Coronavirus Social Media Analytics

Yuyang Zhou(10105592), Jinpei Han(9865011), Yilei Wu(9981601), Hongbo Zhu(10313053)
University of Manchester, UK

October 14, 2020

1 Introduction

The coronavirus has been a global pandemic in 2020. It has changed people's lifestyle in many aspects. Social media is one of the key channels for people to express their opinions on coronavirus outspread. This thesis investigated four proposed questions related to coronavirus using Social Media Analysis on Twitter.

2 Research Questions:

2.1 Health data statistics

During the coronavirus epidemic in the UK, people's major concern has been the capability of the HNS. The NHS publish daily update tweets which contain cumulative confirmed cases, newly confirmed cases, new suspected cases, increased deaths and newly cured cases. However, the tweets are often limited to the text form, which leads to a strong "stacking sense" of the data and cannot present the overall trend of the data. The emergence of graphical tables meets the public's demand for clear and intuitive information and data. Thus, this form is often more acceptable to the public than the text + data form, and the data is displayed at a deeper level so that the audience has intuitive visual thinking on the problem.
The Department of Health and Social Care's Twitter account @DHSCgovuk has been publishing data on the outbreak of new coronary pneumonia since January 25, 2020, and insist on updating it daily including the number of new coronavirus tests, diagnoses and deaths. By accessing the Twitter API and calling the get_timeline function, 3200 historical tweets were obtained from the account @DHSCgovuk. Then the data is filtered through a series of filters, and finally, the epidemic data from January 25, 2020, to April 13, 2020, is obtained from Twitter and processed into charts. The result is shown in Figure 5-7.

2.2 Sentiment Analysis - people's reaction to UK's corona-virus policies

As the coronavirus pandemic spreading across the UK and affecting more and more people, the UK government has been gradually implementing policies to counter the spread of the virus. Since the 16th of March, the UK government started daily briefing to publish and update measures every day. This research question is interested in people's daily reaction to the policy and guidance posted by Prime Minister Boris Johnson. Data in 15 days, between the first day of daily briefing on the 16th of March and the 31st of March, were recorded every day. The data was processed by sentiment analysis which finds the composition of emotions including anger, anticipation, disgust, fear, joy, sadness, surprise and trust within people's tweets.
Since the UK Prime Minister Boris Johnson has been posting the majority of government updates and daily briefing during this time, many people have been posting feedback directly towards Boris Johnson on Twitter. Therefore, 'Boris Johnson' is used as a keyword to see people's view of his policies during this time frame. The major topics people were discussing were found by generating a topic model. The free version of twitter API only supports fetching time-specific tweets in the last seven days, which restricted the pipeline structure. Therefore, the data have to be collected and stored daily, which increased the workload and complexity of the procedure. One thousand tweets with the keyword 'Boris Johnson' were collected and analysed every day during the time frame. The result is shown in Figure 4 in the Appendix.

2.3 Keywords Mining - Word Cloud & Frequency

The focus of this research question is on people's discussion about coronavirus, starting with finding out the corresponding keywords based on the

statistics of word frequency. Then a word cloud is constructed to show what people care about in the epidemic.

First of all, 3200 historical tweets containing the keyword ”COVID-19” were extracted. Then special characters and links in the text were removed. The tm-map() function was used to clean tweets with following steps: remove punctuation, remove English common stopwords, remove numbers and eliminate extra white spaces. In order to get a more meaningful vocabulary, we additionally removed some stopwords such as ”get”, ”may”, ”can” and ”will”. A term-document matrix was formed given the processed tweets to show the most frequently used words and their corresponding frequency. Finally, a word cloud was generated to illustrate the significance of the words.

2.4 Related Sensitive Topic - ”What place discuss COVID-19 most?”

The main topics discussed previously would derive some less relevant topics, where some of them are sensitive and to some extents reflects worldwide people’s attitude towards them, which can be discussed in many aspects. ”ChineseVirus” is selected as a sensitive topic that was recently brought up and controversial.

