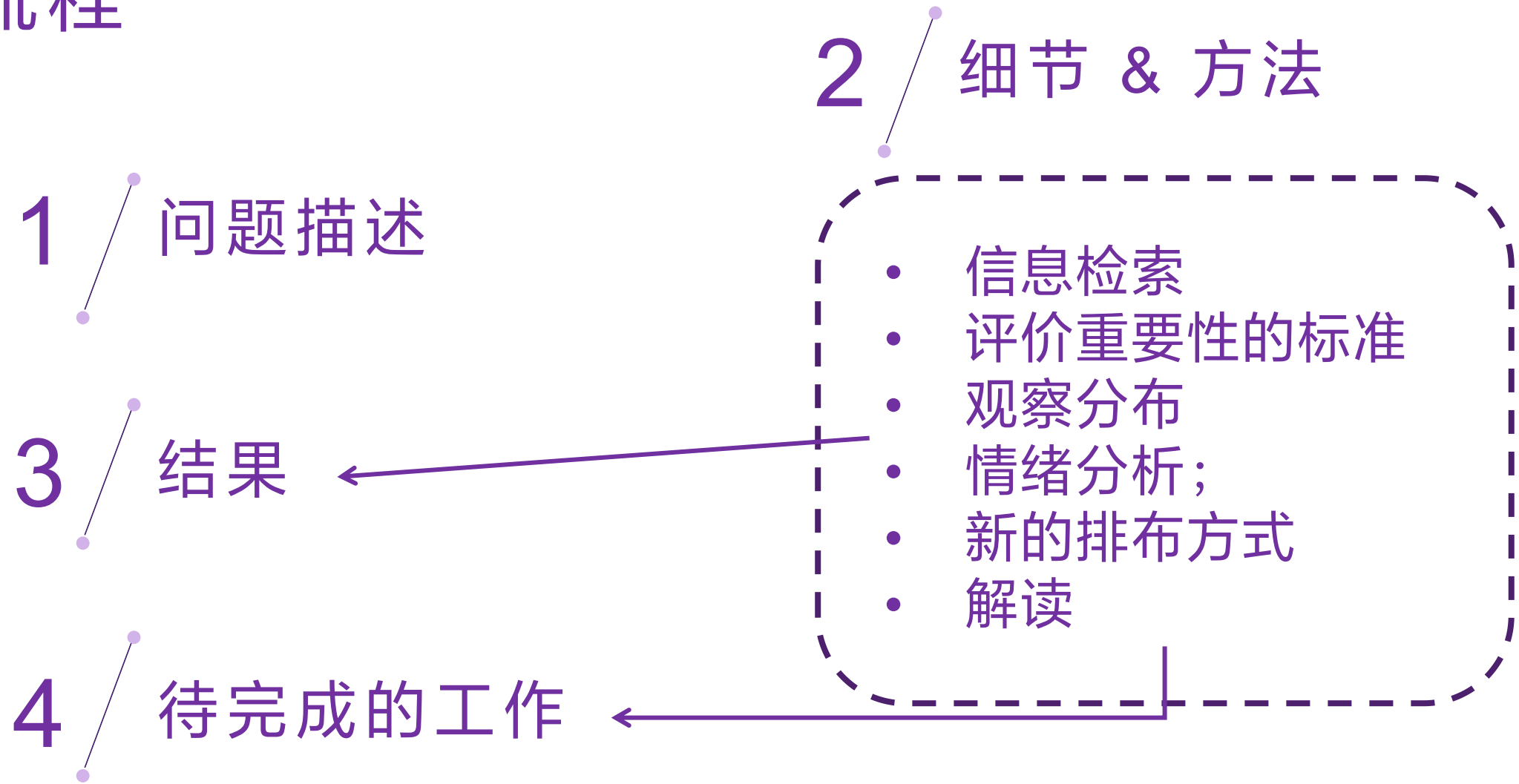
CORD-19　数据分析　&　可视化

流程



1 问题描述

CORD-19 数据集：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。

用以研究 COVID-19 新型冠状病毒相关科研文献的重要性和随时间的情绪变化。

1 问题描述 / 动机

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

2 细节 & 方法

1. 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））：

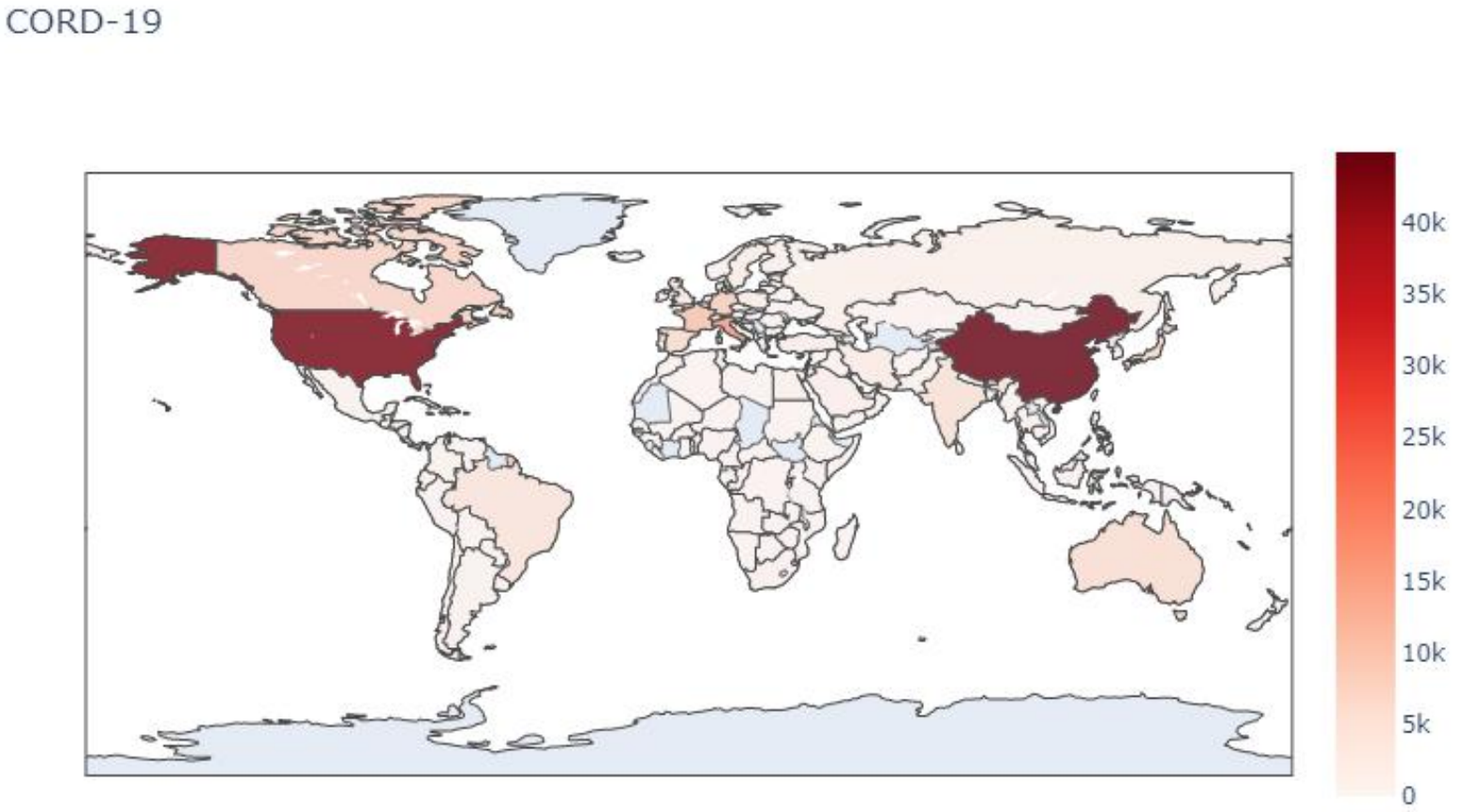
   - 使用 Geopy + Google Cloud Platform (Google Maps API)对各个机构进行定位（由于大部分作者署名习惯只写邮箱不写地址机构，所以检索极其模糊，且由于网络I/O限制，速度极慢，待完成的工作中对具体地区的研究难以进行，最终将地点占有率从不通过API的**55.07%**提高至**62.31%**）；

