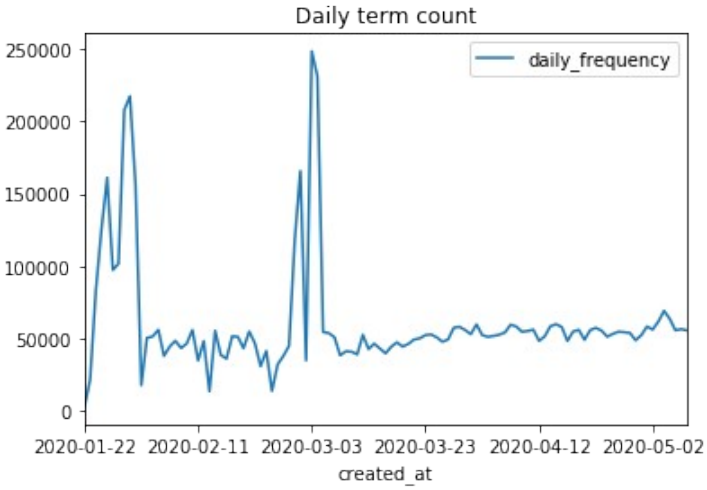
Outbreak and Symptom Term analysis

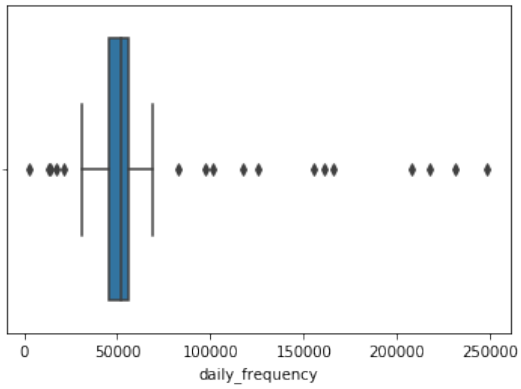
The data used is a subset of the original data from 22-01-2020 to 08-05-2020 and consists of ~11 million rows.



1) Daily Count

```
Unusually large counts(>100000) were observed on -

3      2020-01-25
4      2020-01-26
6      2020-01-28
7      2020-01-29
8      2020-01-30
9      2020-01-31
37     2020-02-29
38     2020-03-01
40     2020-03-03
41     2020-03-04
```



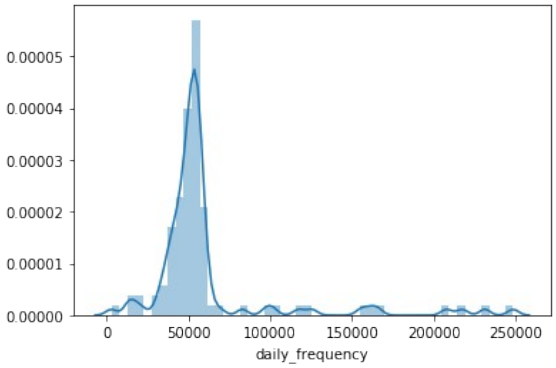
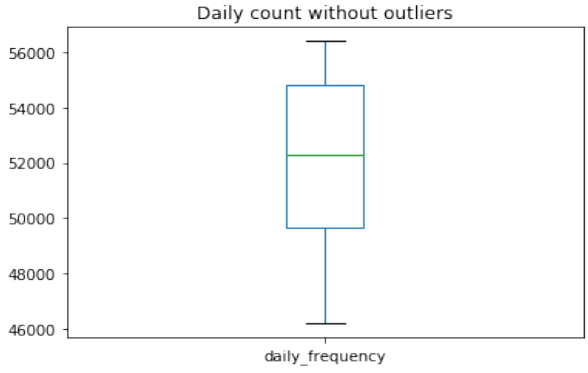


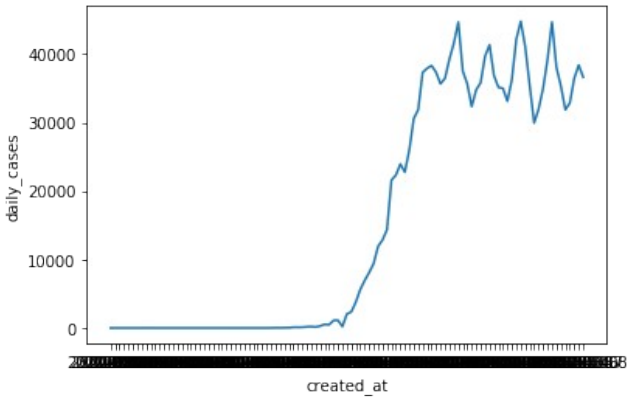
Fig. Distribution plot of daily term count



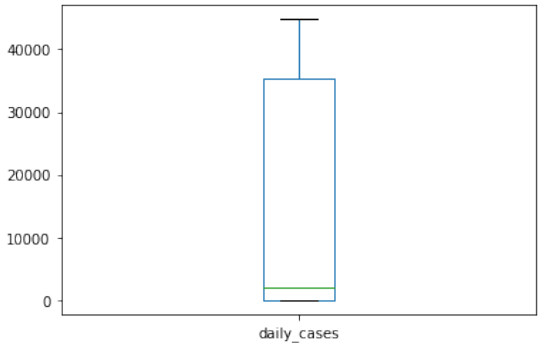
From the boxplot and distribution plot, median count is nearly 52000 and distribution is more or less symmetric if outliers are ignored.

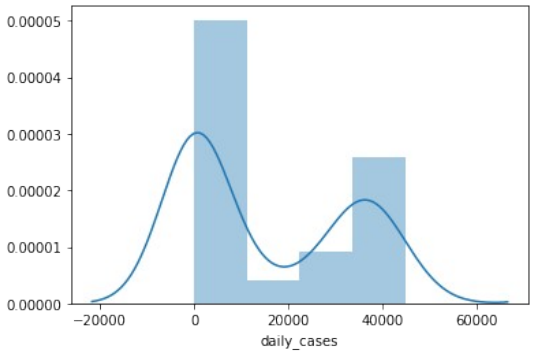
Note : “Daily cases” and “daily deaths” were obtained from https://datahub.io/core/covid-19#readme. Daily deaths and daily cases were obtained for the countries - `'Australia', 'Canada', 'India', 'New Zealand', 'US', 'United Kingdom'`

2) Infection Rate

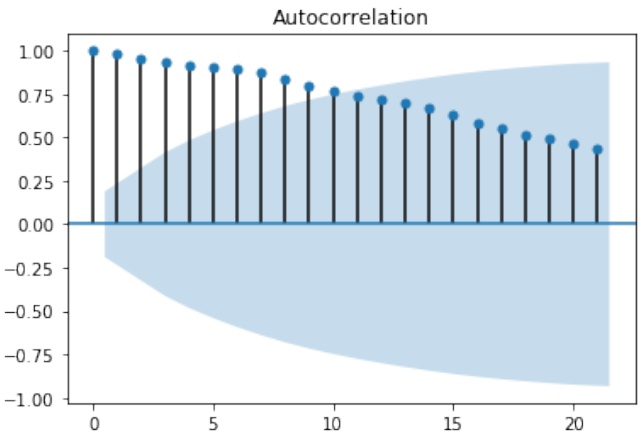


Initially, number of cases reported were low, but as testing ramped up, daily_cases began rising too.





No outlier is apparent from the boxplot, but the distribution is highly skewed with median around 2000.



