目录

一、背景及意义

随着大数据时代的到来, 数据从简单的处理对象开始转变为一种基础性资源。而随着互联网的不断普及，越来越多网民选择利用社交网络展现其对于现实世界的关注，藉由网络表达自身意见、宣泄自身情绪，互联网平台逐渐成为现实事件发展的传感器，而网络舆情也在反映民意、折射现实中承担着越来越重要的作用。

在网络舆论场当中，有一些重要人物扮演着举足轻重的角色，我们称之为“意见领袖”。意见领袖的价值观和情绪，总会通过不同媒介的宣传作用，以及个别因素的带动下，以理性或非理性的形式爆发出来，对民众产生一定的影响。

因此，我们选取了美国总统特朗普这一现象级的政治人物，基于 Twitter 大数据，对特朗普本人、他的社交圈、粉丝、以及与#TRUMP 关键字相关的 Twitter 用户逐一进行了分析。

二、技术路线

我们基于推特 API 及其他渠道获取特朗普本人及其关注者、话题涉及者个人信息、关注信息、推文内容等多种格式数据。在此基础上，经过数据清洗和预处理，将获取到的数据分别用于以下三个方面：

(1) 特朗普推文分析：LDA 主题聚类、情绪分类等；
(2) 特朗普社交网络分析：基于 PageRank 算法的关键人物定位、群体画像等；
(3) 美国民众感情倾向分析：基于情感分类的州(县)级情感倾向分析、城市发言意愿分析等等。

1

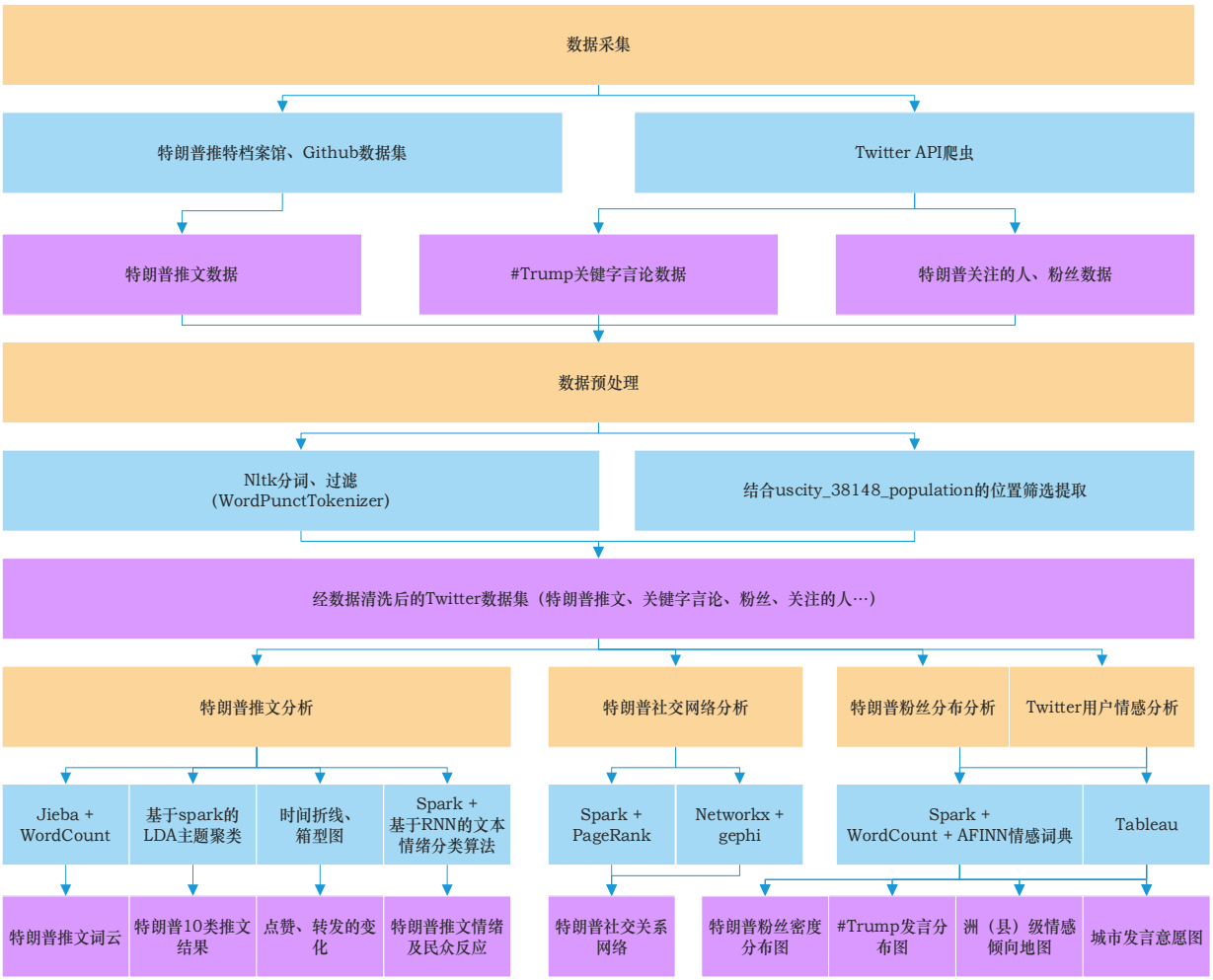


图 2： 技术路线图

三、数据采集

3.1 数据来源

1. 特朗普推特档案馆

http://www.trumptwitterarchive.com/

2009.5.4-2020.11.6 川宝所有的推文及相关信息共 55,090 条

包括推文内容、是否为转发、是否已删除、设备来源、点赞、转发量、日期……

2. GitHub 上的数据集

包括 2018-02-22 至 2018-03-03 共 10 天

226,770 条 Twitter 用户#Trump 言论数据（经数据清洗后）

https://github.com/sxl5507/Twitter-Sentiment-Analysis/blob/master/tweets_processed.txt

3. 使用 Twitter API 编写爬虫程序进行爬取

爬取 45 万余特朗普关注者（followers）数据

爬取 2020-12-25 的 213,881 条 Twitter 用户言论数据（圣诞特辑!!）

4. 实验室学长学姐的大腿

获得特朗普言论数据集（2009-2020）4 万余条，与川普推特档案馆数据集互补使用

3.2 数据选用及预处理

有关特朗普推文，我们采集的 csv 格式的数据包括 id(推特 id), text(文本内容), isRetweet, isDeleted, device(设备), favorites(点赞数), retweets(转发数), 以及 date(日期)的信息。如下图所示：

id	text	isRetweet	isDeleted	device	favorites	retweets	date
13118921906800	Tonight, @FLOTUS and I tested posi	f	f	Twitter for iPhone	1869706	408866	2020-10-02 04:54:06
13122338079914	Going well, I think! Thank you to all.	f	f	Twitter for iPhone	1219870	139605	2020-10-03 03:31:34
13237658842682	WE ARE LOOKING REALLY GOOD	f	f	Twitter for iPhone	961091	109900	2020-11-03 23:15:55
13238640211671	I will be making a statement tonight. A	f	f	Twitter for iPhone	926998	133988	2020-11-04 05:45:53
13235346634539	https://t.co/85ySh1KYkh	f	f	Twitter for iPhone	898911	170287	2020-11-03 07:57:08
11573456925176	A$AP Rocky released from prison ar	f	f	Twitter for iPhone	821423	226235	2019-08-02 17:41:30
12671296442282	The United States of America will be	f	f	Twitter for iPhone	788277	210615	2020-05-31 16:23:43
12663540840361	CHINA!	f	f	Twitter for iPhone	764501	143066	2020-05-29 13:01:56
13121584003529	https://t.co/B4H105KVSs	f	f	Twitter for iPhone	755999	129745	2020-10-02 22:31:56