     （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%**）

2 细节 & 方法

1. 文献数量在地理位置上的表现：



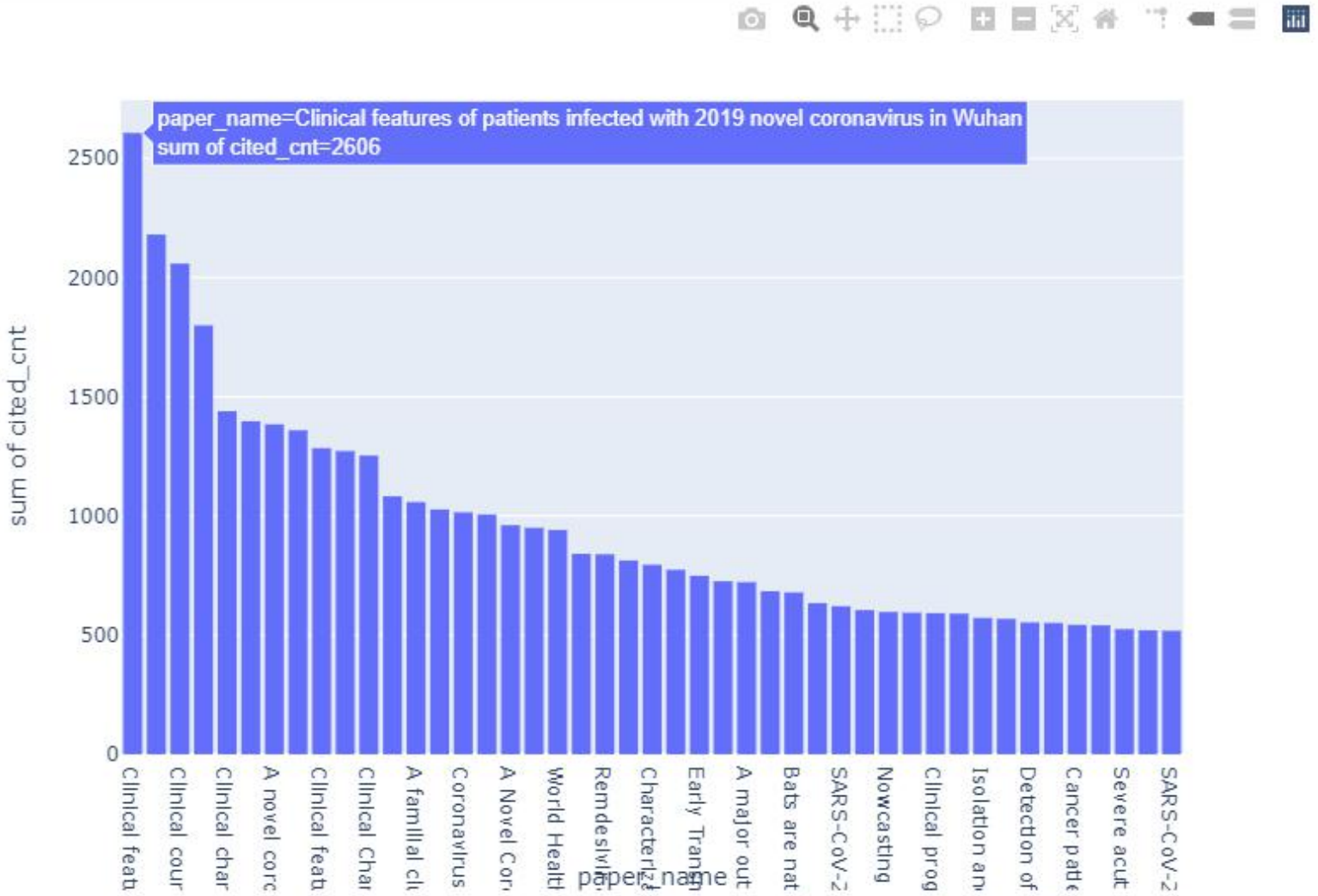
2 细节 & 方法

2. 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，显示文献的重要性。

2 细节 & 方法

	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	1359	2003
5183	Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China	1284	2020
9154	A novel coronavirus associated with severe acute respiratory syndrome	1272	2003
10759	Clinical Characteristics of Coronavirus Disease 2019 in China	1253	2020
1030	Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding	1082	2020
1523	A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster	1058	2020
508	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
1887	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
1051	Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study	813	2020
9155	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
2156	A major outbreak of severe acute respiratory syndrome in Hong Kong	722	1986
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