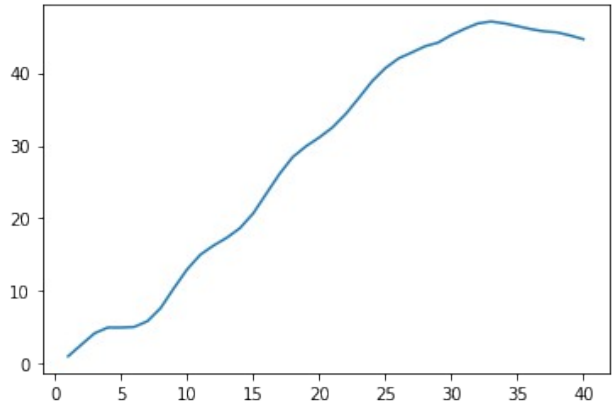
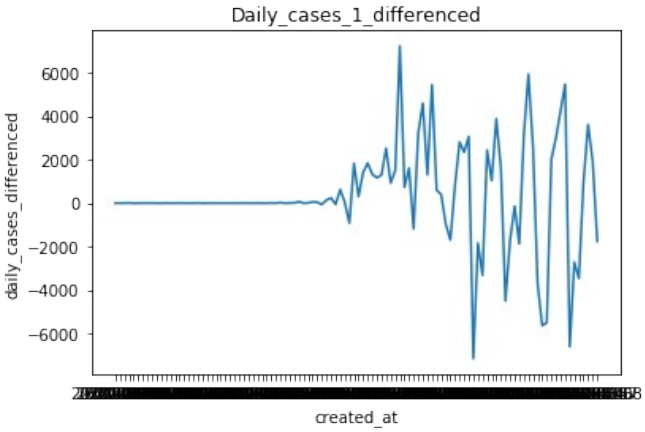
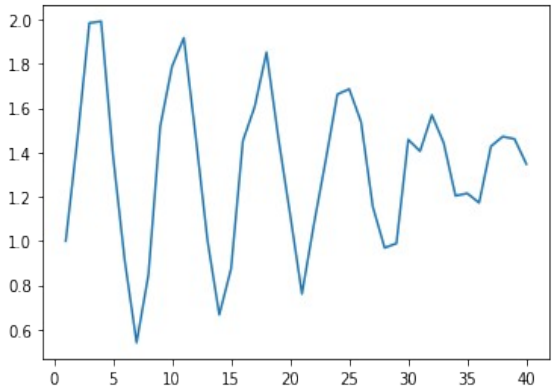


Fig – Variogram analysis

In the ACF plot, the acf values are significant and cutoff at high lags. Also the variogram values also increase with lags. This suggests that the “infection_rate” is nonstatioary and needs at last one order of differencing.

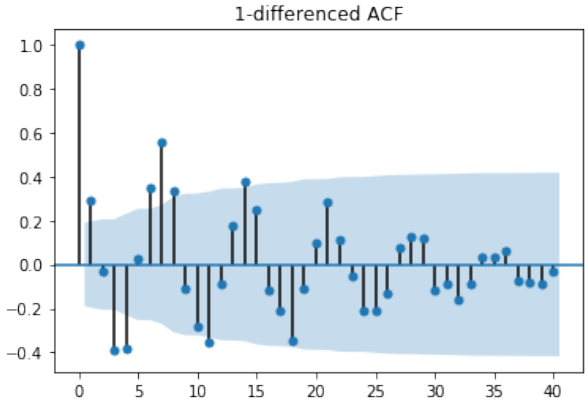
The differenced plot looks like -

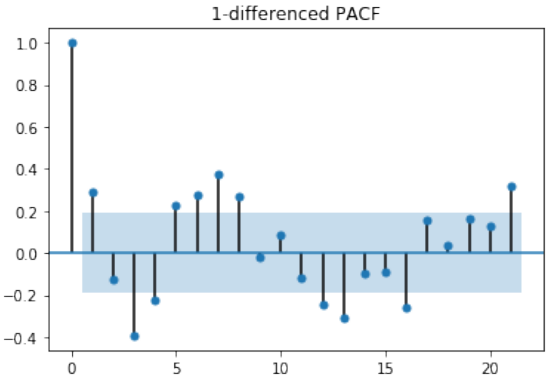




Variogram analysis of differenced series

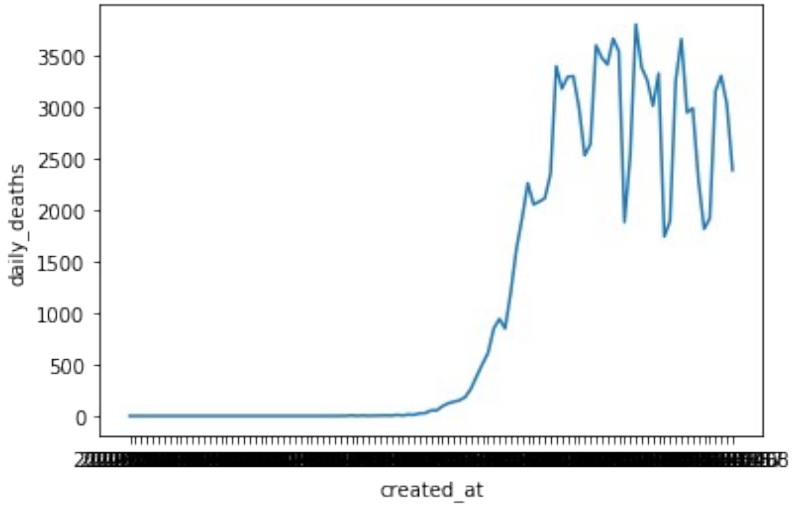
The differenced series looks stationary with mean around 0.



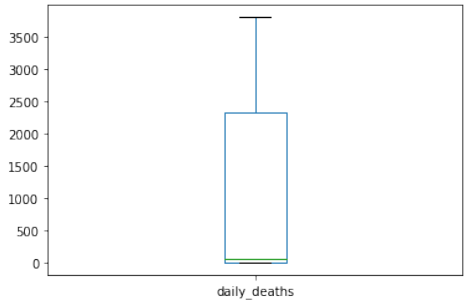


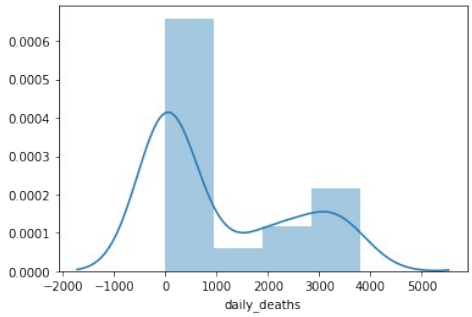
Both ACF and PACF plots show a damped sine wave which suggests an ARMA process of suitable order. Thus, daily_cases may be modelled by ARIMA process.

2) Daily Deaths

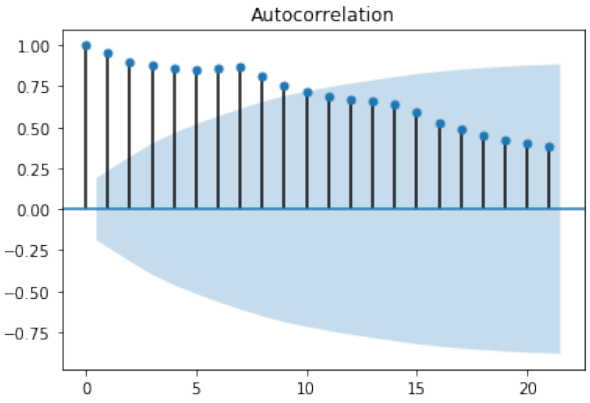


As expected, daily_deaths rise with daily cases.