图 3.1： 特朗普推文数据格式

有关特朗普关注的人以及特朗普粉丝数据，我们申请了 Twitter 官方开发者账号，获取使用官方 API 的权限。

对于知名度较高的人，其关注列表会受到许多人的关注，甚至会被外界认为代表其思想倾向，因而他们的关注列表是精挑细选的，具有很高的研究价值。因此我们爬取特朗普及其关系密切的用户信息来构建关系网络：具体地，首先根据特朗普的 id 爬取他关注的人的 id（51 个），随后将这些人的 id 作为用户 id 再次进行关注列表爬取，获取下一层关注人，具体关系如图 1 所示。最终我们构建了四层关系网络，获取了上万人的关注信息及其地理位置。

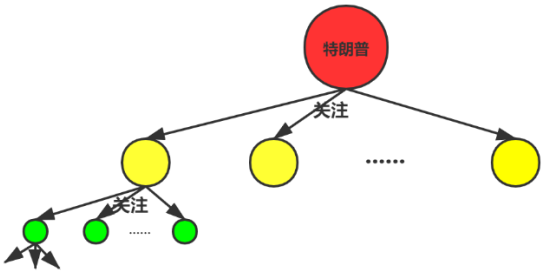


图 3.2： 特朗普相关人物爬取方式

有关 Twitter #Trump 相关言论数据，爬取的信息包括（用户 ID，推文内容，来源设备，地理位置，发言时间）。并使用 *uscity_38148_population.txt* 中美国的地理位置信息列表，提取可以确定实际位置的用户，标明其所在州、县、城市以及该城市对应人口。

有关具体推文内容的使用，因推文中含有大量的网址、图片地址、特殊符号、虚词等没有实际含义的信息。所以需要对数据进行清洗。步骤如下：

3

（1）利用自然语言处理包中的 **nltk** 包将特朗普推文中的文本用空格分隔成初步的单词表。
（2）其中还有一些特殊的符号干扰了文本分析，如：@(后面跟着用户名)、#(推特中用来标注线索的标签)……因此，将初步的单词表再做一次分词，用 **nltk** 中的 **WordPunctTokenizer**，这样可以拆分那些带有标点符号的单词，得到最终的词包。最后还需要过滤掉一些没有实际意义的词：如介词、特朗普的姓名等。最终导出一个词包。

四、数据可视化分析

4.1 特朗普推文分析

4.1.1 特朗普推文词云

数据清洗完成后，利用得到的词包，将词包里的单词进行可视化展示。

图 4.1： 2016 年之前特朗普所有推文的词云

4

图 4.2： 2016 年之后特朗普所有推文的词云

可以看出，不管是 16 年之前还是之后，特朗普说的最多的词就是 **great**。因为他身为一个伟大的人，经常在伟大的一天，干伟大的事。其次，16 年之后 **fake news** 突然出现在了很显著的位置，这很特朗普。

4.1.2 LDA 主题聚类分析

LDA 是基于三层贝叶斯概率模型的主题提取方法，包含了很多上下文的语义信息。LDA 对短文本的分析效果较差，而推特存在词限制恰好属于短文本。所以，需要先通过某种方法，将多个推文连接在一起形成长文本，然后将这些长文本作为 LDA 的输入。

（1）首先用 LSI(一种基于 SVD 的机器学习的方法)对推文进行初步的主题聚类。我们假设有 20 个主题，每条推文会 LSI 模型映射成一个 20 维的向量，作为每条推文的特征表示。

（2）其次，对所有推文进行 K-means 聚类，聚成 1000 个类。保存 1000 个类的聚类中心和所有推特到聚类中心的距离。根据距离远近，将推文分配给 1000 个中心。这样每个中心大概分配 40 条推文，连接起来形成长文本作为 LDA 模型的输入。

（3）利用 LDA 模型抽取 10 个主题，并且返回每个主题的关键词。其主要原理是将第二步中的 1000 个长文本映射成一个 10 维的向量，该向量所有维度中最大的值决定了该长文本是哪一类。

（4）最后利用 python 中的 pyLDAvis 包进行 LDA 聚类的可视化展示，主要原理是利用 PCA 将这些 10 维向量降至 2 维，得到一个可视化网页。由于特朗普于 16 年出任美国总统，在此之前的推文研究价值不高，于是将特朗普 16 年之后的推文进行聚类，对其在总统任职期间发的推文进行分析。

5

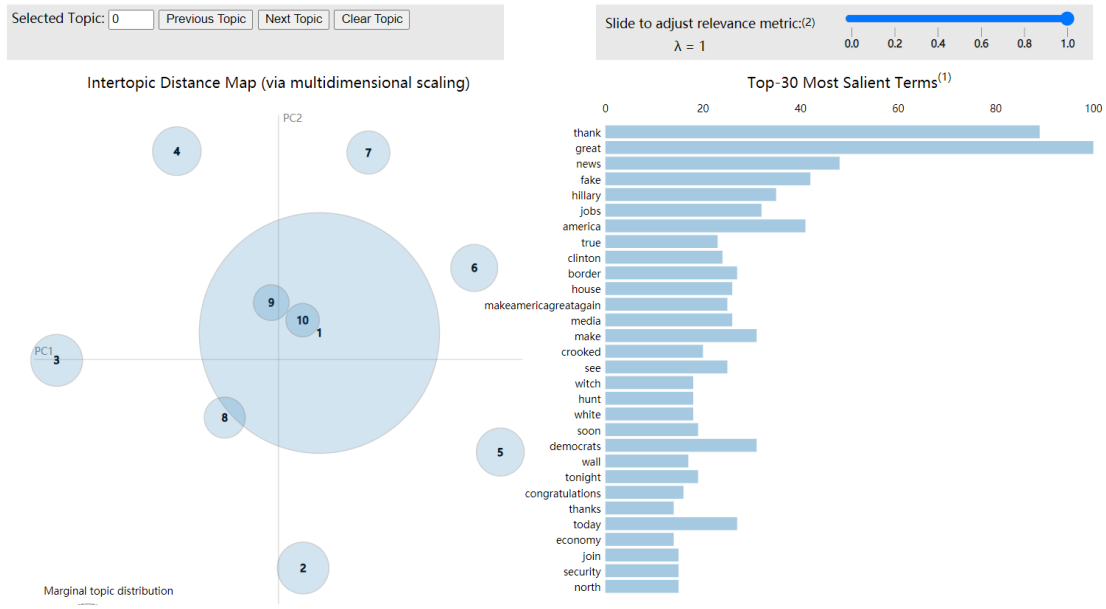


图 4.3：特朗普总统任期内推文聚类结果

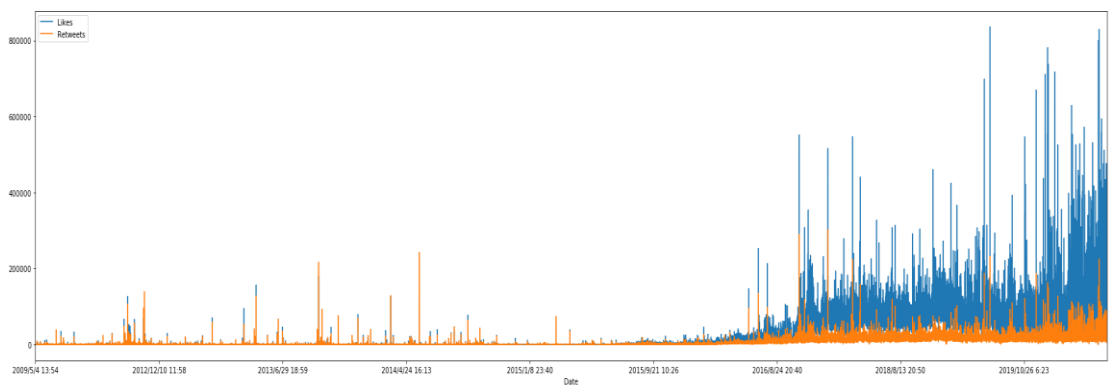
聚类得到的各个主题及分析如下：

类别	高频词汇	主题内容
1	Great	夸耀政绩相关
2	Make America Great Again	宣扬政治口号相关
3	fake news	抨击 CNN 等新闻媒体
4	Border wall Mexico	修建美墨边境墙
5	Thank support	感谢支持者
6	Witch hunt Russia	为“通俄门”辩解
7	Hillary crooked	借助“邮件门”抨击竞争对手：希拉里·克林顿
8	Jobs market economy	经济、市场和就业相关
9	Congratulations winning	恭喜他人赢得竞选

