CORD-19 数据分析 & 可视化

吴双 10164102141

问题描述

本实验研究 COVID-19 新型冠状病毒相关科研文献的重要性的分布和情绪指数随时间的变化。

数据集描述

CORD-19 数据集：CORD-19 数据集是一个开源学术文献数据集，包含了所有研究 COVID-19 新冠病毒的学术文献数据集。其中,我们选取 biorxiv 的文献，共 89486 篇作为研究对象。

问题动机

2020 年的新型冠状病毒是一场非常大规模的传染病，全世界科研人员都在紧急应对，2020 年伊始至今八个月时间，科研文献出现了井喷，我们进行对科研文献的情绪分析相关研究。

在这项研究中,可以发现 COVID-19 相关研究的一些发展趋势，为之后的科研工作的方向起到一定辅助提示作用。

实验方法与细节

由于这项研究在之前没有先例，所以我们根据一般的数据挖掘与商业分析步骤对数据集进行相应处理和分析。在这次试验中，我们将实验过程分为以下几个部分：

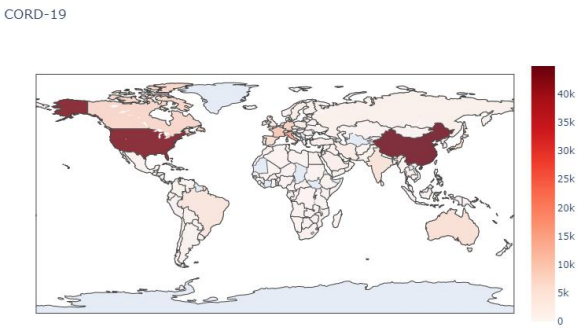
信息检索

首先使用外接 api 和爬虫对文献的信息进行检索(时间,简介,作者(所处国家)):

- 使用 Geopy + Google Cloud Platform (Google Maps API)对各个机构进行定位(由于大部分作者署名习惯只写邮箱不写地址机构,所以检索极其模糊,且由于网络 I/O 限制，速度极慢,待完成的工作中对具体地区的研究难以进行,最终将地点占有率从不通过 API 的 55.07%提高至 62.31%); (尝试通过姓名预测国家的方法是错误的)
- 使用 Gscholar 对文献的年份进行检索(由于 scholar google 对频繁访问的 IP 地址实施了反爬虫机制,所以在中间使用 shadowsocks 对 ip 地址进行了遮挡,但由于大部分文献都有名称或作者名缺损情况,最终将年份占有率提高至 92.70%)

观察分布

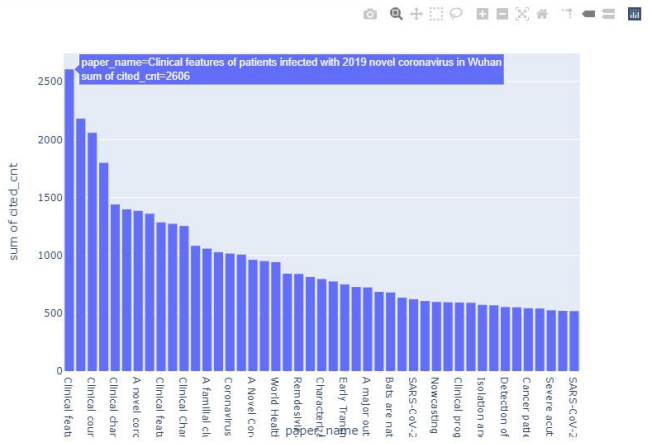
文献数量在地理位置上的表现如右图所示。这里可以看出，美国和中国两国的文献数量最多，均高达 40k 以上。由此可以看出，在这次疫情研究中，中美两国均做出了较为突出的贡献。



文献重要性的标准

我们引入一种较为普遍的评价文献重要性的标准：按照文章在 COVID-19 研究论文中的被引数，显示文献对此次疫情研究的贡献程度。其中做出贡献数超过 1000 的文章如右图所示。