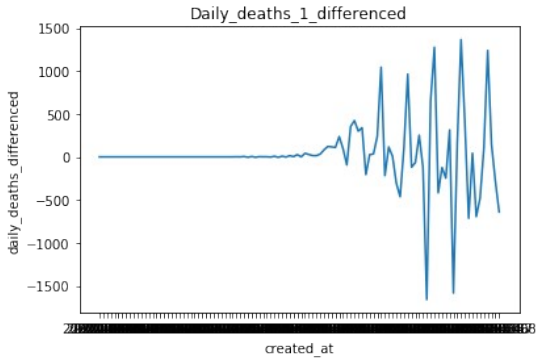




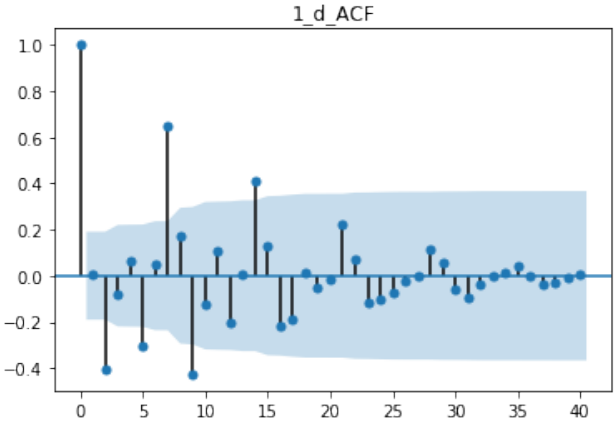
From the boxplot, no apparent outlier is observed. However, the distribution of daily deaths is highly skewed with median around <100 as seen in distribution plot as well as boxplot.

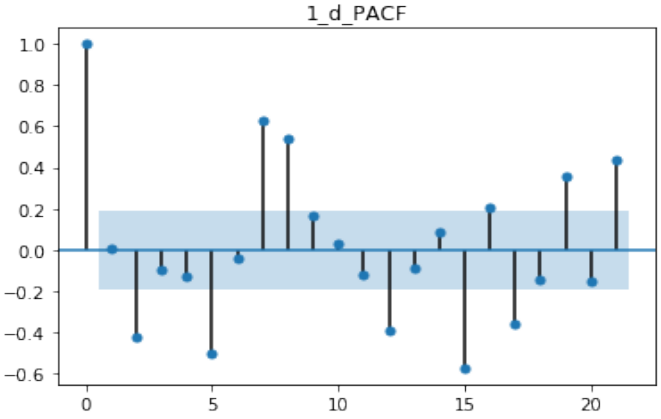


The acf plots show signficant acf values at high lags which suggests non-stationarity in time series.



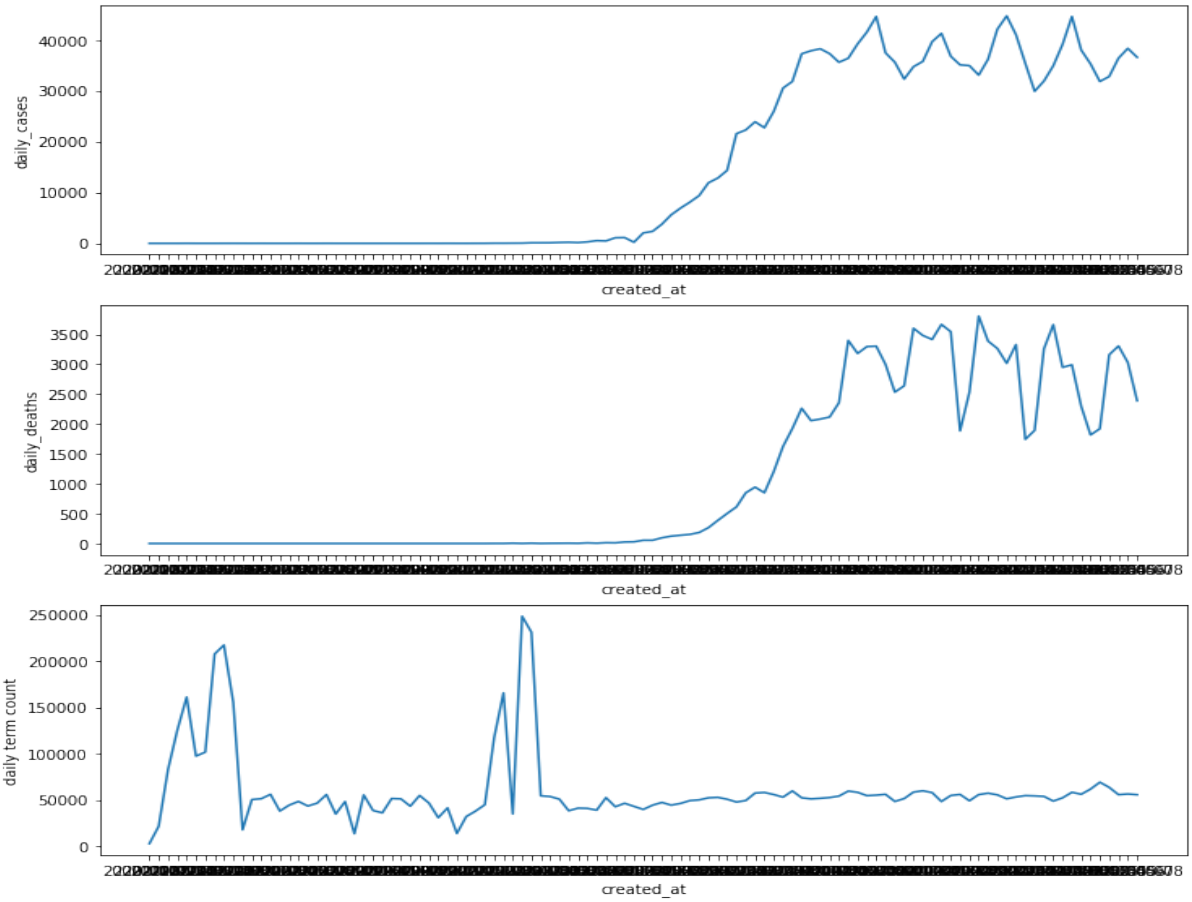
After differencing, daily deaths resembles a stationary time series.

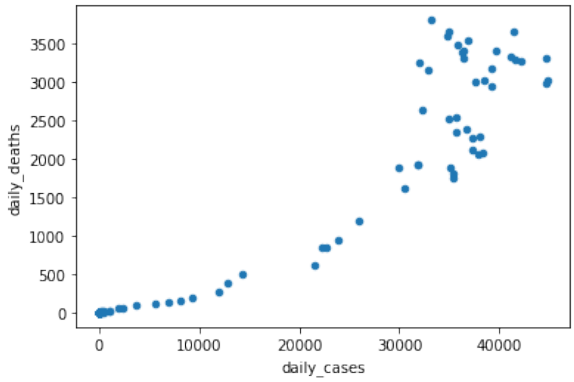




From ACF and PACF plots, it seems the differenced daily_deaths time series follows ARMA process of suitable order.