4.1.3 转发数与点赞数相关分析

1. **转发数-时间折线（橘色）与点赞数-时间折线（蓝色）**

图 4.4: 特朗普推文转发点赞数变化趋势图



论点 1：明显看出 2016 年前后是特朗普推文受关注度的分水岭，而且越往后其推文的受关注度越高。所以可以发现特朗普在当上总统之后发生了疯狂涨粉的现象。

2. 2016 年后特朗普推文的转发数与点赞数的箱型图

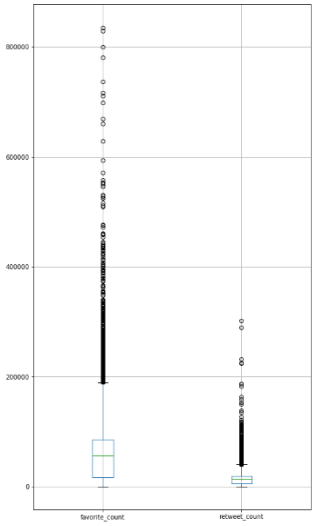


图 4.5： 特朗普推文转发点赞数箱型图

论点 2：图中可以很明显地看出来无论是转发数还是点赞数，都具有明显的大量的异常值，可见特朗普的推文的受关注程度并不稳定，据分析应该是与特定的热点事件相关联，针对热点事件的推文出现的时候就会出现点赞数和转发数的爆炸性增长。这也与图 3 折线图反映的信息相一致。

3. 找出受关注度爆炸性增长的几条推文并分析其主题

我们找出了受关注度最高的 10 条推文并分析了他们的主题。2009 年到 2020 年这 11 年之间受关注度最高的 10 条推文的主题分别是：

第 1 名、第 2 名	特朗普感染新冠
第 3、4、5、6 名	2020 大选相关
第 7 名	歌手洛基回国
第 8 名	反恐
第 9 名	CHINA！
第 10 名	特朗普感染新冠

论点 3：显然最近的新冠和大选是最受民众关注的热点事件，特朗普与之有关的推文都获得了不小的点赞数和转发数。但 CHINA!仅凭 6 个字符就能冲进前十，可见

7

外国人还是很吃特朗普把矛盾转嫁给中国这一套的。我们要放弃幻想，准备战斗。

4.1.4 特朗普推文情绪分析

1. Plutchik 情绪理论与情绪标签的生成

这个理论把文本中蕴含的情绪分为八个维度，另外每个维度都有各自的“程度”，不同的维度之间还有混合情绪，是一个在 NLP 领域相对比较权威的情绪识别模型[1]。在这里我们为了简化对推特文本情绪标签的添加，采用了 8 个情绪维度中程度最适中的 8 种情绪进行分类，分别是 Anger, Anticipation, Jor, Trust, Fear, Surprise, Sadness, Disgust。

我们使用下面这篇论文的方法，把推特文本放入一个 RNN 模型中进行识别，最终得到带有 8 种情绪标签的推特的数据。

Emotion Recognition on Twitter: Comparative Study and Training a Unison Model

Publisher: IEEE　Cite This　PDF

id	link	content	date	reweets	favorites	mentions	hashtags	emotion
1	https://tv	Be sure tc	04/05/2009	510	917			Surprise
2	https://tv	Donald Tru	04/05/2009	34	267			Surprise
3	https://tv	Donald Tru	08/05/2009	13	19			Trust
4	https://tv	New Blog F	08/05/2009	11	26			Joy
5	https://tv	"My perso	12/05/2009	1375	1945			Trust
6	https://tv	Miss USA 1	12/05/2009	29	28			Trust
7	https://tv	Listen to	13/05/2009	15	16			Anticipation
8	https://tv	"Strive fc	14/05/2009	18	27			Joy
9	https://tv	Enter the	15/05/2009	15	9			Surprise
10	https://tv	"When the	16/05/2009	19	47			Trust
11	https://tv	"Don't be	17/05/2009	40	69			Fear
12	https://tv	"We win ir	18/05/2009	74	122			Joy
13	https://tv	"...these	19/05/2009	23	23			Trust
14	https://tv	"Always kr	20/05/2009	28	51			Surprise
15	https://tv	Read a gre	20/05/2009	10	11			Anticipation
16	https://tv	"Keep it f	21/05/2009	61	82			Trust
17	https://tv	Don't forg	22/05/2009	13	5			Joy
18	https://tv	Did you kr	23/05/2009	58	79			Surprise
19	https://tv	"Your high	26/05/2009	50	81			Trust
20	https://tv	Read an ex	27/05/2009	6	12			Surprise
21	https://tv	"You have	28/05/2009	63	59			Trust
22	https://tv	Check out	03/06/2009	15	20			Trust
23	https://tv	"If you c	05/06/2009	13	31			Trust
24	https://tv	Last week	08/06/2009	10	3			Surprise
25	https://tv	Today is I	14/06/2009	26	32			Surprise
26	https://tv	Thanks to	15/06/2009	37	38			Surprise
27	https://tv	RE: FB Var	18/06/2009	19	15			Joy
28	https://tv	- Wishing	21/06/2009	61	123			Joy

图 4.6： 推文情绪分析结果

2. 针对特朗普推文情绪的各种分析

得到情绪标签后，首先我们对特朗普的推文的情绪比例进行分析。得到如下的饼状图。

8

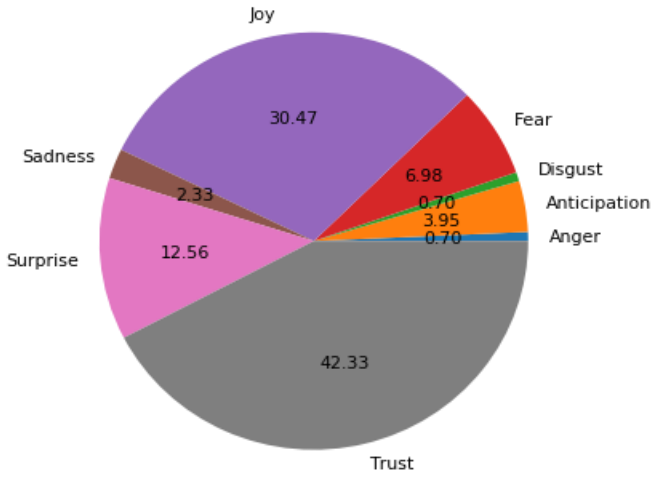


图 4.7： 特朗普推文情绪分布

论点 4：从中可以看出就比例而言，特朗普最喜欢发的情绪依次是：Trust、Joy、Surprise、Fear、Anticipation、Sadness、Anger 和 Disgust。可见特朗普还是个蛮正能量的人，他不喜欢发负能量的推文。