Since the topic is new and controversial, to obtain more information, the geographical location of tweeting users is also recorded for analysing purposes. Obtaining and counting the location information is difficult since not all users filled their location information on their twitter account profile, and even if some people did, the format could be different including upper or lower cases. To alleviate the problem, after filtering information other than English characters, only the country is taken as the information of location, and if there is only a city in the profile, then the city would be recorded as the location. For this topic, 2000 samples are extracted, yet only around a quarter of these (around 500) meet the requirement.

3 Result & Analysis

3.1 Health data statistics

According to Figure 1, the number of positive patients per 100 testers is gradually increasing from the 25th of January to the 4th of April, where it reaches 25%. The diagnosis rate will maintain at this level afterwards. The result shows that the single-day positive test rate is higher than that of cumulative one, which means the cumulative positive test rate will be increased in the future for a short period of time, which also implies the overall number of diagnoses will rise.

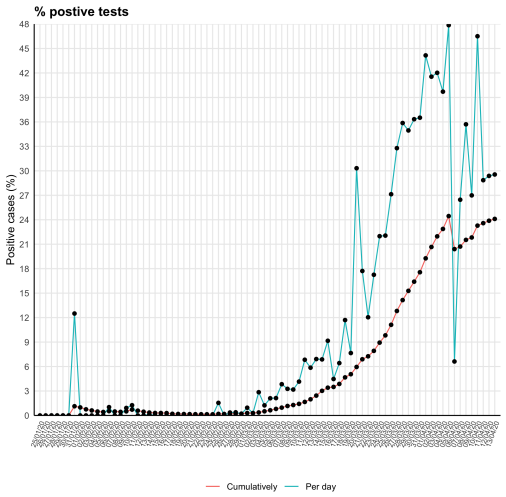

Figure 1: positive test rate

In epidemiology, the basic reproduction number (R0) of infection can be thought of as the expected number of cases directly generated by one case. A simple model was used to calculate the reproductive rates, to show the numerical growth trend of confirmed cases and deaths:

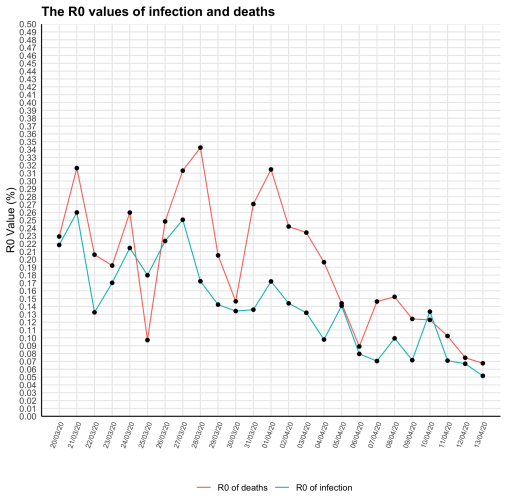

Figure 2: The basic reproduction number (R0) In epidemiology

Conclusions can be drawn from Figure 2:

1.The situation is getting better.

As restrictive measures such as social distancing and lockdown were introduced, both R0 values of infection and deaths are gradually decreasing. It does not mean that the numbers aren’t growing, but the rates of growth are slowing down.

2. The NHS don't have enough tests.
The R0 curve of deaths is above the infection curve while it supposed to be at the same level or lower, which indicates that there are much more infected people than reported and we're not testing enough.

3.2 Sentiment Analysis - people's reaction to UK's corona-virus policies

The result shown in Figure 4 is the daily change of percentage emotion composition of 1000 tweets mentioning Boris Johnson from the 16th of March to 31st of March. Some data labels were placed onto data points to indicate the content of daily briefing posted by the UK government, which is usually the major topic found by topic modelling on that day. During this pandemic, people's major reactions are fear, trust, anticipation and sadness. The result can be used to indicate if people like or dislike the government's action. In most cases, the trust percentage increases whenever The government published beneficial policies. However, people do not always appreciate the government's action. For example, The UK lockdown significantly increased fear rather than trust. Furthermore, on the 28th of March, the PM wrote a letter to every household to urge them to stay at home. Even though it sounds like a positive action, the majority tweets on that day discussed how this action is a waste of money, and this money should have been spent on funding the NHS. Thus decreased the trust level on that day.

