团, 福建, 立法会, 传媒, 内地, 示威者, 反修, 建制, 市民, 运动, 特区政府, 特首, 郑月, 基本法
Topic 14: 指数, 股市, 下跌, 跌幅, 标普, 股票, 国债, 投资者, 抛售, 股指, 暴跌, 美联储, 收益率, 道琼斯, 美国股市
Topic 15: 李文亮, 武汉, 疫情, 微博, 媒体, 社交, 医生, 训诫, 微博上, 愤怒, 帖子, 舆论, 武汉市, 冠状病毒, 去世
Topic 16: 腾讯, 用户, 流量, 平台, 电商, 头部, 上线, 互联网, 广告, 品牌, 抖音, 亿元, 财新, 商家, 营销
Topic 17: 品牌, 沃尔玛, 零售商, 销售额, 门店, 亚马逊, 顾客, 销售, 消费者, 商品, 购物, 连锁, 奢侈品, 电商, 零售
Topic 18: 债务, 违约, 债券, 贷款, 银行, 融资, 债权人, 发行, 投资者, 民营企业, 资金, 评级, 偿还, 亿美元, 偿付
Topic 19: 球迷, 火箭队, 比赛, 球员, 莫雷, 球队, 篮球, 推文, 体育, 休斯顿, 言论, 转播, 道歉, 联盟, 赛季
Topic 20: 伊朗, 伊拉克, 袭击, 德黑兰, 导弹, 美军, 卫队, 伊斯兰, 伊朗人, 革命, 发动, 战争, 军事, 基地, 攻击
Topic 21: 迪士尼, 串流, 娱乐, 制作, 订阅, 节目, 游戏, 媒体, 电视, 电影, 音乐, 服务, 频道, 流媒体, 内容
Topic 22: 亿美元, 基金, 投资, 投资者, 收购, 股票, 交易, 上市, 市值, 股价, 软银, 资产, 利润, 资本, 估值
Topic 23: 习近平, 中共, 领导人, 毛泽东, 领导, 党内, 邓小平, 政治, 会议, 权力, 主席, 中共中央, 北京, 总书记, 党中央
Topic 24: 汽车, 万辆, 电动车, 车型, 电动汽车, 销量, 制造商, 通用汽车, 新能源, 特斯拉, 汽车销量, 电池, 品牌, 车企, 电动
Topic 25: 排放, 气候变化, 气候, 变暖, 二氧化碳, 减排, 排放量, 燃料, 化石, 巴黎, 气体, 煤炭, 能源, 温室, 干旱
Topic 26: 植物, 汉堡, 产品, 肉类, 食品, 生产, 种植, 餐厅, 口感, 食物, 农场, 制成, 味道, 含有, 牛肉
Topic 27: 猪肉, 生猪, 猪瘟, 非洲, 价格, 上涨, 扑杀, 养殖, 价格上涨, 农业, 农民, 肉类, 食品, 涨幅, 通胀
Topic 28: 中国外交部, 驱逐, 华尔街日报, 外交部, 中国政府, 外国, 新闻, 记者, 发言人, 北京, 媒体, 报道, 声明, 美国国务院, 签证
Topic 29: 台湾, 英文, 选举, 国民党, 两岸, 选民, 大陆, 韩国, 台湾人, 总统, 候选人, 台北, 香港, 大选, 选票
Topic 30: 社会, 理解, 体制, 逻辑, 权力, 治理, 制度, 群体, 讨论, 关系, 民众, 层面, 政治, 思考, 话语
Topic 31: 土地, 项目, 城市, 面积, 建筑, 业主, 建设, 住宅, 南方周末, 房屋, 规划, 平方米, 用地, 地块, 记者
Topic 32: 货币, 利率, 央行, 美联储, 美元, 货币政策, 汇率, 降息, 通胀, 贬值, 人民币, 债券, 资产, 经济体, 欧洲央行
Topic 33: 用户, 隐私, 数据, 谷歌, 加密, 科技, 广告, 微软, 服务, 信息, 网络, 伯格, 数字, 系统, 服务器
Topic 34: 欧盟, 欧洲, 英国, 脱欧, 德国, 法国, 成员国, 贸易, 意大利, 俄罗斯, 塞尔维亚, 一带, 协议, 国际, 欧盟委员会