但是用户喜欢特朗普推文带有什么样的情绪呢？于是我们对 8 种情绪的特朗普推文的平均点赞数和平均转发数做了详细分析。得到了如下的条状图。

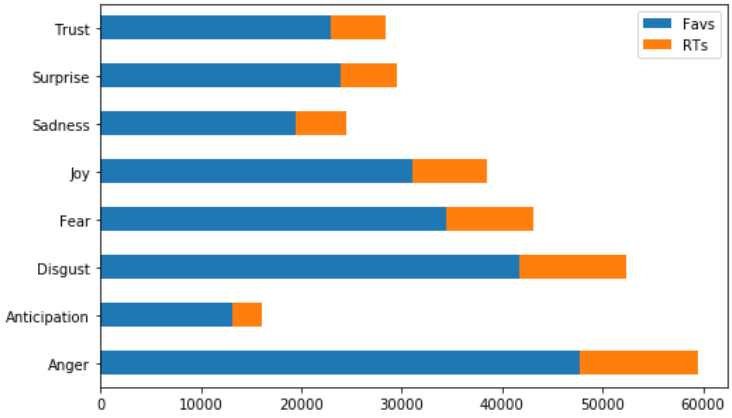


图 4.8： 特朗普不同情绪推文点赞转发数条形图

论点 5：虽然特朗普不喜欢发负能量的推文，但是 Anger、Disgust 与 Fear 等负面情绪对应的点赞数和转发数却明显地高于其余的情绪。表面上看起来是用户们对于情绪化严重的推特更喜闻乐见（符合勒庞的大众心理学理论），但是也不排除这样的原因：能够引起特朗普发情绪化严重的推特对应的社会事件往往同时也同时能够吸引更多的用户关注，所以相应的推特得到了更高的点赞数和转发数。

4.1.5 特朗普发推设备的变迁史

因为收集的数据中有 device 的相关信息，就对特朗普的发推设备做了一些分析。下

9

图可以看到特朗普发推设备的使用比例和随时间的变迁。

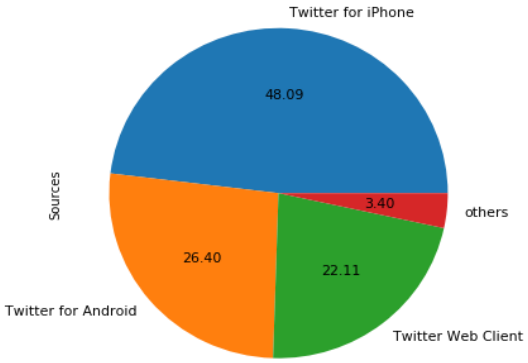


图 4.10： 特朗普使用设备占比饼状图

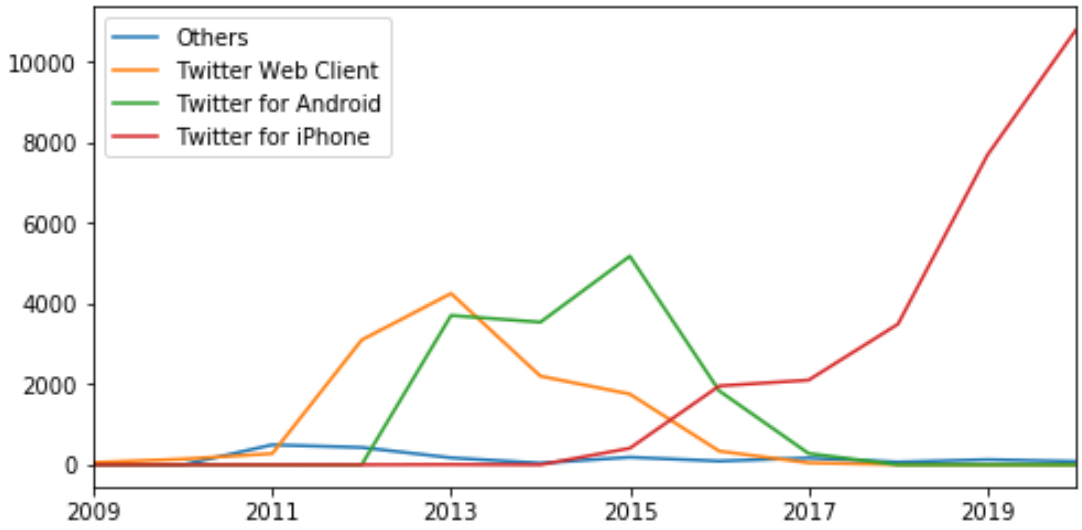


图 4.11： 特朗普设备使用折线图

论点 6：2012 之前特朗普基本没有使用智能手机，一般使用网页端发推。2012 年买了一部 Android 操作系统的手机，然后对网页端的使用率下降（但一直还是用的）。再后来 2015 年他又买了一部 iPhone，于是 2015 和 2016 年他使用三种设备混合发推。有意思的是 2017 年往后（不知道是不是买了一部更好的 iPhone）特朗普迅速抛弃了旧爱网页端和 Android 平台，全力在 iPhone 上发推输出。

4.2 特朗普的朋友圈——社交网络分析

10

4.2.1 节点、权重与网络构建

为了构建关系网络，需要确定哪些人要作为网络的结点，并确定他们的权重来进行聚类。

选取网络结点方面，我们首先确定了特朗普和其关注的 51 个人都在网络之中，随后在第三层和第四层网络中，按照“粉丝数”，“被第一层的 51 个人关注的人数”两个指标进行筛选，最终选取了另外 69 个人来构成一个拥有 121 个结点的网络

确定权重方面，我们采取了推荐算法，并根据推荐人的等级不同（例如特朗普是第 1 级，51 个人是第 2 级等），被推荐人数，粉丝数等因素确定了各结点的权重。

采用 python 的 networkx 库对图模型进行构建，随后输出结点和边文件，并送入 gephi 软件进行可视化，获得网络图以后对所有人进行身份调查，并采用个人照片作为结点外观，最终获得完整网络关系图，如图 4.12：