3.3 Keywords Mining - Word Frequency & Cloud

The word cloud is shown in Figure 8 and the term-document matrix can be found in Figure 9-18 in the appendix. In the word cloud, the higher the frequency of the words, the closer the words to the center, the larger the font. The most frequent and most noticeable word is undoubtedly the keyword employed to retrieve tweets: "covid". Then, other words with relatively high frequency are "health", "world", "coronavirus", "crisis". "funding", "work", "organization" and "deaths", which can be found in Figure 8. From the words presented above, it can preliminarily be concluded that the epidemic situation is gradually becoming a global crisis. People's work are affected, so economic problems ensue. The daily increase of the number of deaths is also receiving continuous attention.

Some non-English stopwords such as: "las" and "con" can be found in Figure 8 even though some other non-English stopwords were removed in the cleaning step. Since Twitter is a global social platform, users from all over the world tweeted with various languages. Some non-English characters can be easily cleaned up, but there are also phrases made up of English letters that are difficult to rule out. Such phrases are likely to be made of stopwords in other languages with no practical meaning. Even if some phrases have their own meaning, it is difficult for us to guess its meaning without context.

3.4 Sensitive Topic

The number of tweets sending by users from different places is counted, and the results are shown in Figure 3. The Figure shows that India is the one who mostly discusses this topic, where the number of tweets is around 100. The result is not strong enough as the data retrieved is incomplete, also since the formatting processing is influencing the result, the result would be more precise if the number of tweets is larger.

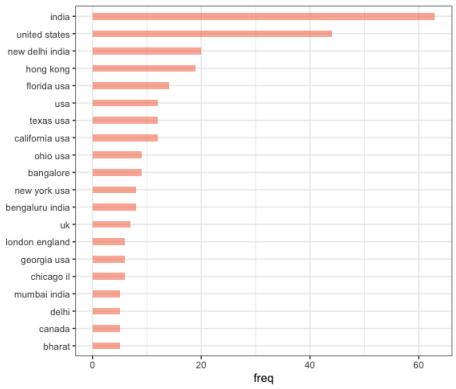

Figure 3: Location distribution of tweets with "#ChineseVirus"

4 Conclusion

Overall, we have designed and implemented an SMA model that analyzes the problems regarding the recent topic "COVID-19", including Health data statistics, Sentiment Analysis, Topic Modelling, Keywords Mining, and Geospatial Analysis, where each of these contains a proper evaluation.