Topic 35: 石油, 沙特, 原油, 油价, 俄罗斯, 能源, 天然气, 王储, 产量, 减产, 美元, 石化, 储量, 油气, 沙特阿拉伯
Topic 36: 新疆, 穆斯林, 维吾尔, 维吾尔族, 拘禁, 少数民族, 宗教, 极端主义, 人权, 伊斯兰教, 民族, 关押, 教育, 文件, 学员
Topic 37: 学生, 学校, 教育, 课程, 大学, 老师, 学院, 学习, 毕业生, 毕业, 商学院, 小学, 教室, 课堂, 该校
Topic 38: 财新, 亿元, 发改委, 产能, 建设, 万吨, 记者, 市场化, 电价, 全国, 改革, 煤炭, 省份, 发电, 产业
Topic 39: 太空, 地球, 月球, 轨道, 发射, 火星, 卫星, 火箭, 宇宙, 太阳, 探索, 人类, 实验室, 载人, 登陆
Topic 40: 华为, 芯片, 设备, 英特尔, 智能手机, 半导体, 任正非, 技术, 电信, 科技, 供应商, 巨头, 美国公司, 服务器, 通信
Topic 41: 飞机, 飞行, 航空, 航空公司, 波音, 乘客, 飞行员, 发动机, 客机, 空中, 航班, 航空业, 机场, 电动, 公里
Topic 42: 警方, 集结, 反对, 蒙面, 公众, 集会, 未经, 游行, 法庭, 非法, 禁止, 批准, 比例, 基本法, 废除
Topic 43: 药物, 细胞, 疫苗, 治疗, 临床试验, 基因, 试验, 研发, 人体, 疗法, 科学家, 免疫, 疾病, 制药, 蛋白质
Topic 44: 总统, 特朗普, 国会, 民主党, 竞选, 共和党, 参议员, 参议院, 选举, 弹劾, 众议院, 白宫, 投票, 法案, 担任
Topic 45: 股东, 股权, 安邦, 持股, 亿元, 股份, 有限公司, 控股, 出资, 集团, 亿股, 增资, 董事, 持有, 董事长
Topic 46: 病例, 疫情, 病毒, 感染, 隔离, 新冠, 确诊, 死亡, 传播, 措施, 意大利, 冠状病毒, 患者, 人数, 入境
Topic 47: 增速, 增长, 同比, 经济学家, 下降, 放缓, 百分点, 下滑, 下调, 预期, 预计, 增幅, 数据, 刺激, 消费
Topic 48: 法院, 法律, 案件, 律师, 司法, 诉讼, 涉嫌, 起诉, 法庭, 刑事, 犯罪, 申请, 调查, 指控, 条款
```

Figure 3.

01. 煤炭大省山西欲转型「氢谷」，但看似成本低廉的煤制氢路径却充满争议和障碍山西省欲打造中国「氢谷」。这个听起来十分时髦的词汇，并不仅仅是山西产业转型选择的方向。广东佛山、山东济南、吉林白城等全国 20 多个地区都在陆续喊出类似的口号，只不过山西省打造「氢谷」的制氢路径不同——煤制氢。作为产煤大省，山西省

02. 大数据风控模式之外，还有依赖客户经理队伍线下搜集风控信息、线上不断完善模型来有效服务小微企业的另一种模式借力金融科技的联合贷款或助贷，是普惠信贷的惟一解决方案吗？答案是不。过去五年间，尽管中国数字金融服务在规模和广度上都取得了令人瞩目的成就，但数字技术的发展并未能完全解决金融服务对中低收入人群的覆盖

03. 欧洲是华为的「粮仓」和「第二本土」。面对美国政府对华为的制裁，欧洲人会作出怎样的选择？瑞士最大城市苏黎世的郊区，电信运营商 Sunrise 总部，前来参观的人络绎不绝。10 月 14 日，在全球移动宽带论坛开幕的前一天，Sunrise 在这里展示了与华为合作的欧洲首家 5G 联合创新中心，用于孵化云