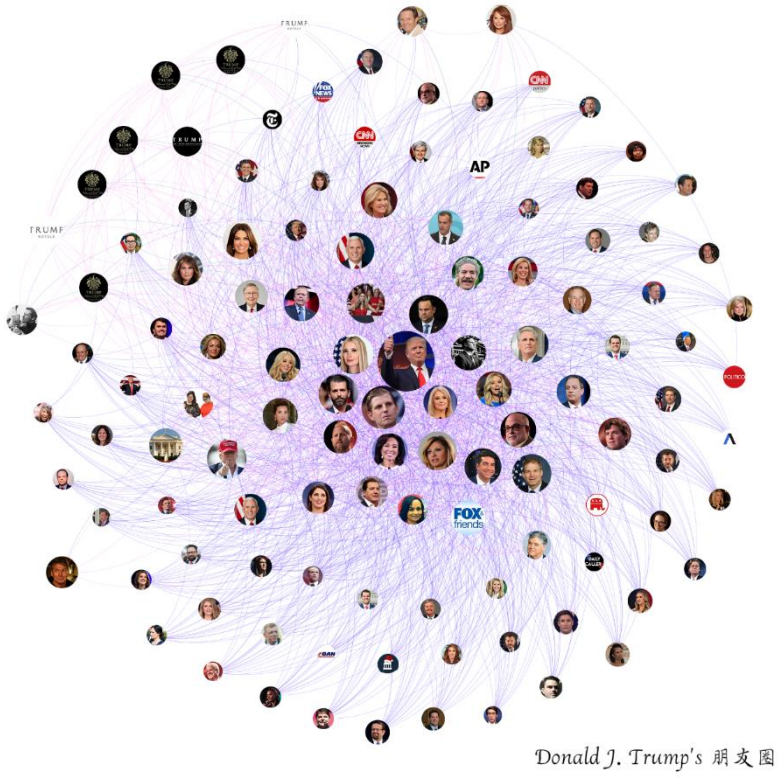


图 4.12： 特朗普社交网络关系图

4.2.2 人物属性分类

通过对所有人的身份调查，可基本将整个关系网络分为如图 4.13 的几个关系

11

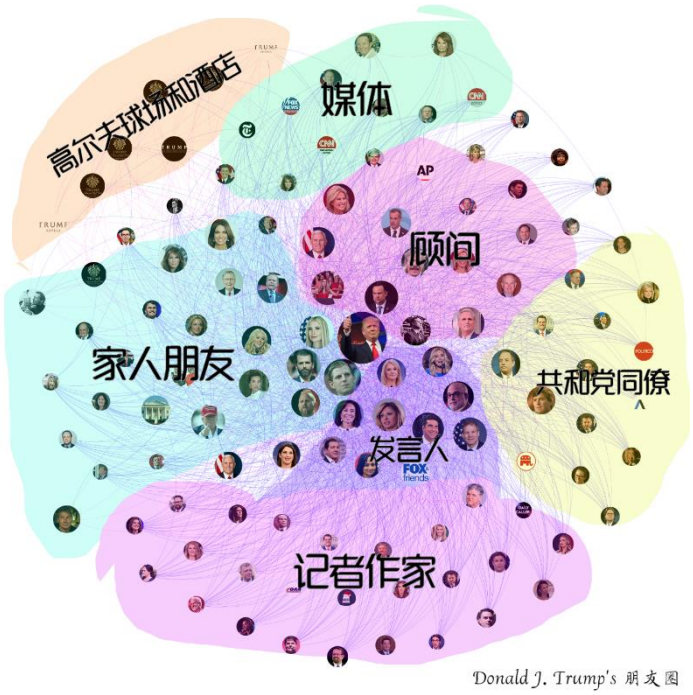


图 4.13： 特朗普社交网络关系分类

4.2.3 意见领袖与人物派别

将各结点信息输入 PageRank 算法获取每个人的权重，最终获得前四名分别为 Eric Trump（二儿子）、Donald Trump Jr.（大儿子）、Brad Parscale（数字总监兼政治顾问，挚友）、Ivanka Trump（大女儿），这四个人在关系网络中实际也离特朗普最近，因而结果和实际情况基本是吻合的。

基本上所有通过筛选进入这张图的人都与共和党有密切的联系，因此可以通过这张关系网络图来判断出哪些人是亲特朗普派，尤其是记者作家，他们的作品以及言论将有极大的偏向性。在关系网络中，FOX 新闻的权重相对于其他如 CNN，纽约时报等媒体的权重要高很多，并且几乎所有的记者都与 FOX 有关系，因此同样可以判断出各大媒体对特朗普的态度。在身份调查过程中，我们发现某些记者或主播表示自己是中立派，但其大量具有偏向性的报道以及在这张图中的位置已经暴露出他们是亲特朗普派的事实，因此社交关系网络同样可以分析出一些隐藏的信息与偏好

12

4.2.4 人物关系

在关系图的核心位置出现了一位女士 Katrina Campins，如图 4，她既不是一名政府工作人员，也不是媒体人员，而仅仅是一个出售豪宅的商人。但她可以处于核心位置，我们分析其主要原因是参加了特朗普举办的一档以挖掘经商天才为目的的综艺节目《学徒》，并取得了不小的成绩。同时这档节目的导演及编剧也被特朗普直接关注，但由于其被较少的亲信所关注所以权重较低（关系图中最上面的两个人，并且为夫妻关系）。这可以看出特朗普对这档节目的重视程度。

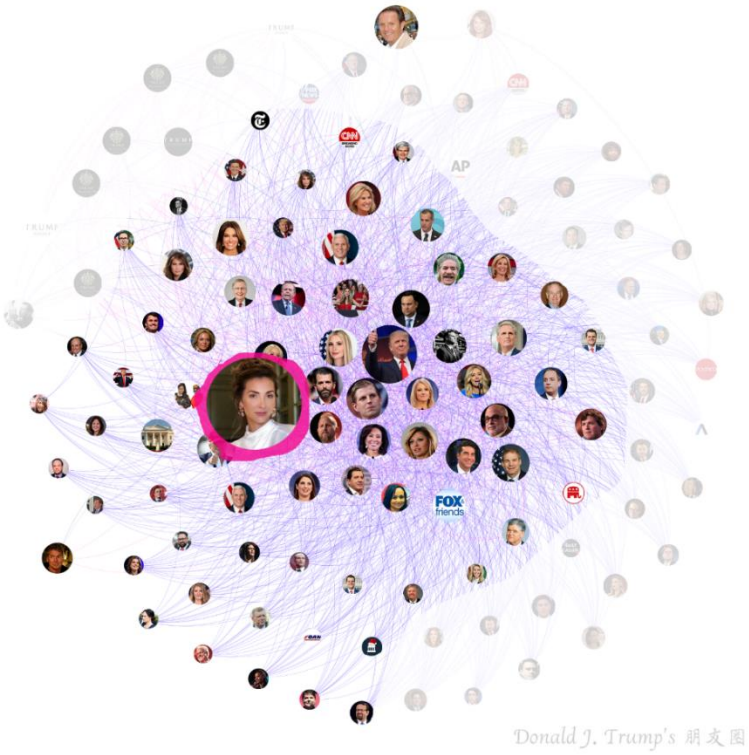


图 4.14： Katrina Campins 与关注她的人

另外，我们发现特朗普的妻子 Melania Trump 与特朗普的距离相较于他与其他儿女的距离要远很多，这或许也可以作为分析他们夫妻关系的一个线索。

同时特朗普的商业账号（高尔夫、酒店等）也被划分到合适的位置，这些账号只被特朗普家族的人所推荐，因而权重较低，处于边缘位置。

13

4.3 特朗普粉丝分布及 Twitter 用户相关发言分析

4.3.1 Twitter 特朗普粉丝密度分布

爬取 45 万余特朗普关注者（followers）数据，其中标明地理位置的有 112,796 人（约 11 万），结合美国的城市、郡县、州的具体信息，使用 spark 根据关注者地理位置对数据进行分类汇总，使用 tableau 绘制出特朗普的关注者的密度分布情况（美国）。

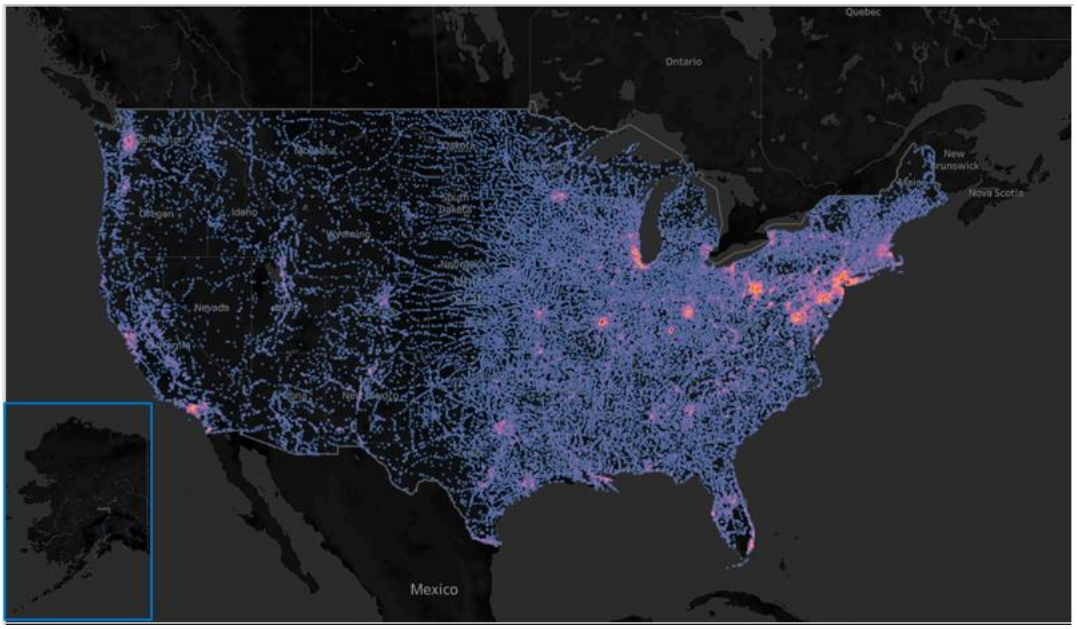


图 4.15：Twitter 特朗普关注者密度分布图（美国）

通过与 2019 年美国各州的 GDP 贡献率的对比，可以发现，特朗普的关注者数量最多的前四个州（加州、德州、佛罗里达、纽约州）的 GDP 贡献率就是全美前 4。