Appendices

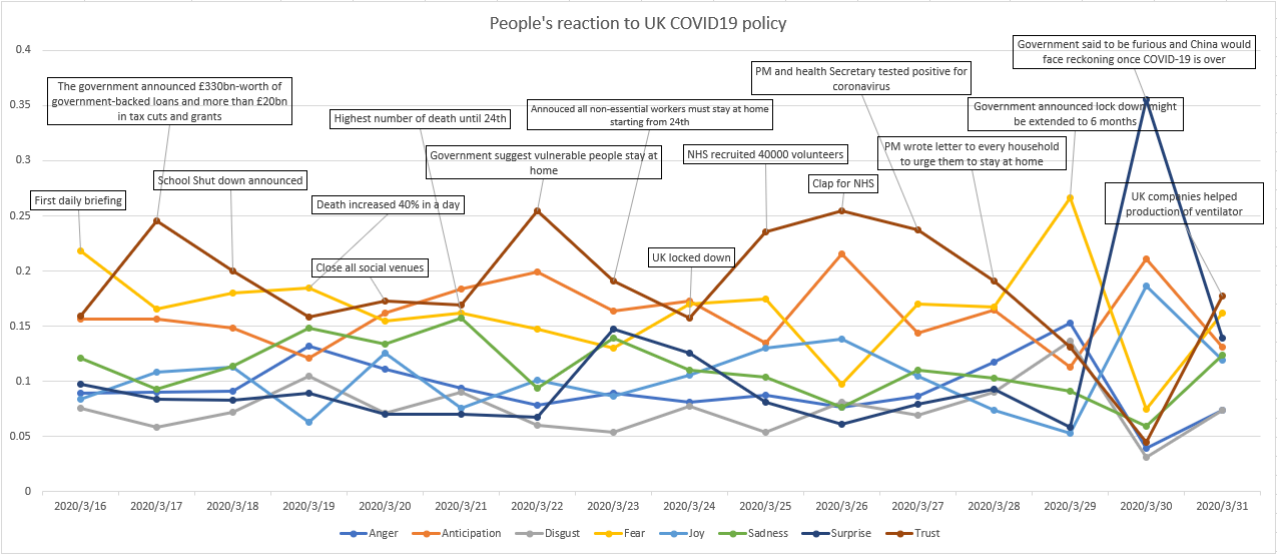

Figure 4: People's reaction to UK COVID-19 policy

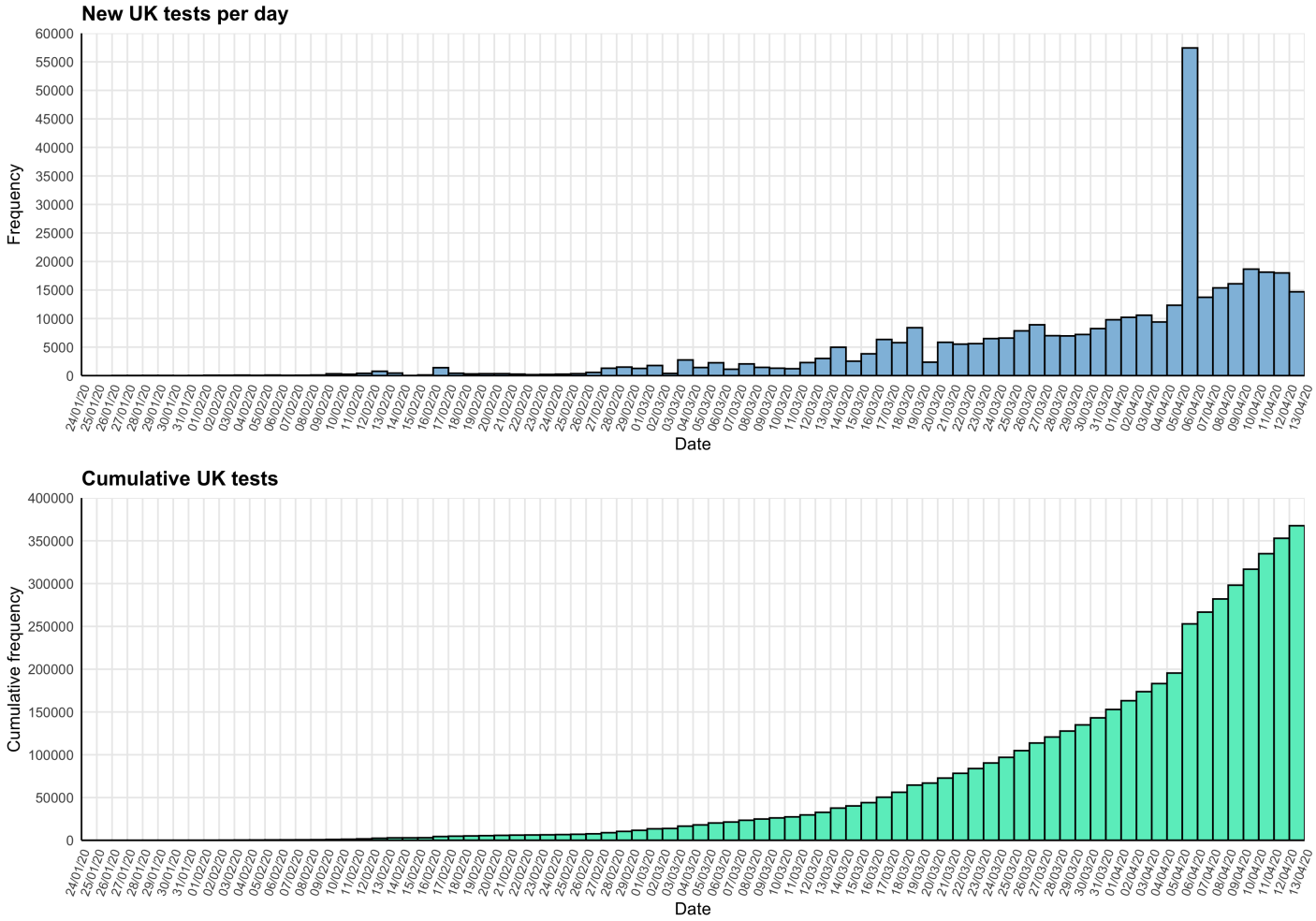

Figure 5: UK COVID-19 Daily Tests data

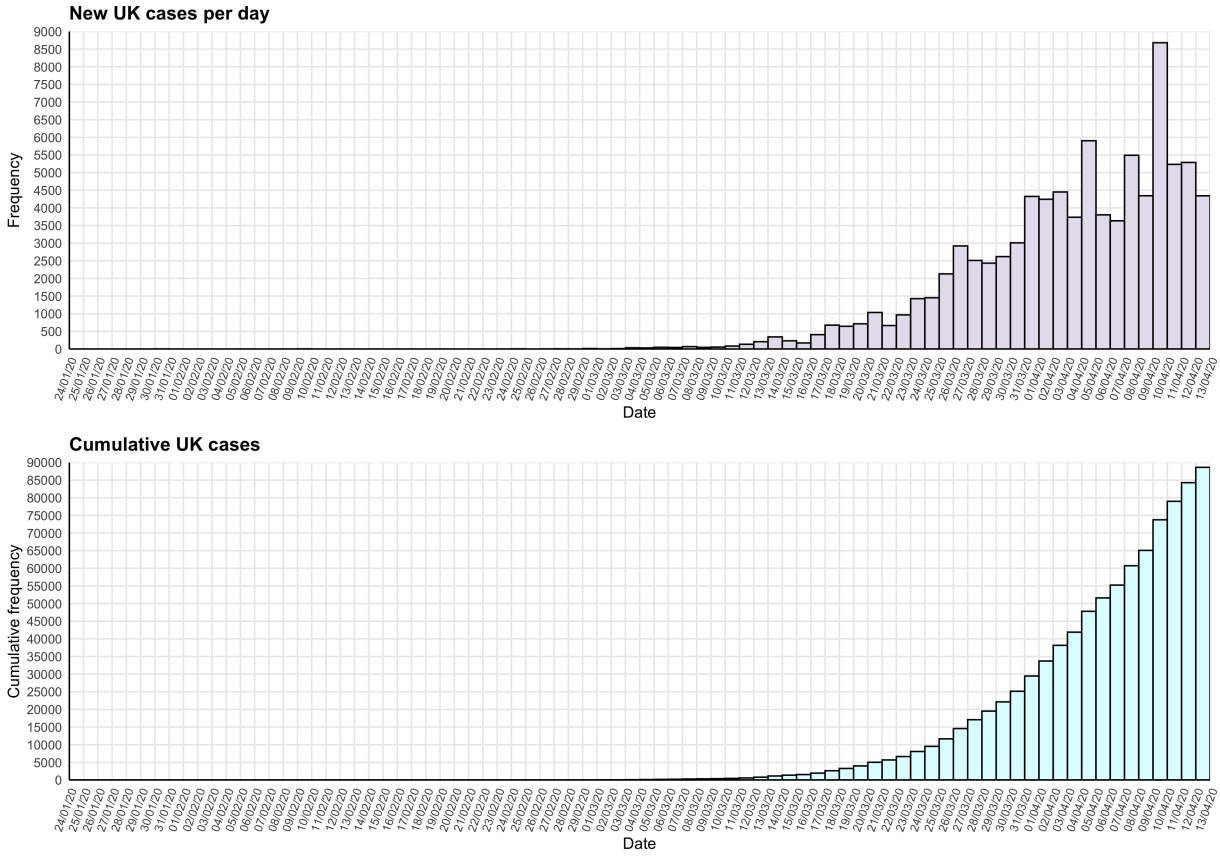

Figure 6: UK COVID-19 Daily Cases data

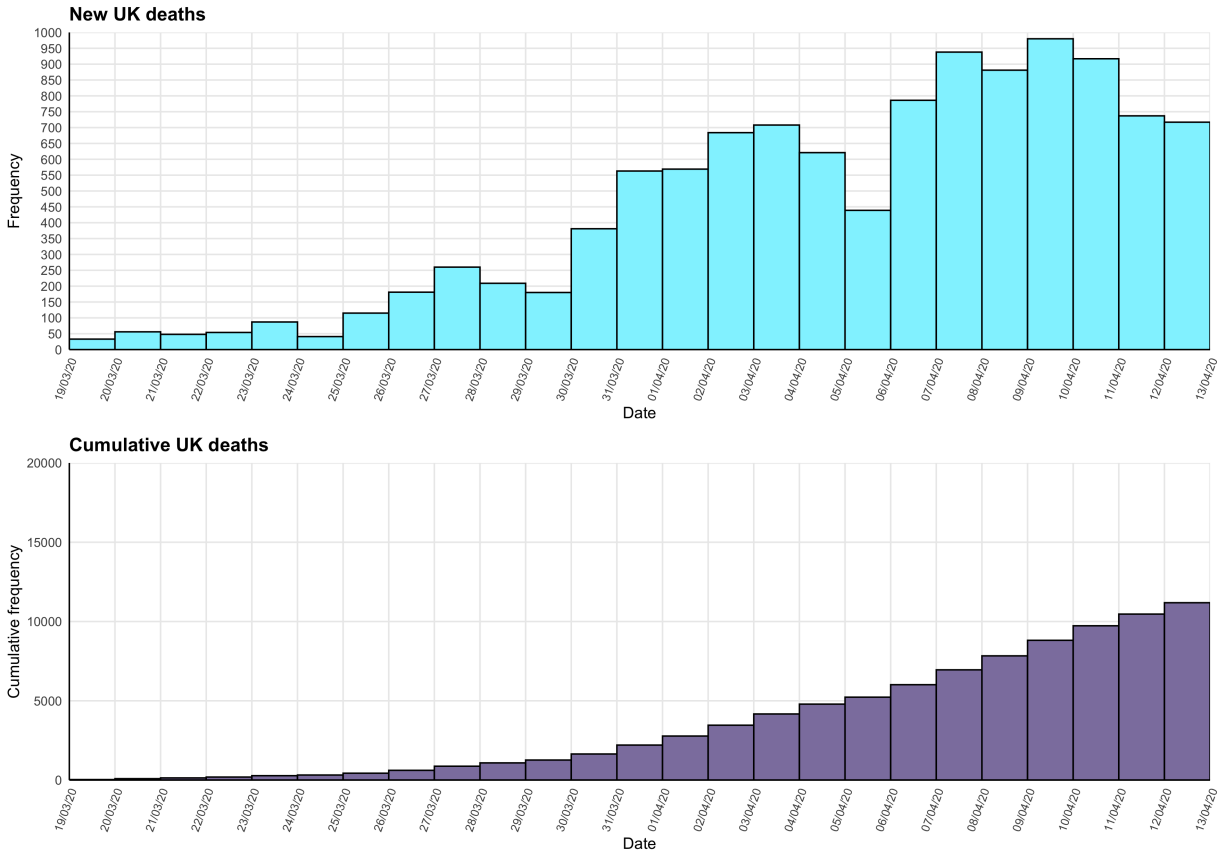

Figure 7: UK COVID-19 Daily Deaths data

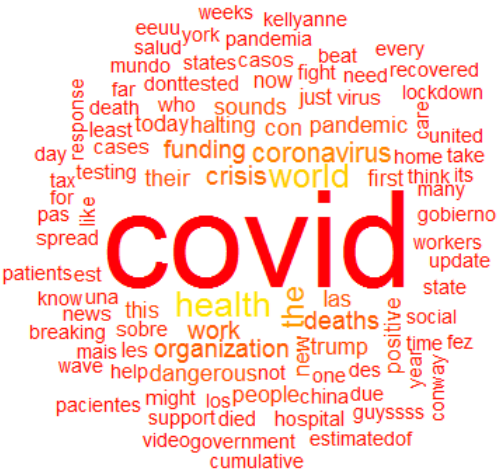

Figure 8: Word Cloud

	word <fctr>	freq <dbl>
covid	covid	1809
health	health	271
world	world	244
the	the	191
coronavirus	coronavirus	160
crisis	crisis	130
funding	funding	115
work	work	113
organization	organization	110
deaths	deaths	109