2 细节 & 方法



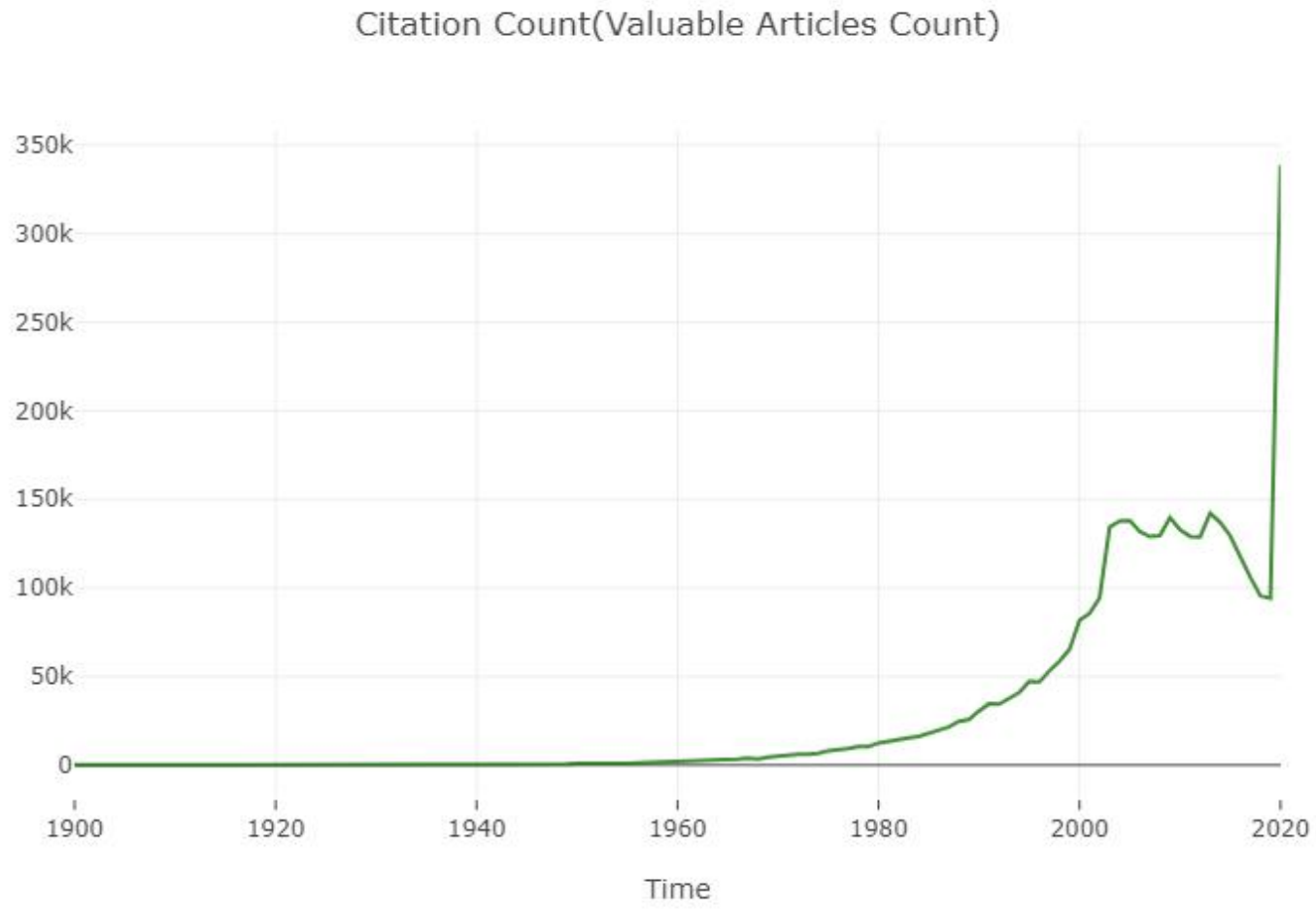
较多是关于中国医疗条件和隔离措施的看法与探讨；
对基本结构进行研究的文章大部分集中在18年之前

不算异常：初期阶段

2 细节 & 方法

3. 对文献被引数量在地理上的分布再次作图，发现大量丢失，故放弃，选则对时间作图

2 细节 & 方法


Citation Count(Valuable Articles Count)
350k
300k
250k
200k
150k
100k
50k
0
1900
1920
1940
1960
1980
2000
2020
Time

2 细节 & 方法

4. 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。

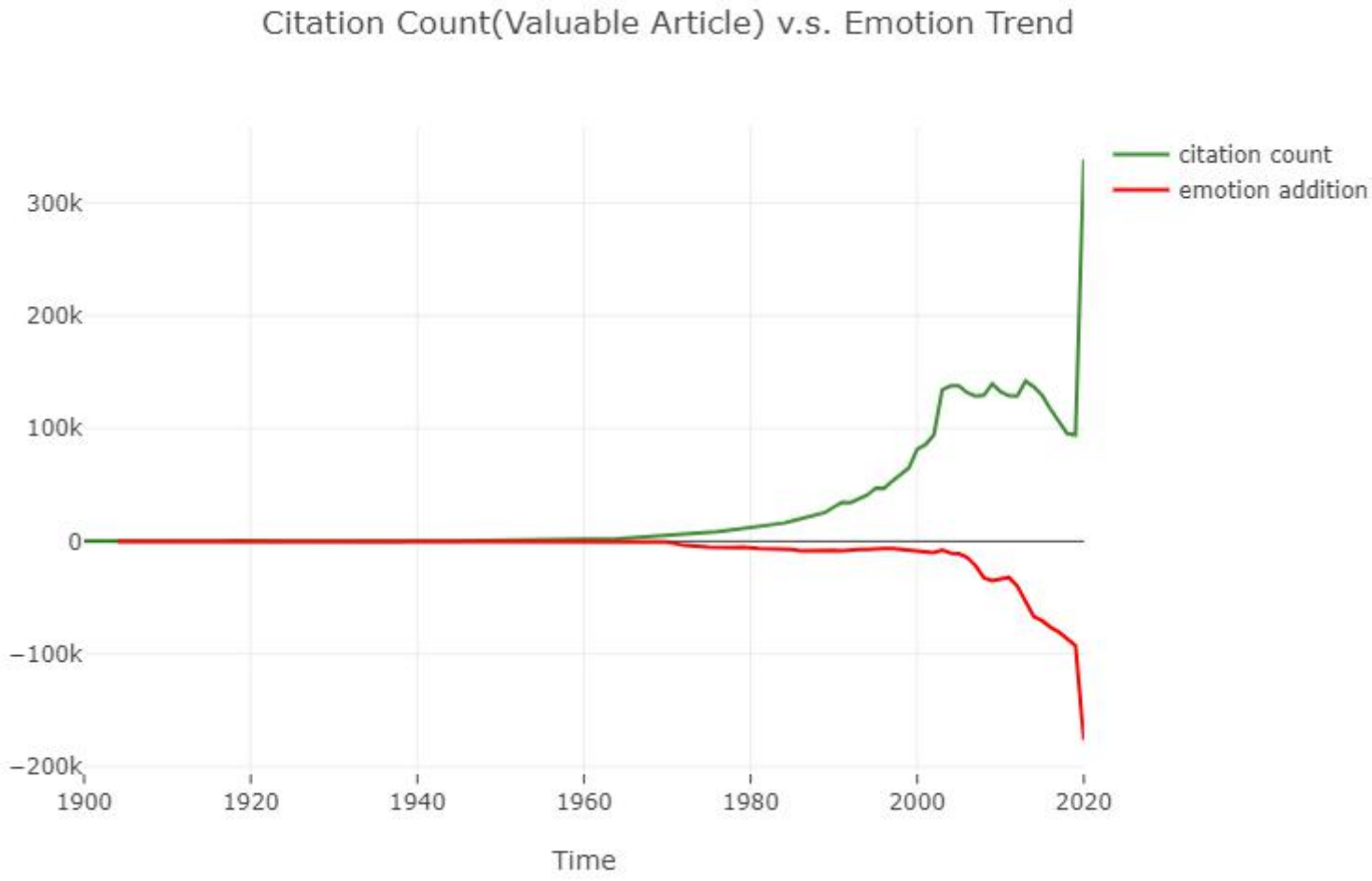
Model: "model"