Time series plots of daily_deaths,daily_cases and daily term count are plotted below





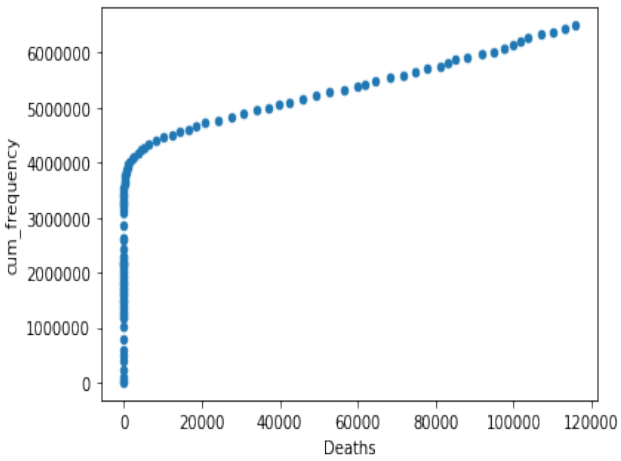
From the above graphs, daily_death and daily_cases look similar and have high correlation coefficient `0.9509412732821179`, which is expected because as daily cases rise, daily deaths increase too.

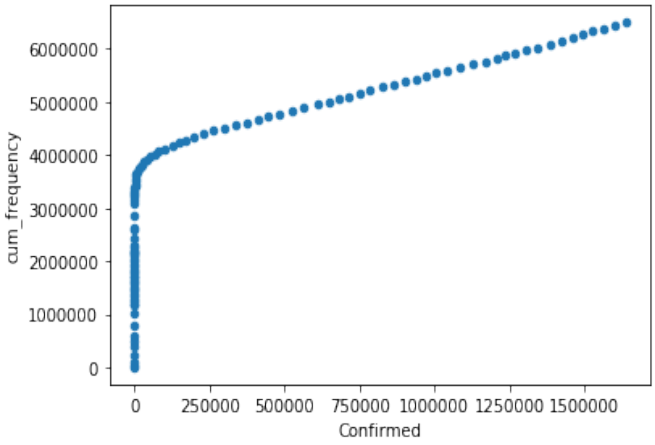
However, daily_count doesn't have much correlation with daily_cases and daily_deaths with the respective correlation coefficients being -

`-0.12905853084101224`

`-0.10939421802504366`

But, there is high correlation between “cumulative term counts” and “cumulative cases” and “cumulative deaths” as shown below





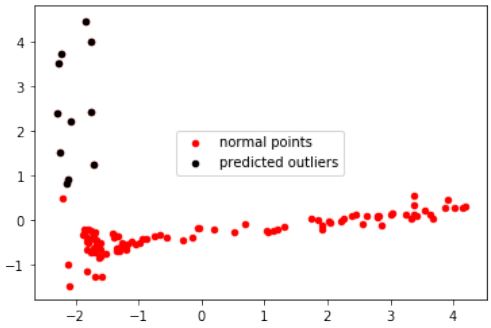
correlation-coefficient : 0.7973005485418727

correlation-coefficient : 0.8281780247336641

Anomaly Detection

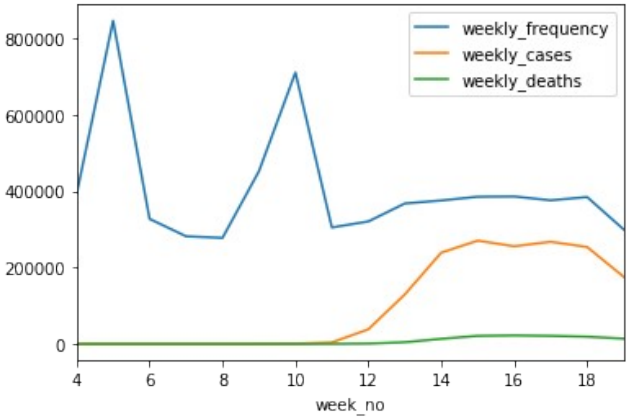
Using the features- daily_cases, daily_term_count, daily_deaths, cumulative_term_count, cumulative_deaths, cumulative_cases

sklearn’s Local Outlier Factor algorithm was used on normalised data. The features were projected on 2-D space using PCA.

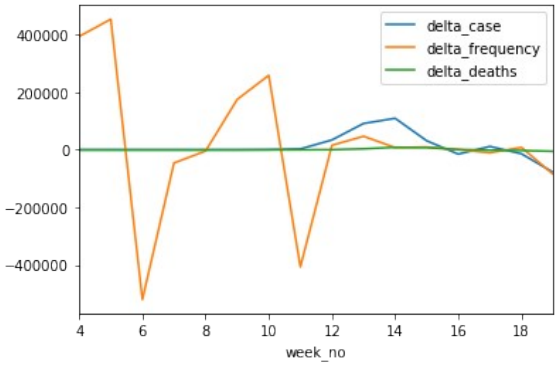


The algorithm classified those datapoints as outliers for which daily_cases and deaths were few, but cumulative and daily frequency of terms were high.

Weekly analysis



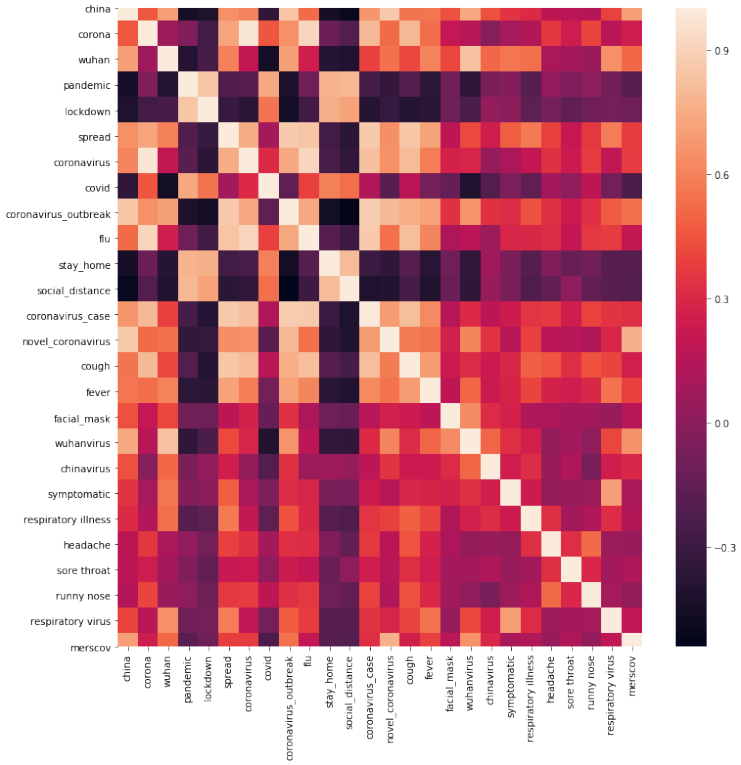
Weekly data represents a similar trend as daily data. Weekly count of terms have unusually high peaks at week no.s 5 and 10, but is below 400000 in he remaining weeks.
Weekly cases show a high correlation with weekly_deaths with correlation coefficient being 0.9754733321404416, whereas weekly_cases and weekly_deaths continue to have low correlation with weekly term count.



Weekly change in term counts fluctuate between positive and negative values, whereas weekly change in deaths and cases is nearly zero across the weeks. However, there is surge in weekly cases from week 12 to week 14.