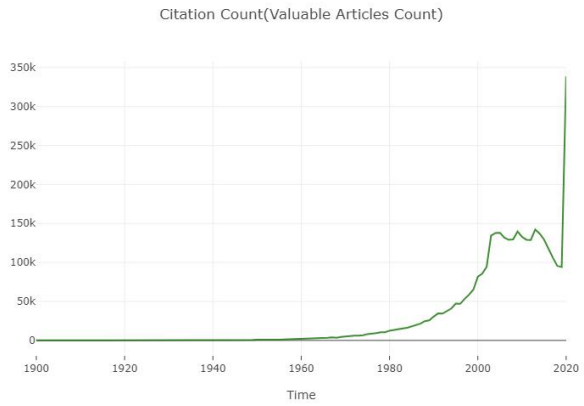
	paper_name	cited_cnt	year
3224	Clinical features of patients infected with 2019 novel coronavirus in Wuhan	2606	2020
760	Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study	2180	2020
288	Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study	2058	2020
2488	A pneumonia outbreak associated with a new coronavirus of probable bat origin	1799	2020
296	Clinical characteristics of coronavirus disease 2019 in China	1439	2020
636	Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia	1397	2012
2174	A novel coronavirus from patients with pneumonia in China	1384	2019
7098	Identification of a novel coronavirus in patients with severe acute respiratory syndrome	1309	2003
5183	Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China	1284	2020
9134	A novel coronavirus associated with severe acute respiratory syndrome	1272	2003
10738	Clinical Characteristics of Coronavirus Disease 2019 in China	1253	2020
1039	Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding	1082	2020
1023	A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster	1058	2020
568	Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia	1027	2020
9228	Coronavirus as a possible cause of severe acute respiratory syndrome	1015	2003
2917	Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus	1006	2003
1857	A Novel Coronavirus from Patients with Pneumonia in China	961	2019
5907	A new coronavirus associated with human respiratory disease in China	950	2020
4682	World Health Organization	941	2019
11958	Pathological findings of COVID-19 associated with acute respiratory distress syndrome	841	2020
6997	Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro	839	2020
1031	Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study	813	2020
9133	Characterization of a novel coronavirus associated with severe acute respiratory syndrome	795	2003
4346	Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China	774	2020
1890	Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia	749	2020
4914	Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation	726	2020
2106	A major outbreak of severe acute respiratory syndrome in Hong Kong	722	1996
1748	COVID-19: consider cytokine storm syndromes and immunosuppression	684	2020
649	Bats are natural reservoirs of SARS-like coronaviruses	679	2005
4114	The molecular biology of coronaviruses	634	1997
5437	SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor	621	2020



值得注意的是，其中较多是关于中国医疗条件和隔离措施的看法与探讨；对基本结构进行研究的文章大部分集中在 18 年之前。不过由于这些研究还都处于初期阶段，所以这一现象其实不算意料之外。

在之后的处理中我们对文献被引数量在地理上的分布再次作图，发现数据集中有关地域信息大量丢失，遂放弃这一方面的研究，在此呼吁广大科研工作人员在署名的时候可以更加明确一点。当然这里需要考虑科研工作人员跨地域的合作或者交流的情况，我们在之后的工作中希望可以引入各机构的科研人员流动数据，当然这是一项极大的工程，需要很大的工作量。

对时间作图，结果如右图所示。可以看出科研贡献度并不是一个逐年上升的趋势，而是有一定的波动性的。在 2019 年前后甚至处于一个较为低谷的状态。直到 2020 年新冠疫情爆发，也自然出现了井喷式增长。



情绪分析判断

我们引入简单的神经网络对摘要进行训练，对文献的摘要进行情绪判断。神经网络的结构如下图所示：

Model: "model"

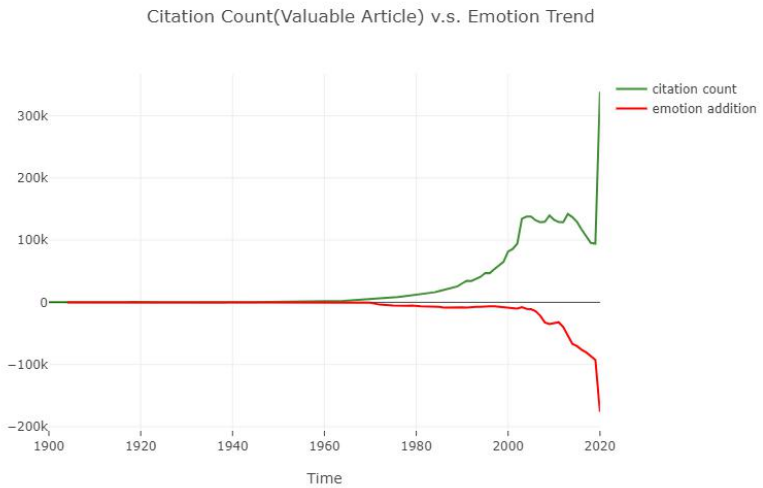
Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 100)]	0	
embedding (Embedding)	(None, 100, 500)	5000000	input_1[0][0]
spatial_dropout1d (SpatialDropo	(None, 100, 500)	0	embedding[0][0]
bidirectional (Bidirectional)	(None, 100, 256)	644096	spatial_dropout1d[0][0]
conv1d (Conv1D)	(None, 98, 64)	49216	bidirectional[0][0]
global_average_pooling1d (Globa	(None, 64)	0	conv1d[0][0]
global_max_pooling1d (GlobalMax	(None, 64)	0	conv1d[0][0]
concatenate (Concatenate)	(None, 128)	0	global_average_pooling1d[0][0] global_max_pooling1d[0][0]
dense (Dense)	(None, 2)	258	concatenate[0][0]