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

2 细节 & 方法

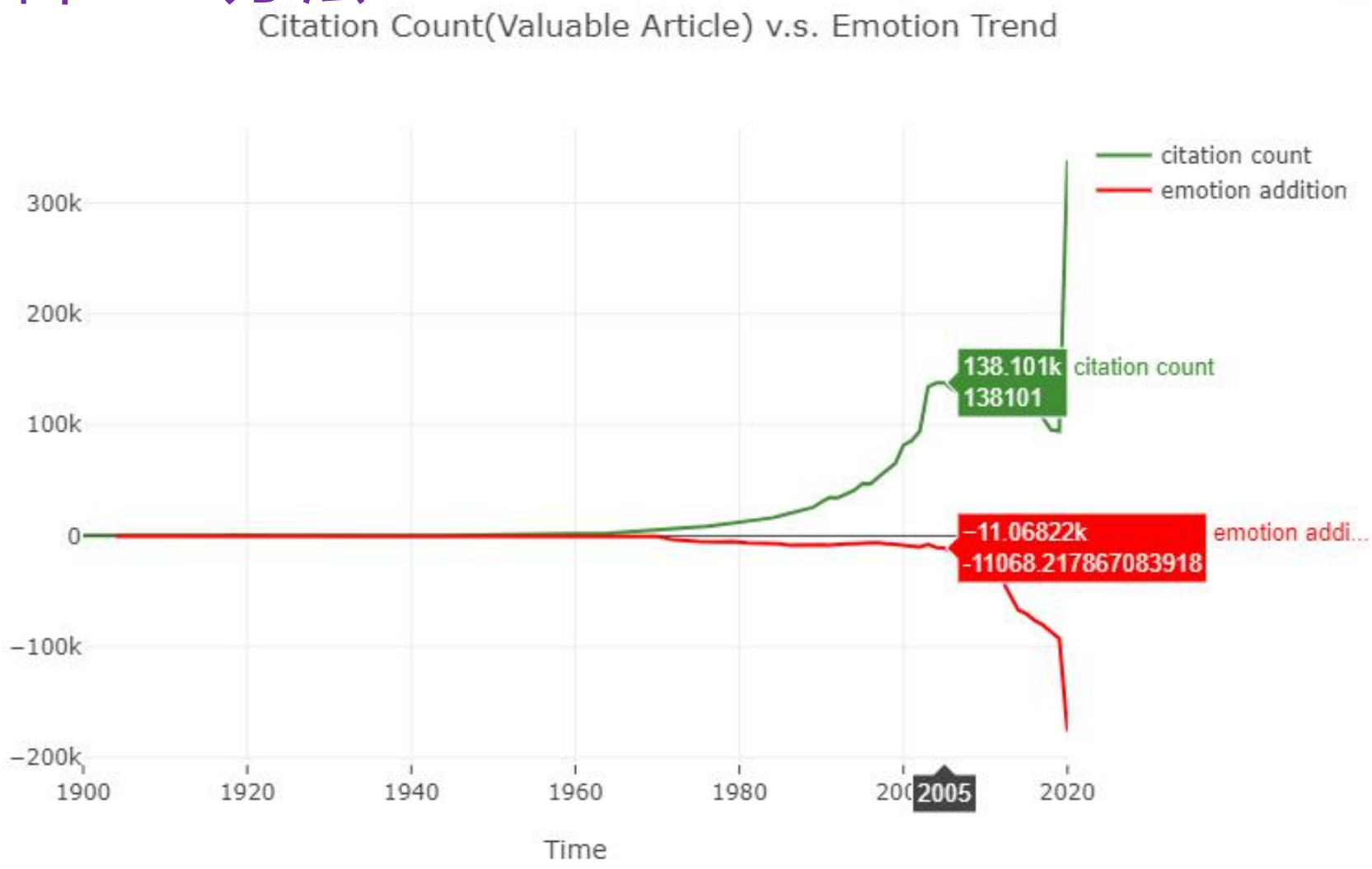
结果：


Citation Count(Valuable Article) v.s. Emotion Trend
citation count
emotion addition
300k
200k
100k
0
−100k
−200k
1900
1920
1940
1960
1980
2000
2020
Time