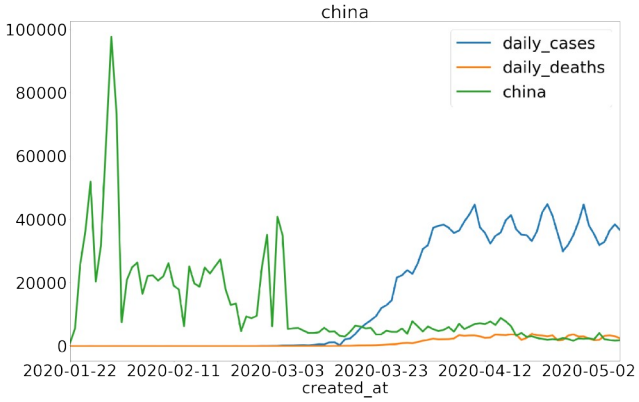
Termwise-Analysis

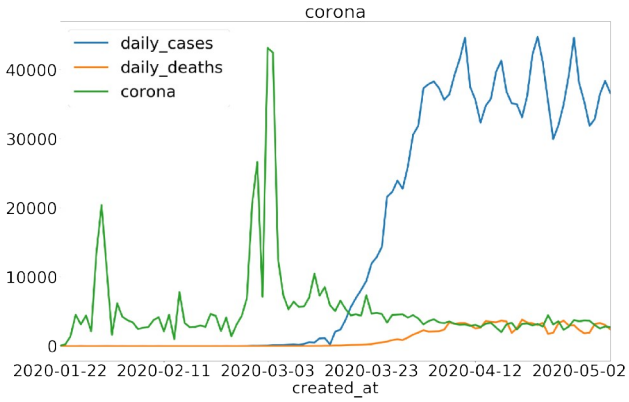
A set of outbreak and symptom terms were identified and the frequency of occurences of those terms in individual tweets were obtained. From the set of terms, top frequently occuring terms were chosen on which further analysis has been done

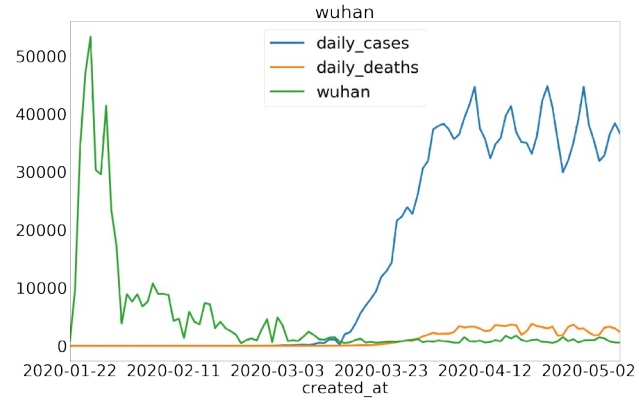


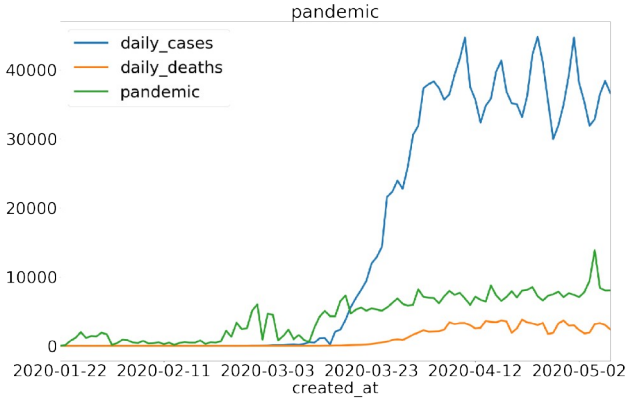
It can be seen that some terms have terms have high correlation with other terms whereas some have little or no correlation at all.

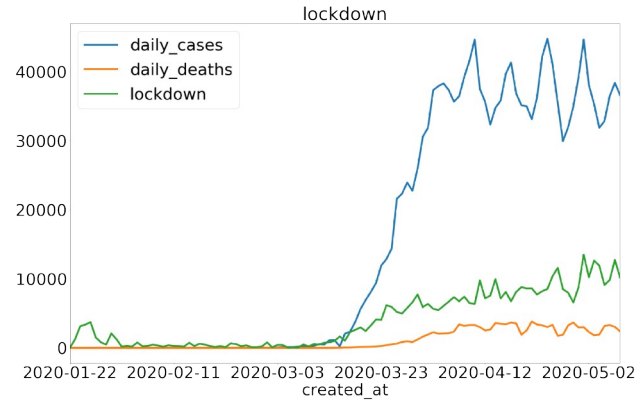
The time series plots below is plotted for each of those chosen terms:

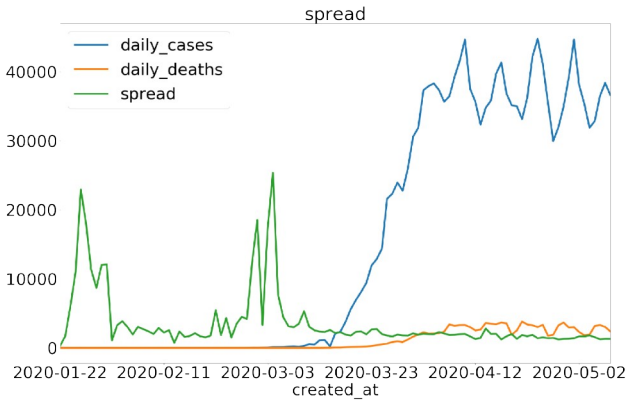


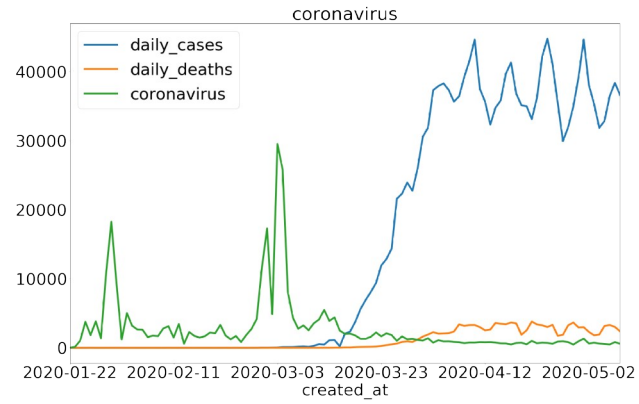


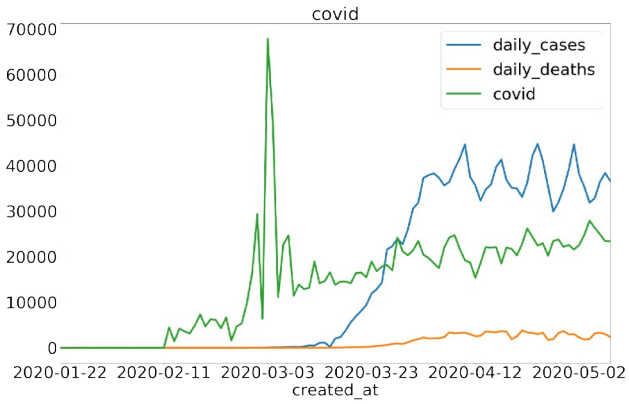


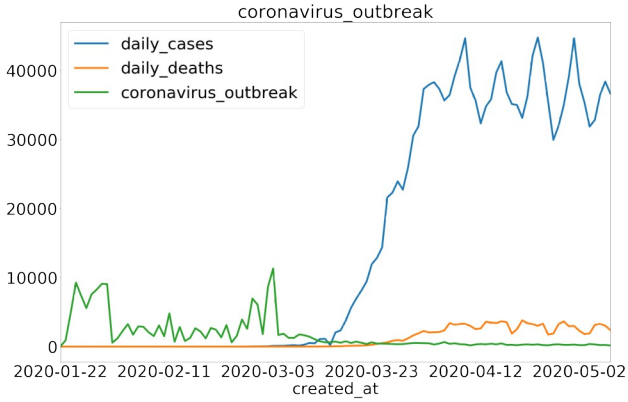


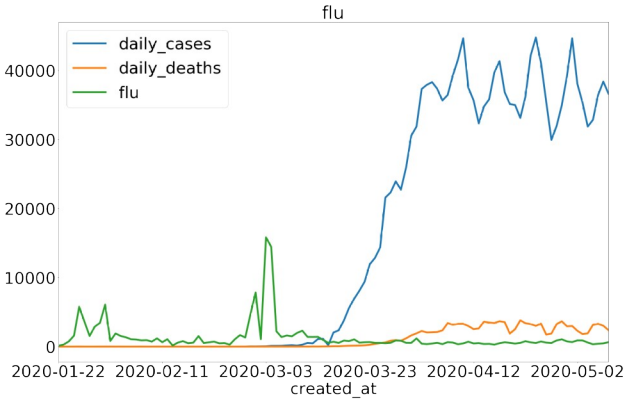


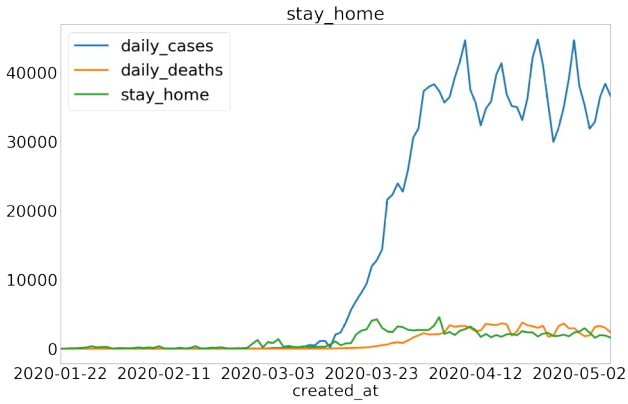


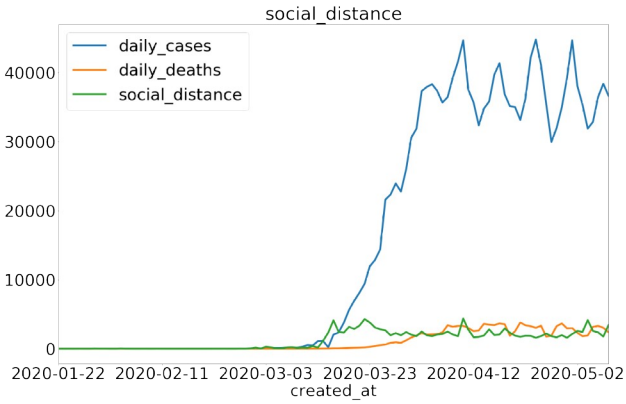


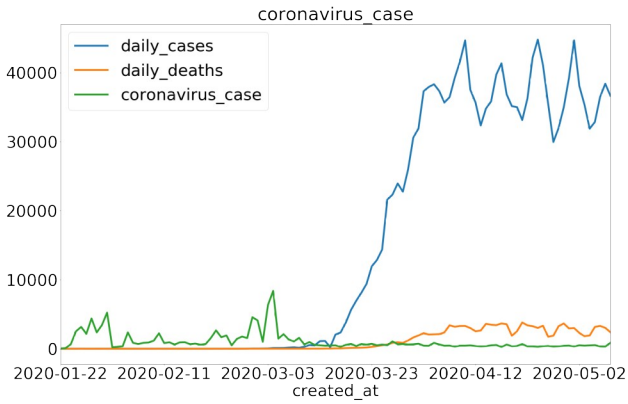


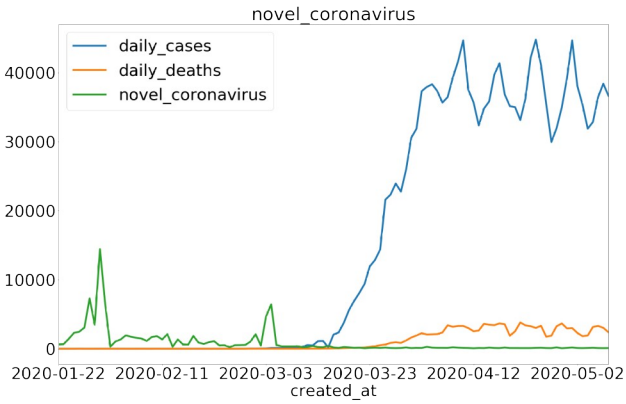

coronavirus_outbreak
daily_cases
daily_deaths
coronavirus_outbreak
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

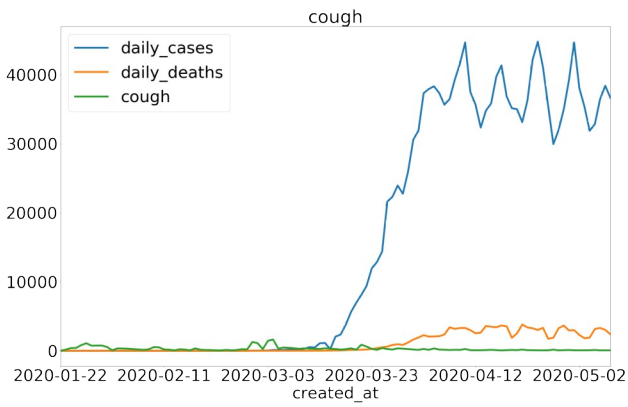

flu
daily_cases
daily_deaths
flu
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

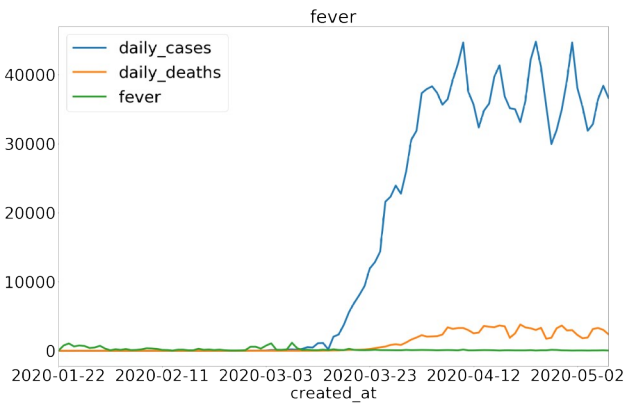

stay_home
daily_cases
daily_deaths
stay_home
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

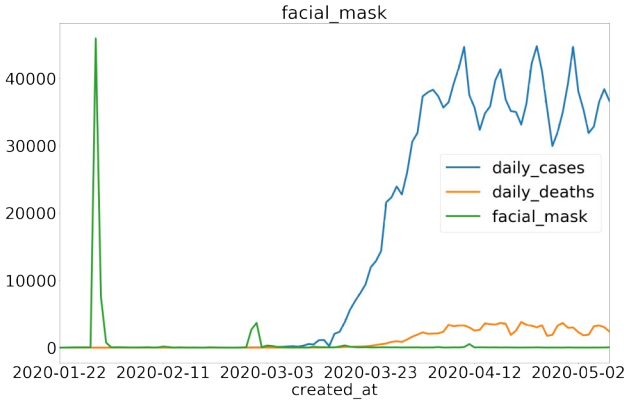

social_distance
daily_cases
daily_deaths
social_distance
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