Total params: 5,693,570
Trainable params: 5,693,570
Non-trainable params: 0

可以看出这是一个较为简单的神经网络。我们使用 Twitter 中被人工标签化的 tweets 进行情感二分类模型的训练，经过 1 epoch 的训练，验证集上的准确率达到 76.72%。对文献的标题和摘要进行情绪分析，得出的结果进行时间上切片累和，并向后累和，最终得到情绪随时间波动图。

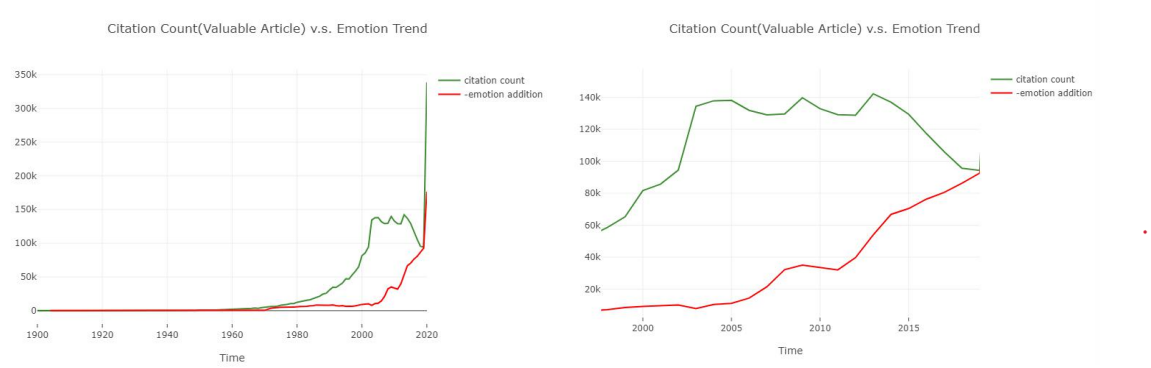
将贡献度变化曲线和情绪变化曲线随时间的变化情况置于同一个折线图中进行比较。展示的结果如下：