2 细节 & 方法

结果：

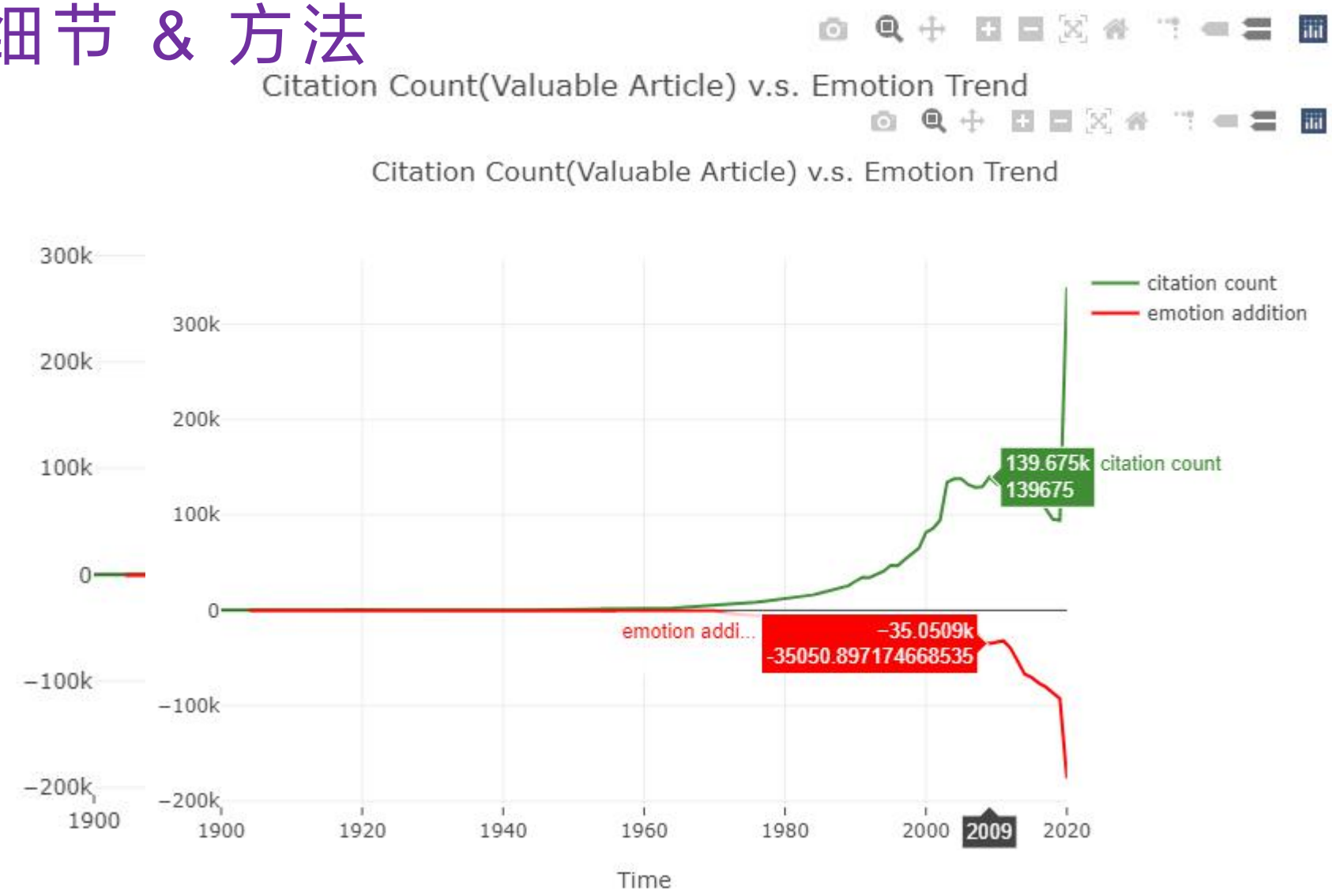
1. 2005
2. 2009
3. 2013



2 细节 & 方法

结果：

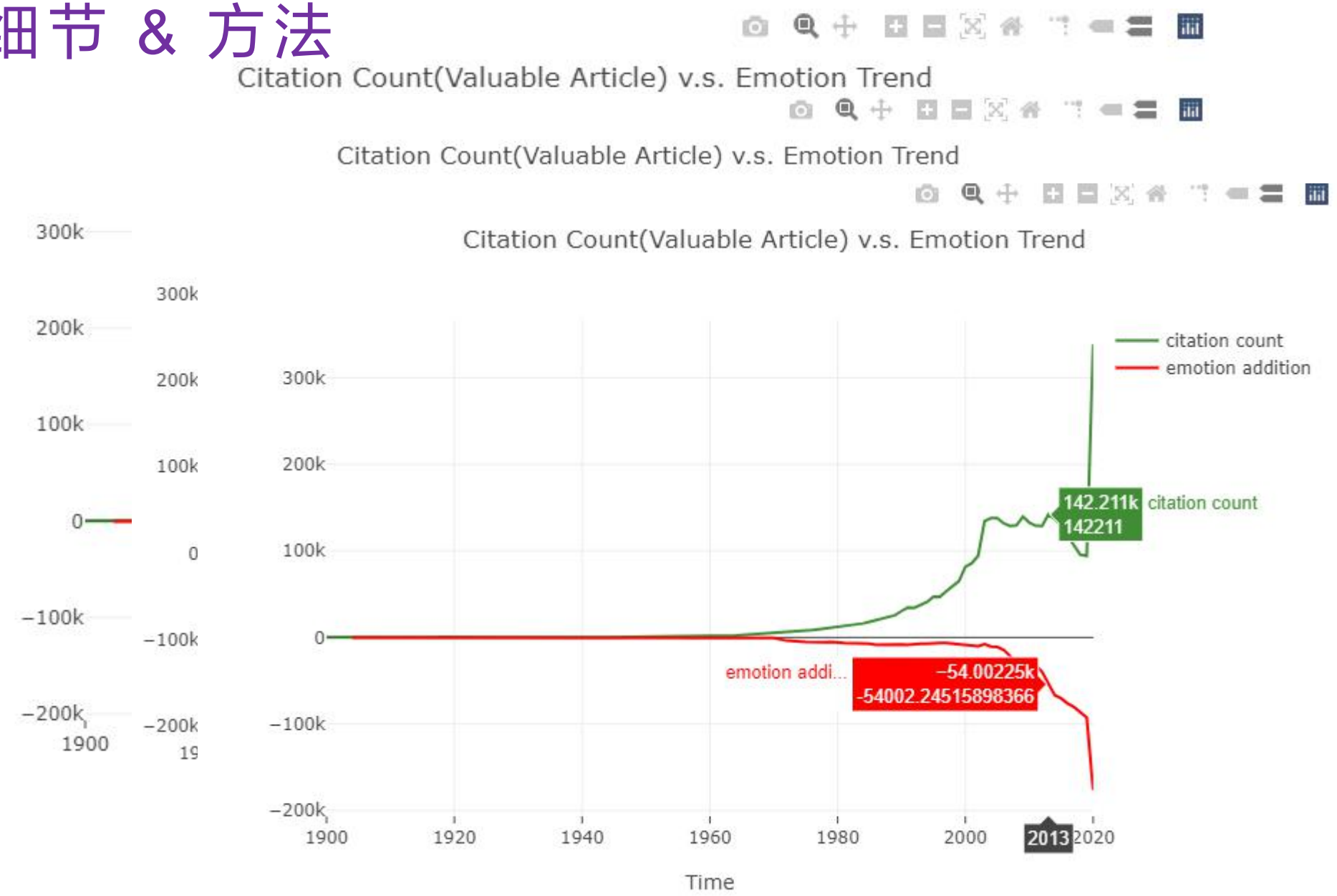
1. 2005
2. 2009
3. 2013



2 细节 & 方法