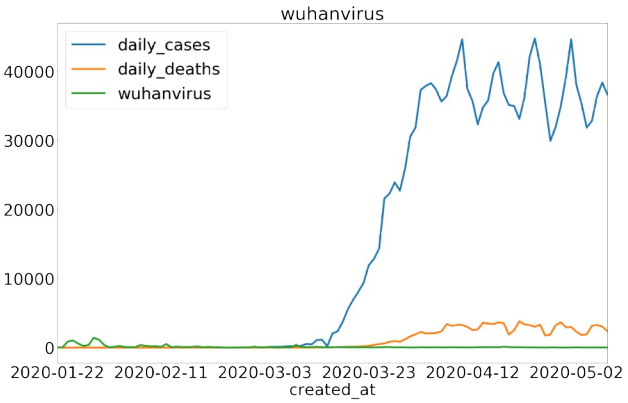

coronavirus_case
daily_cases
daily_deaths
coronavirus_case
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

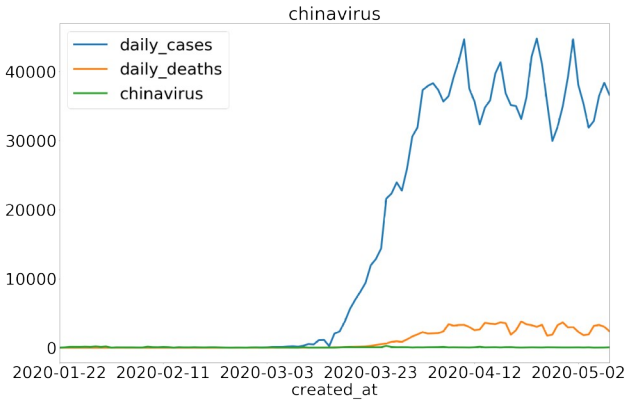

novel_coronavirus
daily_cases
daily_deaths
novel_coronavirus
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

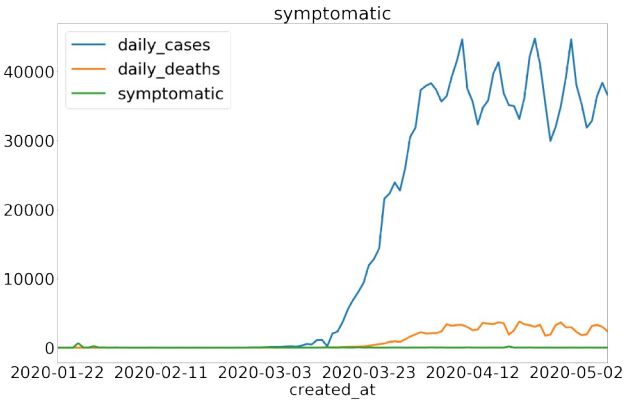

cough
daily_cases
daily_deaths
cough
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

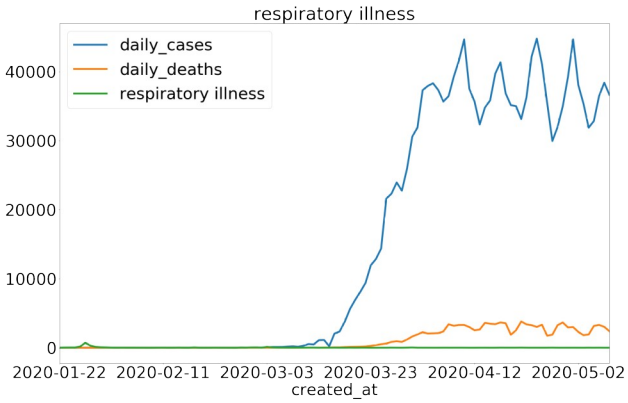

fever
daily_cases
daily_deaths
fever
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

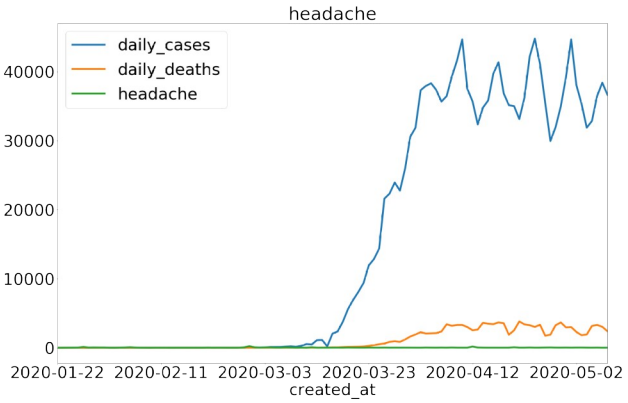

facial_mask
daily_cases
daily_deaths
facial_mask
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

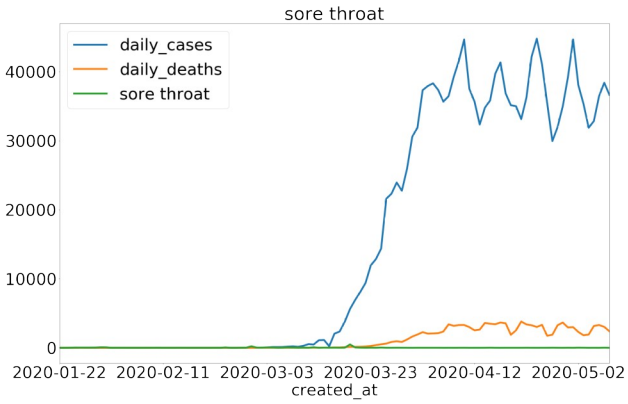

wuhanvirus
daily_cases
daily_deaths
wuhanvirus
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

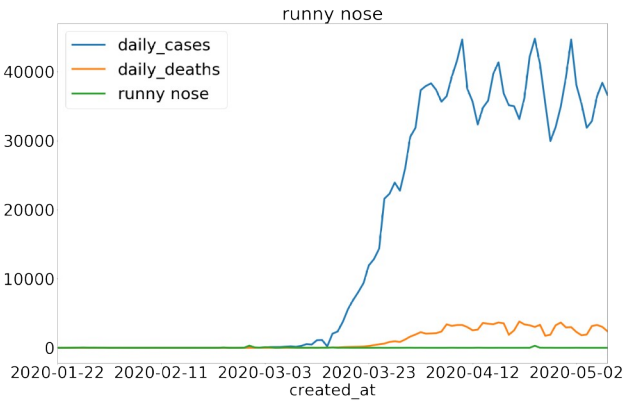

chinavirus
daily_cases
daily_deaths
chinavirus
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

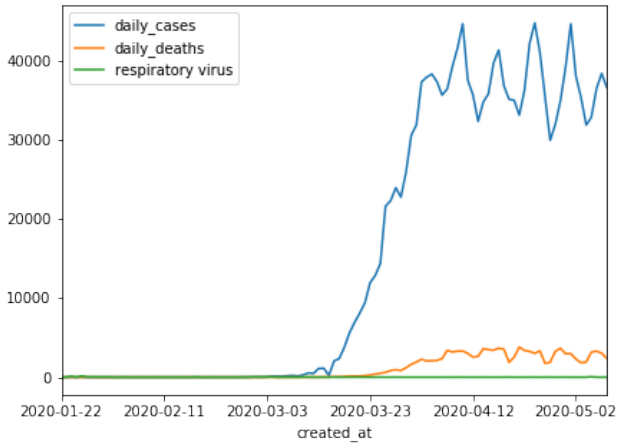

symptomatic
daily_cases
daily_deaths
symptomatic
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at