Figure 9: term-document matrix

	word <fctr>	freq <dbl>
las	las	101
their	their	98
new	new	97
con	con	97
halting	halting	96
dangerous	dangerous	95
sounds	sounds	94
pandemic	pandemic	88
trump	trump	86
this	this	79

Figure 10: term-document matrix

	word <fctr>	freq <dbl>
people	people	78
positive	positive	67
just	just	62
now	now	62
first	first	60
today	today	59
tested	tested	55
who	who	55
sobre	sobre	54
cases	cases	52

Figure 11: term-document matrix

	word <fctr>	freq <dbl>
testing	testing	51
one	one	49
una	una	47
les	les	47
virus	virus	43
fight	fight	43
casos	casos	43
might	might	41
china	china	41
like	like	39

Figure 12: term-document matrix

	word <fctr>	freq <dbl>
died	died	39
least	least	39
hospital	hospital	39
spread	spread	38
est	est	37
news	news	37
help	help	36
states	states	36
social	social	36
due	due	36

Figure 13: term-document matrix

	word <fctr>	freq <dbl>
support	support	35
dont	dont	35
home	home	35
state	state	34
pandemia	pandemia	34
update	update	34
workers	workers	33
time	time	33
gobierno	gobierno	32
think	think	32

Figure 14: term-document matrix

	word <fctr>	freq <dbl>
care	care	32
death	death	32
des	des	31
many	many	31
need	need	31
breaking	breaking	31
beat	beat	31
government	government	30
york	york	30
for	for	30

Figure 15: term-document matrix

	word <fctr>	freq <dbl>
salud	salud	30
year	year	30
cumulative	cumulative	30
know	know	30
response	response	30
guyssss	guyssss	30
patients	patients	29
mundo	mundo	29
pas	pas	29
los	los	29

Figure 16: term-document matrix

	word <fctr>	freq <dbl>
estimatedof	estimatedof	29
kellyanne	kellyanne	28
not	not	28
tax	tax	28
mais	mais	28
lockdown	lockdown	28
pacientes	pacientes	28
united	united	28
wave	wave	28
recovered	recovered	28

Figure 17: term-document matrix

	word <fctr>	freq <dbl>
far	far	28
conway	conway	27
video	video	27
take	take	27
fez	fez	27
its	its	27
every	every	26
weeks	weeks	26
eeuu	eeuu	26
day	day	26

Figure 18: term-document matrix