结果：

1. 2005
2. 2009
3. 2013



2 细节 & 方法

4. 我们似乎得出了一个科研的前进速度与科研人员的悲观情绪成正相关的结果

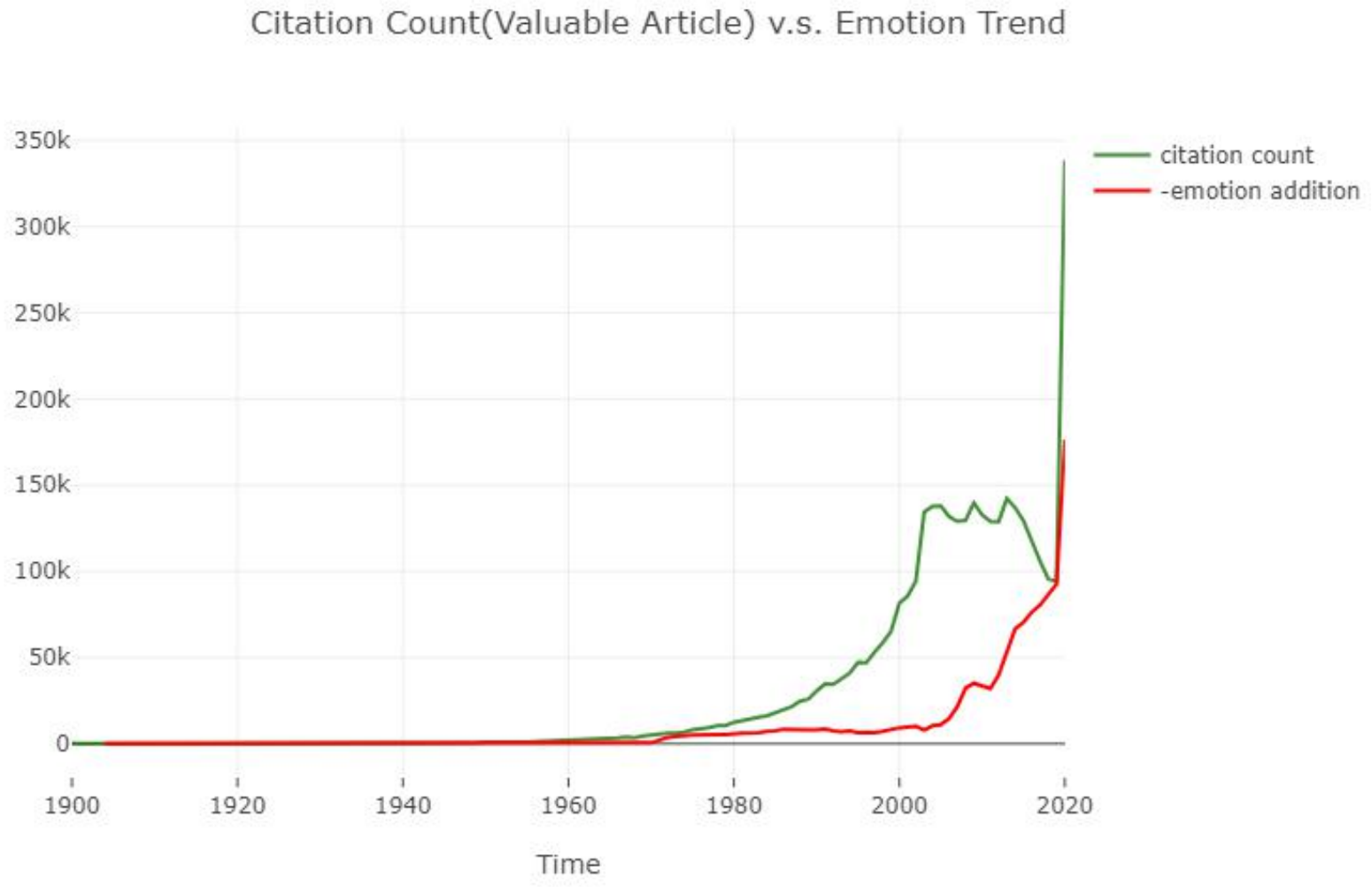
   即“科研人员大多心（xin）怀（qing）怜（bu）悯（hao），

   这里特指生科”