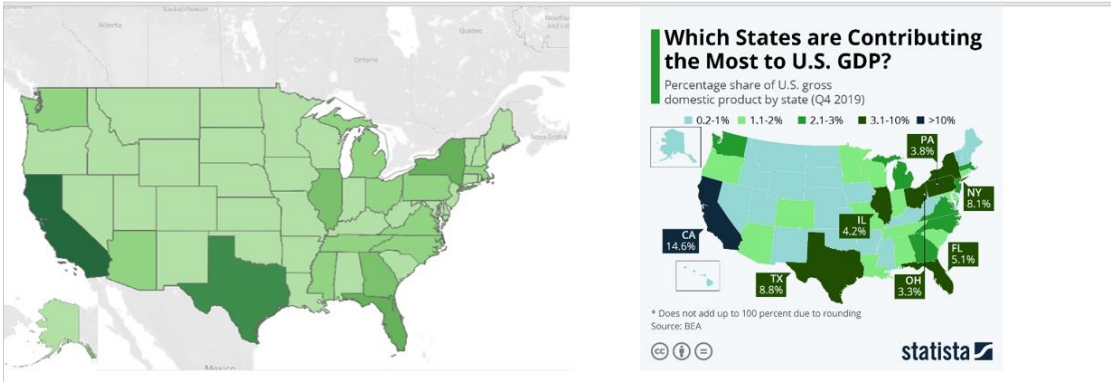


图 4.16：Twitter 特朗普关注者密度-全美 GDP 分布图

通过与美国人口分布的对比，特朗普关注者数量最多的前四个州的人口也是最多的，这也符合基本常理。因此可以初步得出的结论是，特朗普总统 Twitter 关注者的分布基本正比于美国的人口分布，但关注者数量排名第一的加州是第二名德州的 1.44 倍，两州的人口比

14

例是 1.38：1。我们推测这可能与加州的产业特性、移民数量相关：加州拥有全美最高的移民数量占比（27%），加州华裔人口占全美总华裔人口的 36.9%。

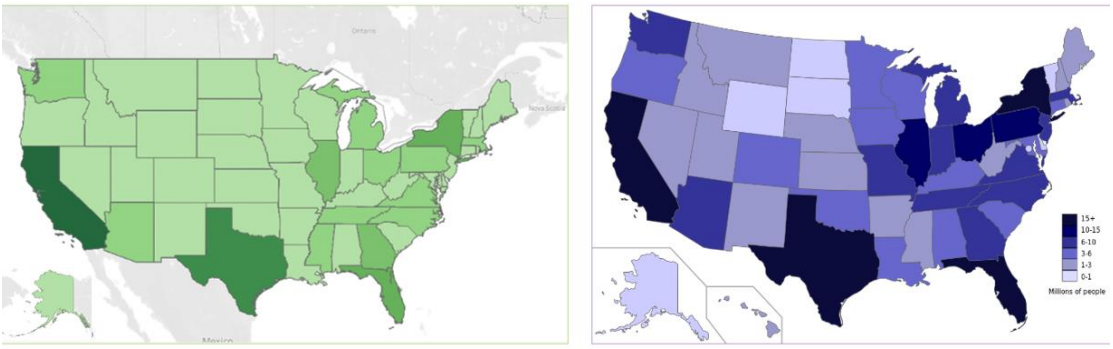


图 4.17：Twitter 特朗普关注者密度-全美人口分布图

4.3.2 美国 Twitter 用户对于特朗普的情感倾向

使用关键字#Trump 爬取 Twitter 用户的推文内容，以用户言论内容为基础分析美国民众对于特朗普的情感倾向。

同样使用美国的城市、郡县、州的具体信息，对应相应的城市人口数和经纬度坐标信息，绘制出发表#Trump 相关言论的 Twitter 用户的密度分布情况，由图可见，2018 年与 2020 年美国各区域网民对于#Trump 的发言频率基本持平。

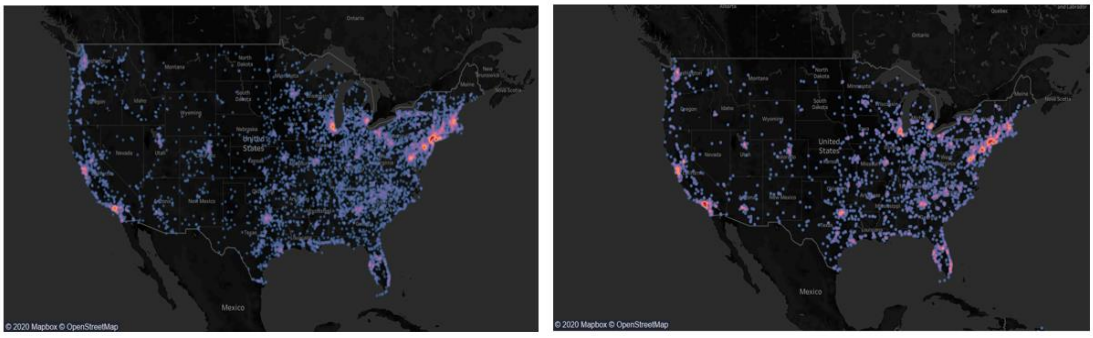


图 4.18：Twitter 特朗普关注者密度变化对比

使用 AFINN 情感词典结合 Spark 对所有经过数据预处理的用户言论进行情感极性评分，评分介于-5 和+5 之间。这些词语由 Finn Årup Nielsen 在 2009-2011 年手动标记。并将情感得分以 county（县）为单位计算总和，在地图上标出：红色为正、蓝色为负、颜色越深表示情感越强烈。

2018 年 2 月 14 日，美国佛州一所高中发生枪击案。我们推测这种恶性事件对于人们在#Trump 发表相关评论时的情感表达具有一定影响。

而在 2020-12-25 圣诞节这天，因为 Twitter 用户的发言包含了大量 merry, blessing, celebrating, thankful, great, love, enjoy 这样的词汇，所以整体的情感倾向略显积极。

15

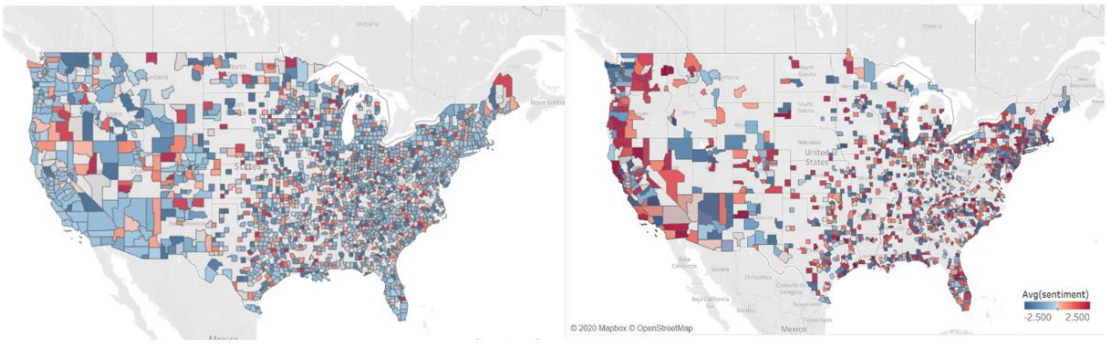


图1：Twitter用户#Trump发言情感倾向地图（2018.2.22-3.2）　图2：Twitter用户#Trump发言情感倾向地图（2020.12.25）

图 4.19：推特用户发言情感倾向地图

4.3.3 情感倾向地图 & 2020 年美国大选结果对比分析

将美国 2020 年大选结果与 2020.12.25 的州级情感倾向地图进行对比，可以发现情感倾向基本与大选结果布局相同，存在的差异主要还是因为 API 条件所限，爬取数据量不足。