可以看出，整体贡献度随时间变化趋势和情绪指数随时间变化趋势相反，有效文献贡献度呈总体上升趋势而情绪指数总体呈下降趋势。这里认为这两点因素具有同源相关性。



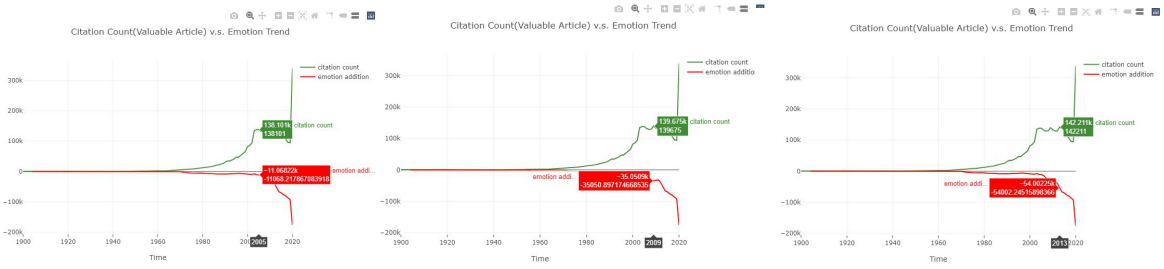
其中，我们将时间尺度缩

小，对 2000 至 2020 这 20 年间的两条曲线的关系进行比较。



为了更好比较对应关系，我们将情绪曲线取反，将两个曲线的总体趋势设为相同之后更加可以比较细节上的对应关系，结果如上图所示。

较为明显的，我们可以看出在 2005 年、2009 年以及 2013 年出现了三次贡献度的高峰，而情绪上也均出现了较为明显的低谷，可以发现具有一定的对应性。



在这一点上其实也并不难理解：

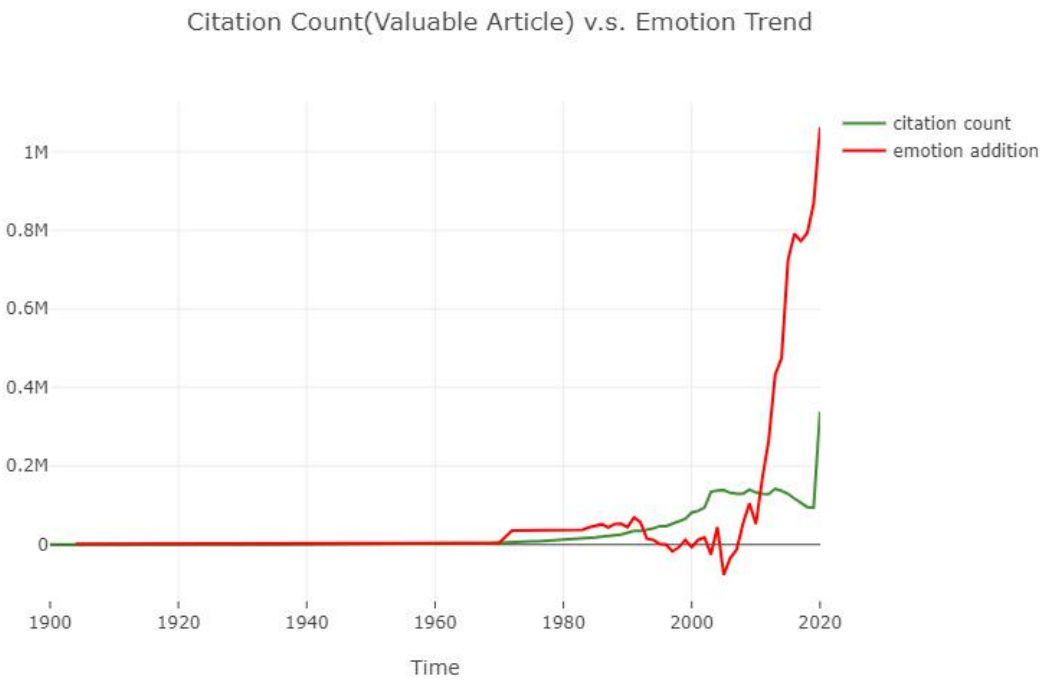
2005 年、2009 年和 2013 年之前均发生过一次大规模的流感，分别是 2003 年的非典型肺炎、2008 年的甲型 H1N1 流行性感冒以及 2013 年的甲型 H7N9 流行性感冒。

这里我们理解为，在每一次流感结束，社会整体的情绪都处于一个较为焦虑且悲观的状态，而科研也受到这一情绪波动的影响，所以会出现较早期的一个悲观情绪的涌现；而且，由于疫情的传播，使得科研所需的数据更加易得，社会对科研工作的需求更加急切。以上两点，都为科研贡献度的三次波峰提供了理由。

新的排布方式

我们引入一种新的排布方式：按照（贡献度*情绪）对时间进行排布。

由于我们之前的结论里认为这两点是有同源相关性的，所以初步认为这一方法可以看作线性相关相乘，对结果没有太大影响。



结果和我们所设想的并不一样，这里没有反转曲线，却出现了情绪曲线失控地上升的情况。

这里我暂时的理解为在 2003 年抗非典成功以后，相关领域的优秀科研人员的科研信心在稳步增加。具有重要贡献度的科研成果大部分都会回归于一种较为乐观的情绪。之前的相反结果很可能是因为大量贡献度较低的工作呈消极态度。

实验结果总结

1. 中美两国是抗疫研究第一线的两个大国；

2. 研究尚处于初期阶段，还需要时间；
3. 优秀的科研人员在成功后的正反馈更强烈；
4. 请善待每一个医疗相关工作者或研究员。

未来工作

1. 未来工作中，我会尝试更详细的检索方式，争取对具体区位和结果做出具体分析；
2. 多选几个 SOTA 神经网络再试试；
3. 避免在预处理的时候和内存问题纠缠不清。