04. 如果你之前认为贸易战对跨境贸易而言很糟糕……集装箱船领航员、海关官员、物流专家、卡车司机和仓库夜间值班人员：所有这些人都擅长对付和国际贸易相关的麻烦，不论是罢工还是贸易战。但是，随着有关今年全球 GDP 大跌的预测浮现，即便是他们最有创意的想法也无法维持 25 万亿美元的商品和服务在世界各地继

05. 一位陨落的企业明星反省自己在德国商业中的失败经历传媒集团贝塔斯曼的前老板、曾经在柏林乃至好莱坞都广受赞誉的托马斯·米德尔霍夫 (Thomas Middelhoff) 于 8 月 20 日发布了新书《有罪》(Guilty)。这并不是法律意义上的认罪——他仍然认为自己因逃税和违反信托获刑三年的判罚过重了

06. 更为微妙的控制依然存在中国西北部喀什的一所穆斯林再教育中心最近刚刚清空，数百个废弃的金属床架杂乱地堆放在草地上。床架上的红色贴纸上写着：识错，认错，悔过。中国官员说，这类中心 (北京方面描述为职业技术学校) 的学员已经全部结业。人权组织及西方政府则称，近年来，遍布新疆的数十所此类再教育中心曾扣押约

07. 3000 亿元央行专项再贷款，如何投向新冠肺炎疫情防控最急需资金的企业，如何防范利益博弈和道德风险，阳光化运作最关键2 月 4 日，汉口银行汉阳支行的客户经理李卓俊戴着口罩出门了。此时，他所在的湖北省武汉市的街道上已空空荡荡，路上有人拦车询问，他说：「我是去办贷款业务的。」李卓俊只身赶往的，是九州通

08. 漏洞都在 TikTok 系统的核心部分根据以色列网络安全公司 Check Point 周三发布的研究报告，在全球拥有上亿用户、深受青少年喜爱的智能手机应用 TikTok 存在严重的安全漏洞，黑客可以利用这些漏洞操纵用户数据并泄露个人信息。这些漏洞可以让攻击者向 TikTok 用户发送带有恶意链接的消

09. 金融机构的风险根源，多在公司治理上——大股东占款、内部人控制，甚至形成「内部人+腐败官员+不良企业」的「黑三角」，应及时识别、及早预案近日，甘肃银行 (02139.HK) 处于风口浪尖，不过在当地货币监管当局、地方政府、大股东的帮助、维护之下，局面已趋稳定。4 月 1 日，甘肃银行股价跳水 43.4

10. 短期数据的快速增长与市场的高度关注并不代表产品的成功，合理有效地均衡各方利益与诉求才是产品发展的关键近期，平安健康推出的「i 动保医疗」，又一次搅动了健康险的市场。在价格战激烈的氛围下，平安健康为这款百万医疗险标注的价格是——免费领：凡 18—50 周岁、有社保的用户，只要加入就可领取有效期一个月的

Figure 4.

4.1 Visualization

Top Term Weights in Topics with Bar Chart

We have looked at the topic descriptors with the rankings of terms in each topic. However, they do not show the strength of association for the different terms in a given topic. We can represent the distribution of the weights for the top terms in a topic using a matplotlib horizonal bar chart (shown in Figure 5) or pyLDAvis.

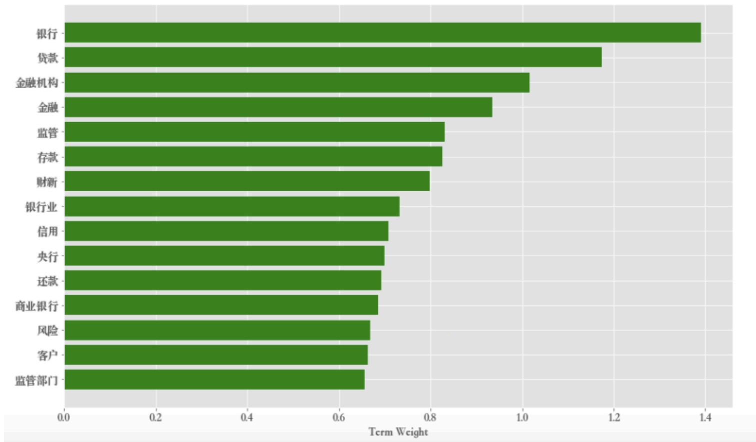

Figure 5. Weights for the top 15 terms in topic 1.