   实际上，这里只能说明这两个现象有相关性，很可能是同源相关性。

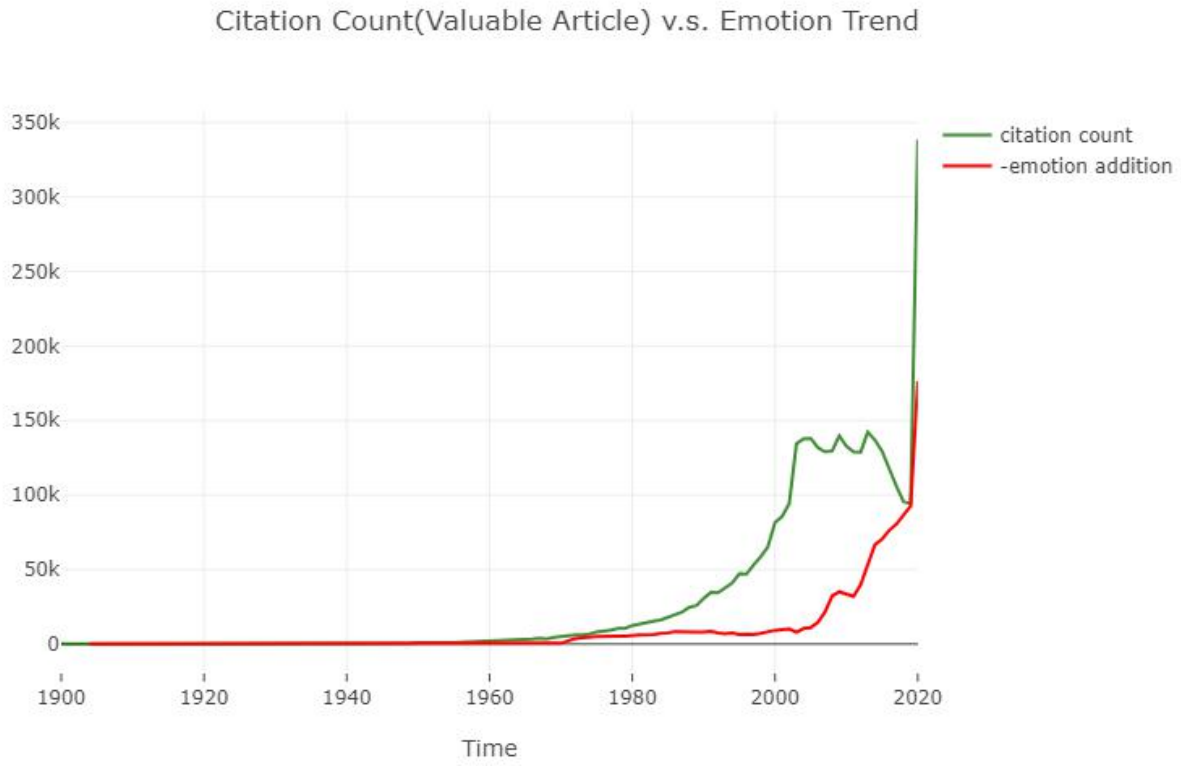
2 细节 & 方法

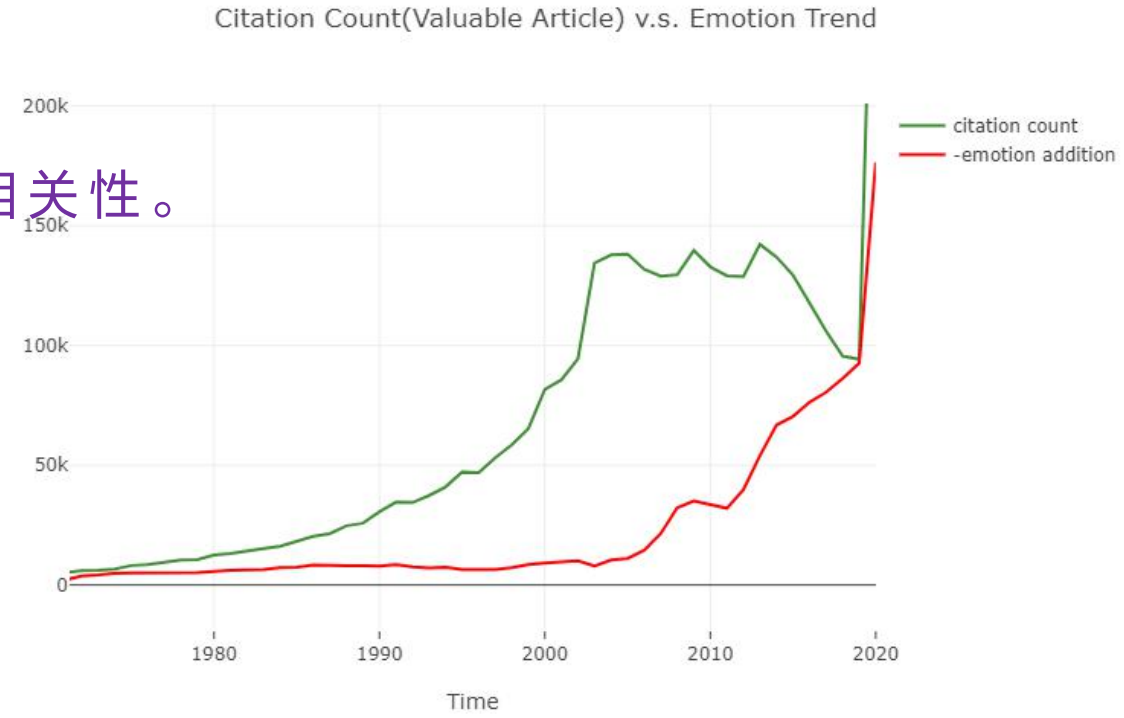
这里反转了一下，以更好的表现相关性。



2 细节 & 方法

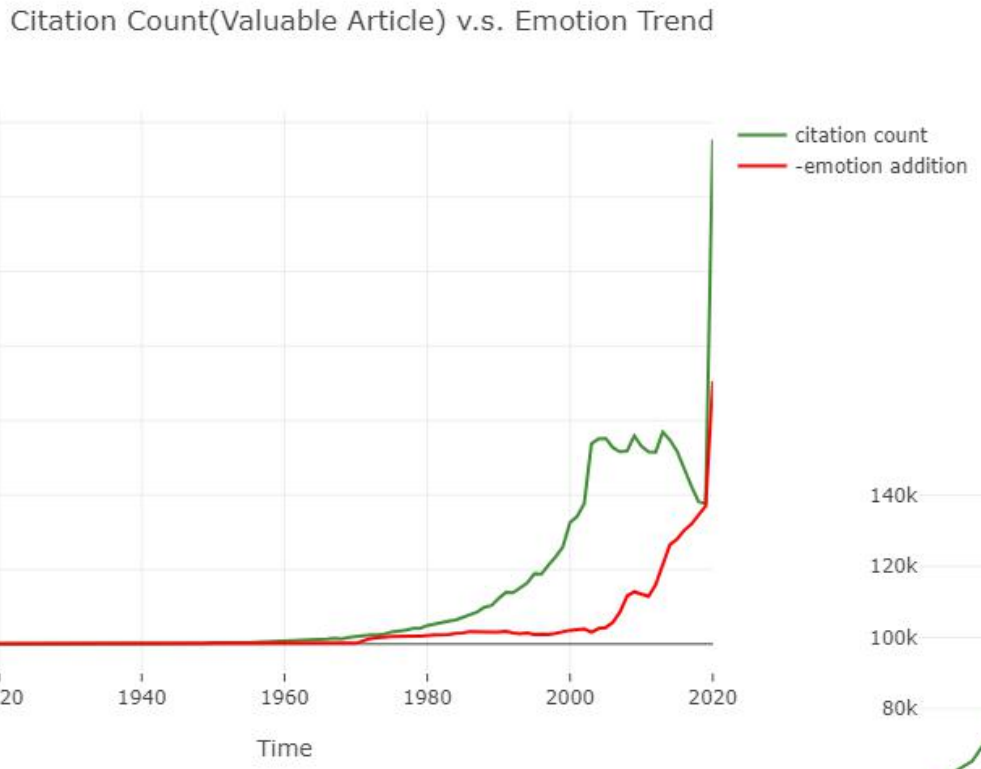
这里反转了一下，以更好的表现相关性。

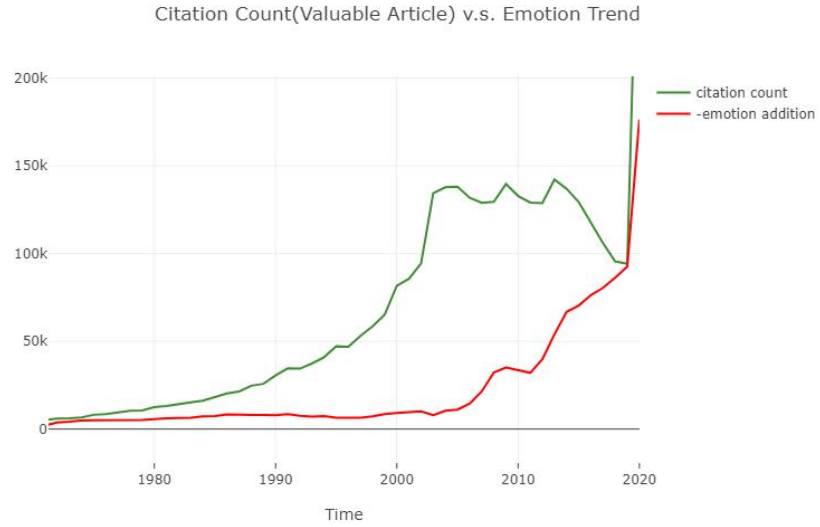


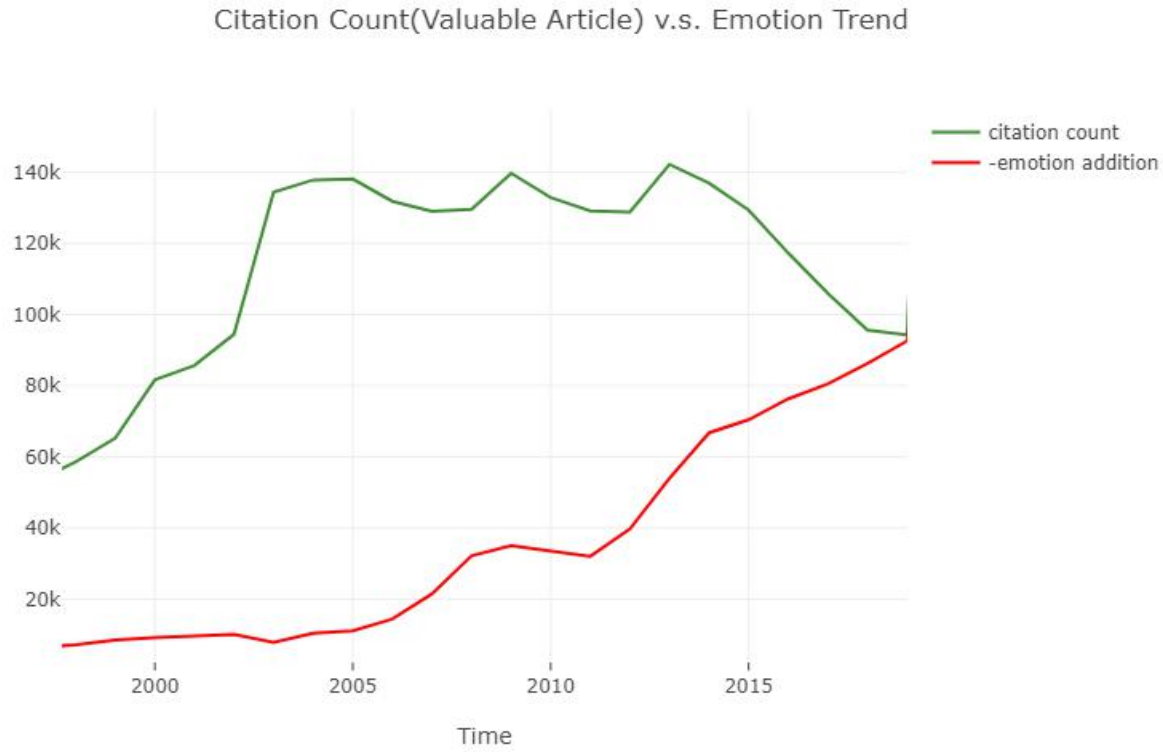


2 细节 & 方法

这里反转了一下，以更好的表现相关性。