然后经过仔细对比，还可以发现一些比较有趣的事情：比如 2020 年的关键摇摆州内华达州，该州在最后时刻翻，且选票被发现存在严重的作弊行为，有 60 万张选票无法做选民签名匹配，还有未成年人票、死者票、和非本州人选票等违规票种。但是在 25 日的统计中，平均情感值为+2.00。 设想如果数据量大到一定程度的话，可以根据情感倾向地图去初步鉴别可能存在选票造假的州。

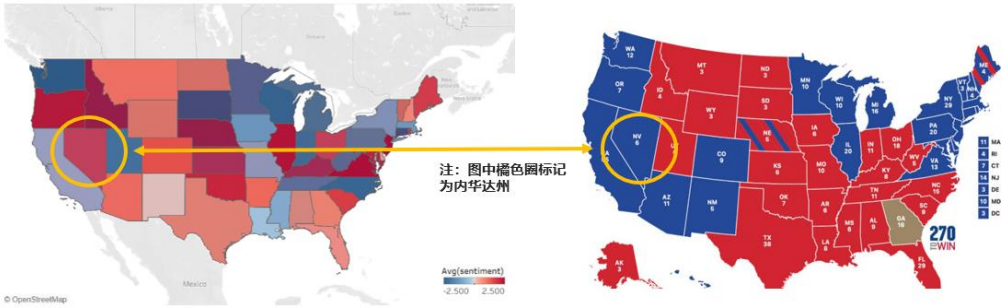


图1：Twitter用户#Trump发言情感倾向州级地图（2020.12.25）　图2：2020年美国大选结果

图 4.20：Twitter 情感倾向-美国大选结果对比

4.3.4 美国各城市 Twitter 用户发言意愿分析

以城市为单位，计算该城市发言记录数/城市人口总数，在图中以圆圈大小表示，可以通过这种方式衡量哪些城市对于#Trump 相关的内容更有发言意愿和其对应的感情倾向，图中的圆圈越大表明该城市发言意愿越强烈。经过分析证实，美国网民对于#Trump 相关的负面内容有较大的发言意愿。

16

同时，这种方法使得一些对于#Trump 发言异常积极的用户显得格外突出：比如缅因州的 Palmyra，密苏里州的 Stoutsville…

图 4.21：全美发言意愿地图及异常用户举例

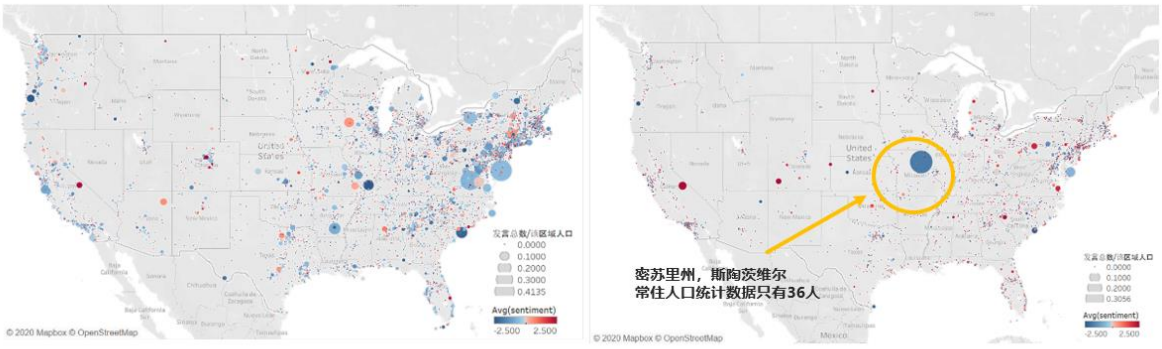
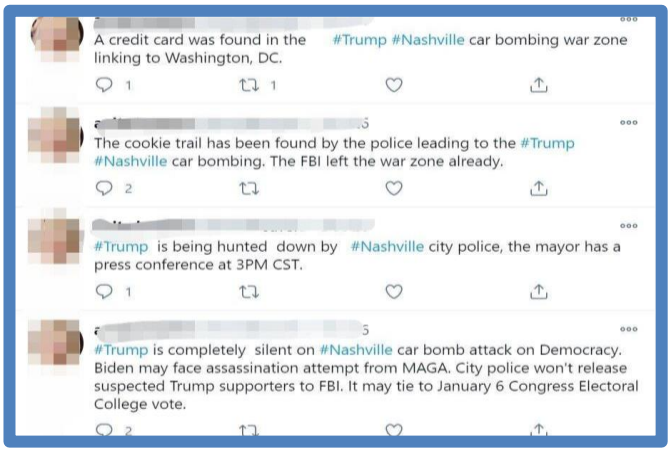


图1：各城市Twitter用户#Trump发言意愿统计图（2018.2.22-3.02）　图2：各城市Twitter用户#Trump发言意愿统计图（2020.12.25）



上图：Stoutsville 的某位用户在 12 月 25 日美国纳什维尔汽车爆炸案发生之后疯狂发推责备特朗普政府。这种方法经过进一步改造后，或可以用于基于舆论大数据的舆论实时监督，可以及时发现带节奏的异常用户群体。

五、大数据分析算法

5.1 Wordcount

我们设计了基于 spark 的 wordcount 算法用于分析推特中的词频统计，以作为后续分析的初步试探。

基于 spark 的 wordcount 算法流程图如下：

核心代码为：

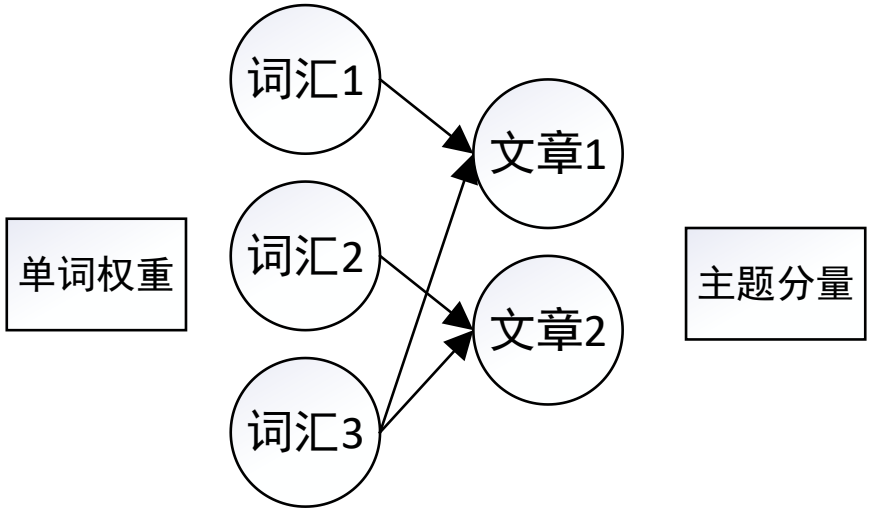
17

```
inputdata.flatMap(lambda x: x.split(" ")).map(lambda x: (x, 1)).reduceByKey(lambda a, b: a + b)
```

得到词频统计后，在此之上可按照词汇的正负性等进行进一步的细粒度划分。

5.2 LDA 主题聚类

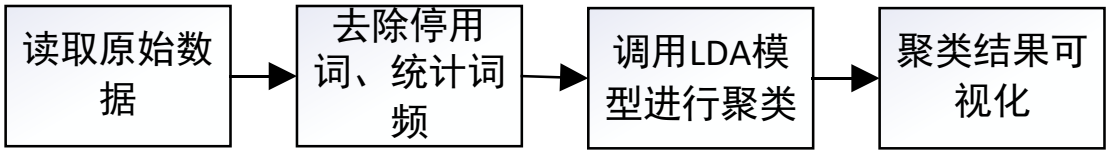
我们设计了基于 spark 实现的 LDA 算法，LDA 是一种经典的主题模型，常用于文本分类任务。该模型是包含词、主题、文档在内的三层贝叶斯概率模型，该模型使用词袋模型，将每篇文档视为一个词频向量，从而得到文档到主题的多项式分布和主题到词的多项式分布。