PyLDAvis[10]

First, same as Figure 5 shown above, pyLDAvis allows us to select a topic to reveal the most relevant terms for that topic. In Figure 6, Topic 48 is selected, and its 30 most relevant terms populate the bar chart to the right (ranked in order of relevance from top to bottom).

Second, on the left panel, the visual features provide a global perspective of the topics and allows us to verify if the topic model is a good one. The areas of the circles are proportional to the relative prevalences of the topics in the corpus. A good topic model should have fairly big, non-overlapping bubbles scattered throughout the plot instead of being clustered in one quadrant[11]. In this 48-topic model fit to the news articles corpus, fairly big bubbles spread out on the whole space, but several of them do overlap with one another, implying the model can be improved in future study.

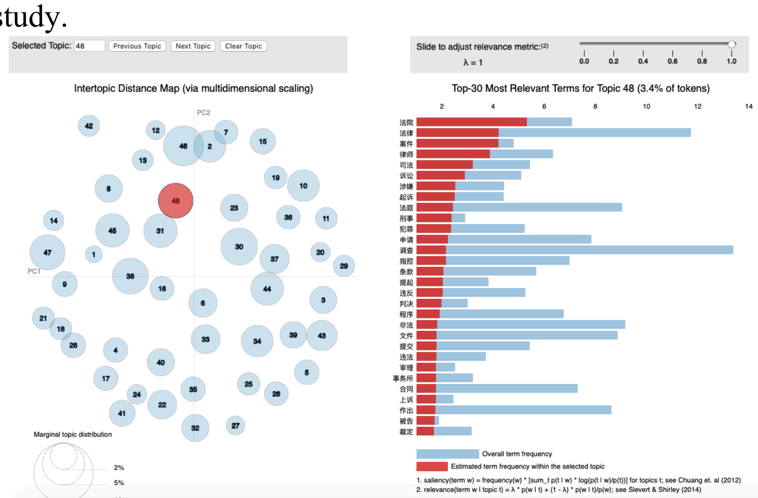

Figure 6.

[10] PyLDAvis is a Python library for web-based interactive visualization of usually a fitted LDA topic model, but it can also be used on NMF models. It helps us to interpret the topic model by providing a global view of the topics (and how they differ from each other) while also allowing for a deep inspection of the terms most highly associated with each individual topic.

[11] Tunazzina, "Yoga-Veganism: Correlation Mining of Twitter Health Data," 5.

In addition, pyLDAvis allows us to select a term (by hovering over it) to reveal its conditional distribution over topics, a feature I utilize as an indicator of the direction of observation on topic significance over time. For example, in Figure 7, “李文亮” (Li Wenliang), the most relevant term for topic 15, is selected. In the majority of this term’s occurrences, it is drawn from 2 topics located in the upper right-hand region of the global topic view: topic 15 and 2. Upon inspection, this group of topics can be interpreted broadly as a discussion of spreading of COVID-19 within China. It somewhat suggests that I can further investigate if the trends of topic significance for the two topics follow similar contours.

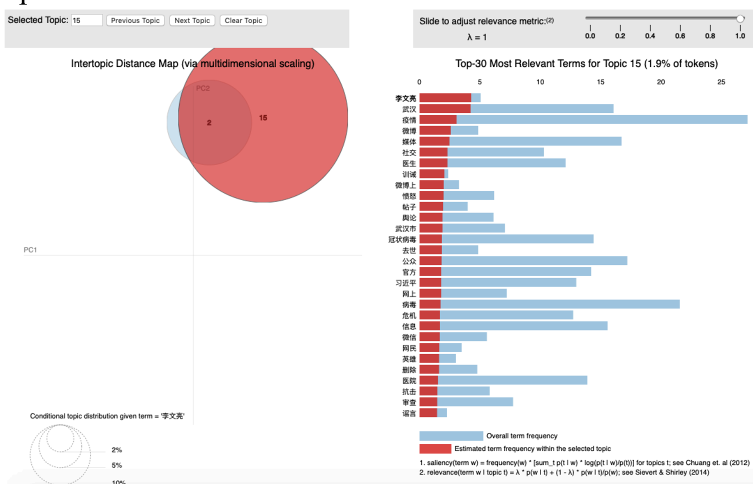

Figure 7.