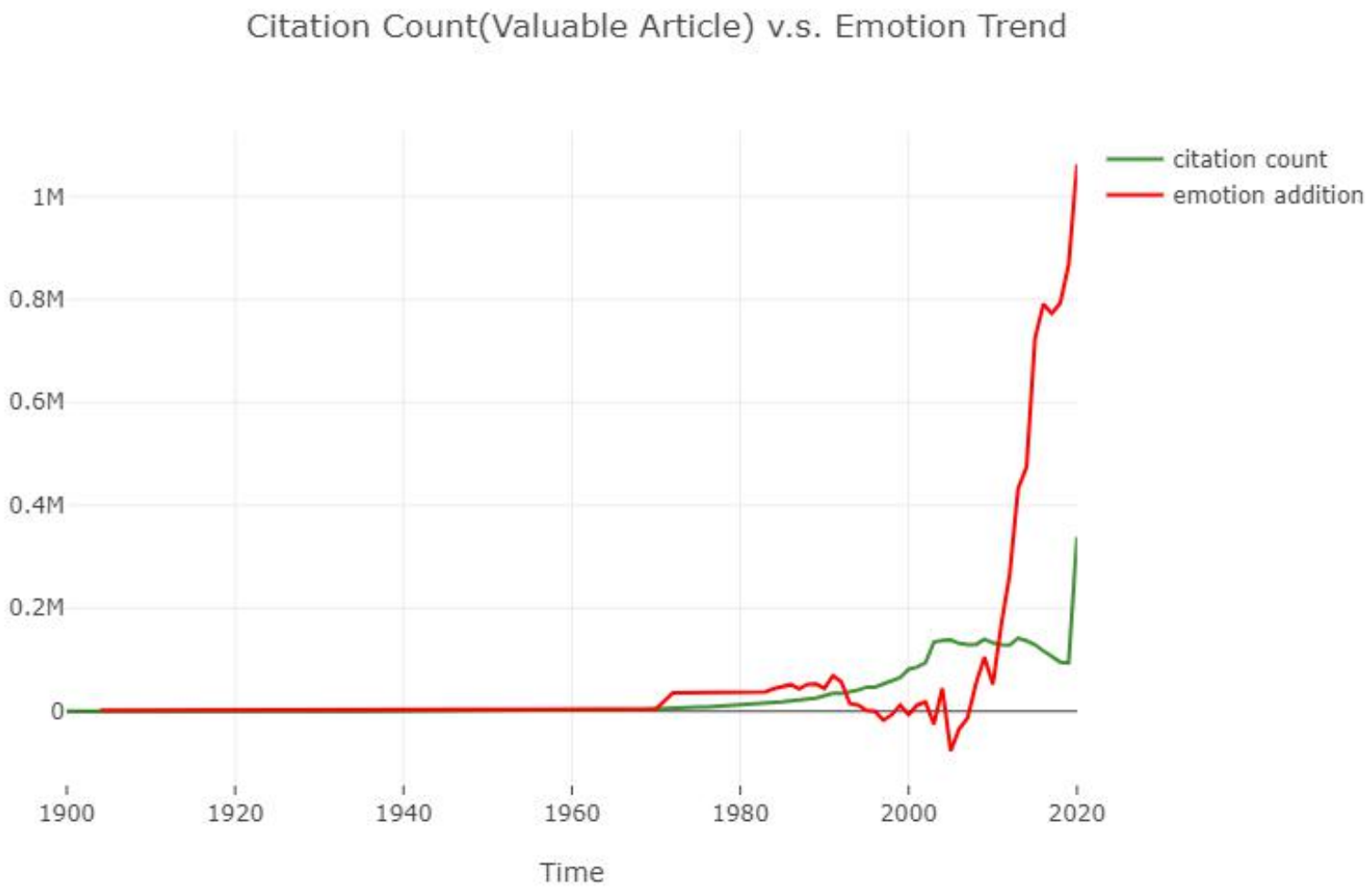




2 细节 & 方法

5. 按照（重要性*情绪）对时间进行排布（之前的结论里认为这两点是有同源相关性的，所以认为可以看作线性相关相乘，没有太大意义）

2 细节 & 方法



结果和我们所设想的并不一样，这里没有反转，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。

3 结论

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；生化都是伤心人

4 待完成的工作

1. 更详细的检索；
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