Spark 中的 Mlib 库中已经包含了 LDA 模型，可以直接进行使用。

对于我们的任务而言，我们希望对特朗普的所有推特内容进行聚类，尝试对其推文的主题进行分析。因此，我们借助 spark 中的 lda 模型实现了推特文本主题聚类算法。

具体地，首先读取存储在 csv 中的特朗普推特内容数据，接着，对读取的数据进行分词、去除停用词等常规预处理操作，生成 RDD 后调用 LDA 模型进行无监督的聚类，流程图如下：



核心代码如下：

```
text_data = spark.read.format("csv").option("header", True).load("./data.csv")

# 分词
tokenizer = ft.RegexTokenizer(inputCol='documents', outputCol='input_arr', pattern=r'\s+|[,.\"]')

# 去停用词
stopwords = ft.StopWordsRemover(inputCol=tokenizer.getOutputCol(), outputCol='input_stop')

# 统计词频
stringIndex = ft.CountVectorizer(inputCol=stopwords.getOutputCol(), outputCol='input_indexed')

# 选择模型
clustering = clus.LDA(k=10, optimizer='online', featuresCol=stringIndex.getOutputCol())
```


18

5.3 PageRank 算法

PageRank 算法用于分析特朗普推特社交圈的网络结构。
具体地，算法流程为：首先，根据网络结构初始化邻接矩阵，其次，按照如下矩阵运算公式进行迭代：

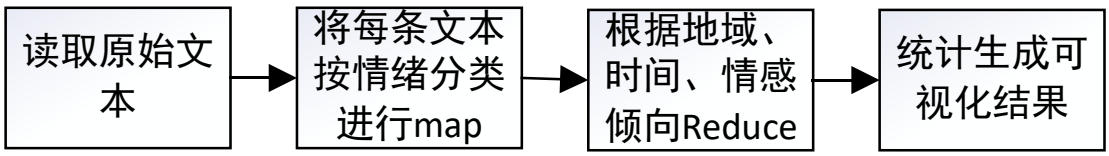
$$\vec{r} := (qA + \frac{1-q}{N} I)\vec{r}$$

其中，q 取 0.85，N 为网络节点数，为 121，I 为全 1 矩阵。
最后，特征向量 r 收敛后，即得到各节点的 RageRank 值。
核心代码如下：

```
A = np.zeros((120,120))
data = csv.reader(open('relation.csv','r',encoding='utf8'))
head = next(data)
nameArray=[]
for row in data:
    if(row[0] not in nameArray):
        nameArray.append(row[0])
    if(row[1] not in nameArray):
        nameArray.append(row[1])
data1 = csv.reader(open('relation.csv','r',encoding='utf8'))
head = next(data1)
for row in data1:
    source = nameArray.index(row[0])
    target = nameArray.index(row[1])
    print((source,target))
    A[source,target] = 1
    A[target,source] = 1

M = graphMove(A)
pr = firstPr(M)
p = 0.85  # 引入浏览当前网页的概率为p,假设p=0.8
rank = pageRank(p, M, pr)  # 计算pr值
result = {}

results = csv.writer(open('result.csv','w+',newline="",encoding='utf8'))
for i in range(120):
    result[nameArray[i]] = rank[i]
    results.writerow([nameArray[i],rank[i]])

print(result)
```

5.4 基于 RNN 的文本情绪分类算法

我们使用 RNN 分类模型对文本进行情感倾向的识别。
我们将单条文本进行 01 分类，0 代表消极负面的情感倾向，1 为积极正面的情绪倾向。
首先，我们将爬取到的用户推特文本存储为 CSV 格式，读取生成 RDD 后，借助实现的情绪分类模型对文本进行分类和 Map 操作，最后按照地域、时间、情感倾向进行 Reduce。流程图如下：



核心代码如下：

19

```
def emotion_recognition(tweet, model):
    tweets = []
    tweets.append(tweet)
    predictions = model.predict_classes(tweets)
    emotion = predictions.values[0][1]
    return emotion
```

```
def func(x):
    emotion=emotion_recognition(x)
    return (emotion,1)

print(time.strftime('%Y-%m-%d %H:%M:%S',time.localtime(time.time())))

conf = SparkConf().setAppName("wordcount").setMaster("local[*]")
sc = SparkContext(conf=conf)
sc.setLogLevel("ERROR")
inputdata = sc.textFile("data.txt")
output = inputdata.flatMap(lambda x: x.split("\n"))\
.map(lambda x:func(x)).\
reduceByKey(lambda a, b: a + b)

result = output.collect()
```

六、结论

6.1 特朗普推文

1. 特朗普在当上总统之后发生了疯狂涨粉的现象；
2. 特朗普的推文的受关注程度并不稳定，与特定的热点事件相关联，针对热点事件的推文出现的时候就会出现点赞数和转发数的爆炸性增长；
3. CHINA！仅凭 6 个字符就能冲进点赞转发前十，可见外国人还是很吃特朗普把矛盾转嫁给中国这一套的。我们要放弃幻想，准备战斗；
4. 特朗普不喜欢发负能量的推文，但用户们对于情绪化严重的推特更喜闻乐见；
5. 2017 年后，特朗普迅速抛弃了旧爱网页端和 Android 平台，全力在 iPhone 发推输出。

6.2 特朗普社交网络

1. Eric Trump（二儿子）、Donald Trump Jr.（大儿子）、Brad Parscale（数字总监兼政治顾问，挚友）、Ivanka Trump（大女儿），这四个人在社交关系网络中离特朗普最近，和实际情况基本是吻合的；

20

2. 在社交关系网络中，FOX 新闻的权重相对于其他如 CNN，纽约时报等媒体的权重要高很多，并且几乎所有的记者都与 FOX 有关系；
3. 特朗普十分重视综艺节目《学徒》，同时这档节目的导演及编剧也被特朗普直接关注，这是商人 Katrina Campins 女士处于社交关系网核心位置的原因；
4. 特朗普的妻子 Melania Trump 与特朗普的距离相较于他与其他儿女的距离要远很多。

思考：

在日常生活中，我们很难对一个人的社交网络进行可视化，但大数据时代提供了各种类型的数据供我们来评估两个人的关系，从而对大社交网络进行可视化。而通过关系网络，我们可以发现不同的关系集团，同时发现他们之间的链接点；除此之外，根据社交网络的聚类结果也可以获取一些人的隐藏信息及偏好，对用户画像的刻画提供帮助。

6.3 特朗普粉丝分布及用户情感地图

1. 特朗普的关注者数量最多的前四个州（加州、德州、佛罗里达、纽约州）的 GDP 贡献率就是全美前 4；
2. 特朗普总统 Twitter 关注者的分布基本正比于美国的人口分布，但是关注者数量排名第一的加州是第二名德州的 1.44 倍。这可能与加州的产业特性、移民数量相关：加州拥有全美最高的移民数量占比（27%），加州华裔人口占全美总华裔人口的 36.9%；
3. 2018 年与 2020 年美国各区域网民对于#Trump 的发言频率基本持平。其中，纽约州和加州遥遥领先；
4. 恶性事件与节日均对于人们在#Trump 发表相关评论时的情感表达具有一定影响；
5. 2020 州级情感倾向基本与 2020 大选结果布局相同，一些存在偏差的州或被爆出选票造假，设想如果数据量大到一定程度的话，可以根据情感倾向地图去初步鉴别可能存在选票造假的州；
6. 美国网民对于#Trump 相关的负面内容有较大的发言意愿，一些对于#Trump 发言异常积极的用户显得格外突出，这种方法经过进一步改造后，或可以用于基于舆论大数据的舆论实时监督，可以及时发现疯狂带节奏的异常用户群体；

21