Topic Prevalence over Time

Topic prevalence over time is worth exploring because it often serves as a mechanism for identifying spikes in discourse and for depicting the relationship between the various discourses in a corpus. Those are information of my interest. Topic prevalence over time is not, however, a measure returned with the standard modeling tool. In order to quantify it, I shift focus from topic composition in terms of words to document composition in terms of topics and perform some computations.

The method I use is calculating normalized or proportional weights of topics. First, I divide the 275 days from 2019-07-20 to 2020-04-20 (the time period in which our texts are produced) into 18 15-day periods. I identify the 15-day period in which an article was published and sum up the weights of articles published in the same period by topic. Then I normalize those values by dividing them by the sum of all the weights in that period so that they total to 1.

For my relatively small corpus comprised of a wide range of content, a stacked bar chart (Figure 7) and an area plot (Figure 8) provide a nice overview of variation in topic prevalence over time.

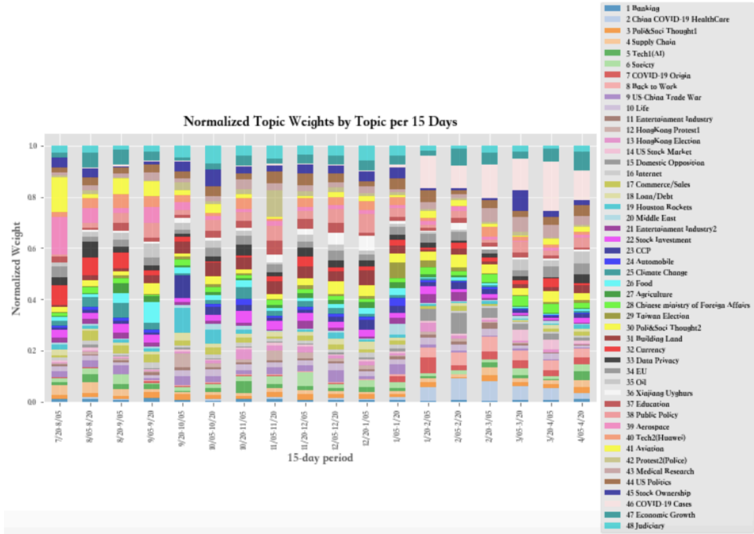

Figure 7.

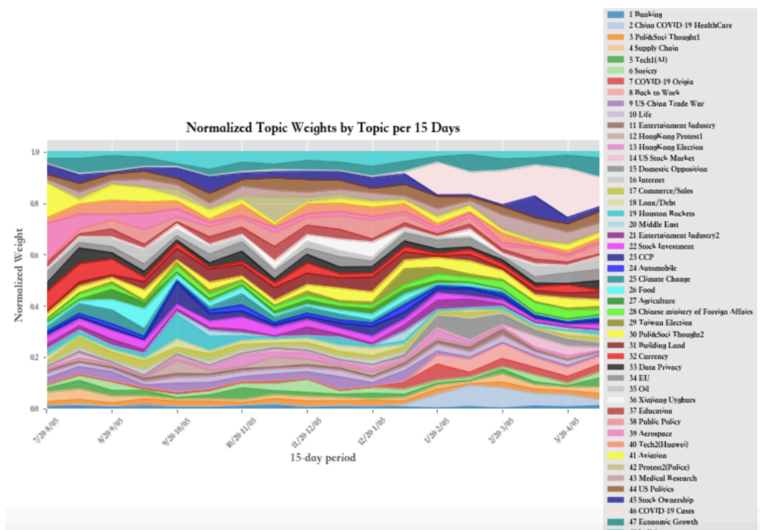

Figure 8.

Would a certain group of topics show any relationships in terms of topic prevalence over time? Following up on the implication of pyLDAvis visualization, I plot out the change of normalized topic weights for topic 2 and 5 and notice their positive correlation in topic significance from 01-20-2019 to 02-20-2020. The topics based on news coverage clearly suggests that domestic opposition against Chinese regime reached its peak when the COVID-19 started to reveal itself to be deadly and highly infective in China but the population didn't receive adequate government alert. The sudden decrease in domestic opposition since late February could be a result of timely crackdown and censorship by the Chinese government[12].

[12] Ruan, Knockel, and Crete-Nishihata, "Censored Contagion: How Information on the Coronavirus Is Managed on Chinese Social Media."

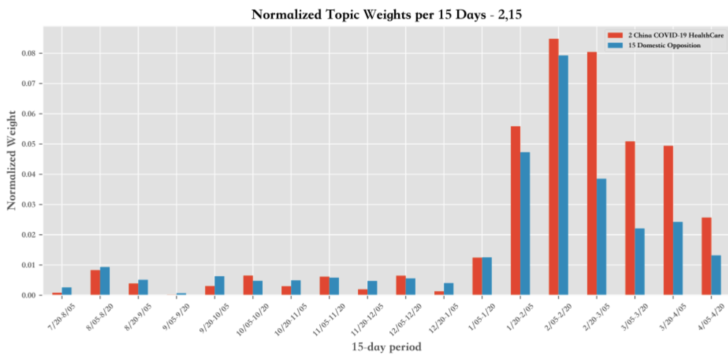

Figure 9.

4.2 Discussion

- Stopwords selection greatly influences the performance of a topic model. In future research, I will keep amending and improving my custom word list, for example, by adding in proper nouns specifically for the corresponding corpus like “蔡英文”(Tsai Ing-wen) and “钻石公主号”(Diamond Princess). And I will further investigate certain stopwords’ influence on model building.
- Due to time constraint, I wasn’t able to try topic modeling via LDA. But LDA with Mallet is a good alternative of NMF and worth testing, especially in terms of improving model coherence.
- As described in my project proposal, I expected the corpus consisting of news articles from the whole period of 2018-2020, but later I realized that Nei.st didn’t start consistently fetching news until late July 2019. As a result, my actual corpus is fairly small. This misjudgment should be taken as a lesson in future data collection. And I will increase the corpus size by extracting more news articles to train a better model. I will divide the corpus into training and testing set to further observe the model behavior for evaluation.
- It would be great to design a tool for interactive topic prevalence tracking and visualization that is compatible with python in the future. Below are among the difficulties I encountered in tracking and visualizing topic prevalence: 1) I have no way to easily detect connections between topic prevalences of topics but by my own perception and reasoning; 2) I can only observe topic prevalences of a group of topics by repetitively plotting them out. No existing python library is specialized in helping to track and visualize topic prevalence so modelers can only show it by static plots which restricts representation and interpretation in a variety of ways.

5 Conclusion

In this study, I apply topic modeling to discover hidden topics and explore variation of topic prevalence over time. Firstly, I outline the process of data collection and text pre-processing specifically for Chinese news reports. Subsequently, I introduce non-negative matrix factorization and utilize it to a corpus of all $\approx 1.5k$ Chinese language news articles from multiple news outlets between July 2019 to April 2020. The topic modeling method allows me to unveil both niche topics related to individual major events and broader topics related to everyday life and certain industries. Finally, I employ different visualizations on the model and discover topic prevalences over time and interesting correlation between prevalences of different topics.

References

Greene, Derek, and James P. Cross. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." Political Analysis 25, no. 1 (2017): 77–94. doi:10.1017/pan.2016.7.

Guan, Qin, Sanhong Deng and Hao Wang. "Chinese Stopwords for Text Clustering: A Comparative Study." Data Analysis and Knowledge Discovery 1 (2006): 72-80.

Islam, Tunazzina. "Yoga-Veganism: Correlation Mining of Twitter Health Data." ArXiv abs/1906.07668 (2019): n. pag.

O'Callaghan, Derek, Derek Greene, Joe Carthy, and Pádraig Cunningham. "An analysis of the coherence of descriptors in topic modeling." Expert Syst. Appl. 42, 13 (2015): 5645–5657. doi: /10.1016/j.eswa.2015.02.055.

Ruan, Lotus, Jeffrey Knockel, and Masashi Crete-Nishihata. "Censored Contagion: How Information on the Coronavirus Is Managed on Chinese Social Media." The Citizen Lab, March 4, 2020. https://citizenlab.ca/2020/03/censored-contagion-how-information-on-the-coronavirus-is-managed-on-chinese-social-media/.

Wang, Quan, Zheng Cao, Jun Xu, and Hang Li. "Group matrix factorization for scalable topic modeling." In Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12). Association for Computing Machinery, New York, NY, USA, 375–384. doi:10.1145/2348283.2348335.