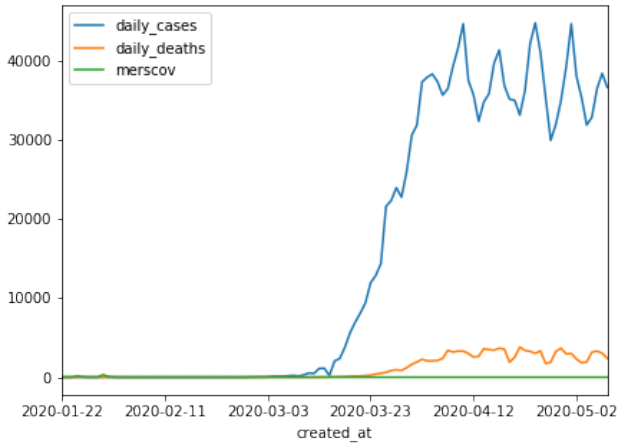

respiratory illness
daily_cases
daily_deaths
respiratory illness
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at


headache
daily_cases
daily_deaths
headache
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at


sore throat
daily_cases
daily_deaths
sore throat
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at


runny nose
daily_cases
daily_deaths
runny nose
40000
30000
20000
10000
0
2020-01-22 2020-02-11 2020-03-03 2020-03-23 2020-04-12 2020-05-02
created_at



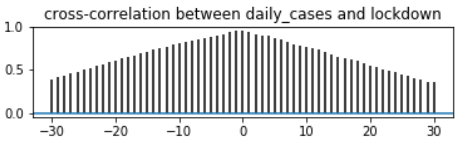


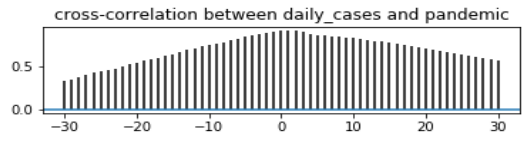
From the time series plots above, it can be seen that certain keywords like “china”, “corona” were highly used in tweets before cases and deaths started being reported. But, as the number of cases began rising, their usage started declining.
However, usage of keywords like “stay_home” , “social_distance” started picking up as cases and deaths started rising.
This pattern in frequency of keywords therefore reflects the perception of disease in conjunction with the evolution of the outbreak.
To see the correlation of the individual time series pattern of words with infections and deaths, cross-correlation plot was plotted for each term to identify the lags at which they correlate the most with the targets “infection” and “death”.

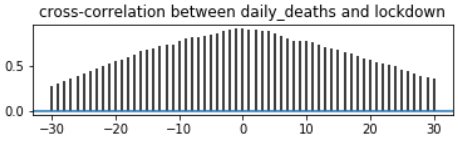
Cross correlation between two time series is obtained by taking sample correlations of the two series at different lags of response variables. If the correlations are significant at higher lags of response, it suggests that lagged effect of predictors on the response is more than contemperous effect. Cross correlation plot can therefore be used to identify the appropriate lags to be used for predictors. The plots below were obtained using “xcorr()” function in matplotlib in python.

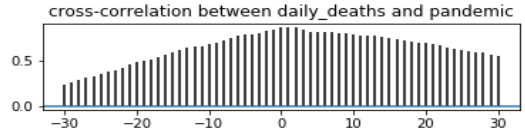
Terms showing similar pattern and range of values were then clubbed together as single term. Altogether 8 such “clubbed_terms” were obtained:

1)Term 1



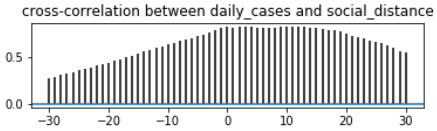


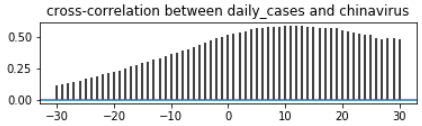


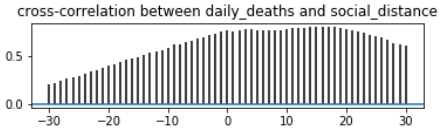


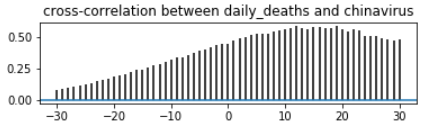
“pandemic” and “lockdown” show similar pattern. The correlation is maximum at lag 0 and decreases as it is shifted forward or backward. Therefore, these terms need not be lagged.

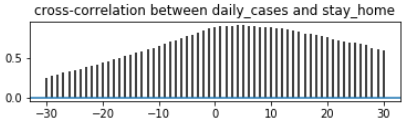
2) Term 2

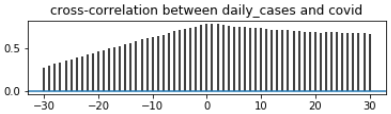


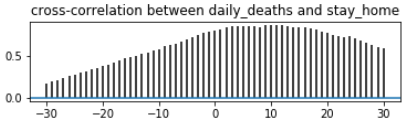


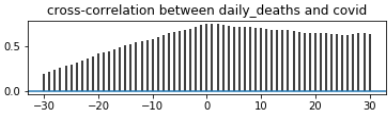




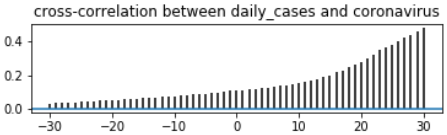


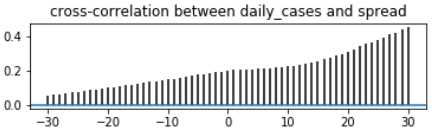


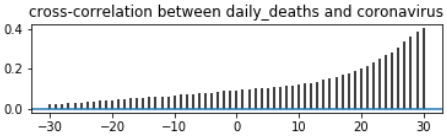


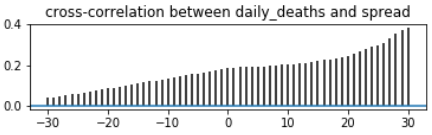


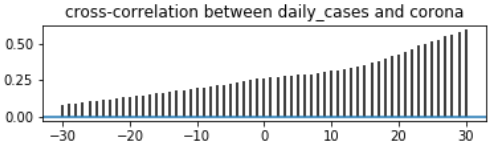
3) Term 3

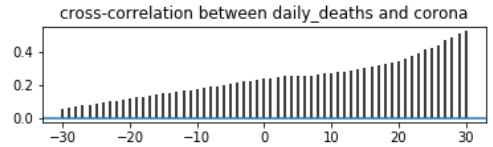








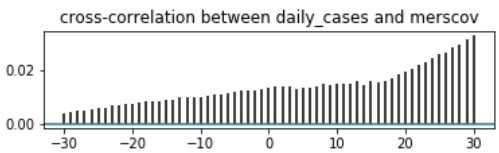


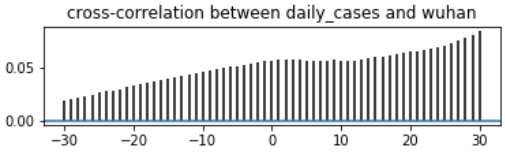


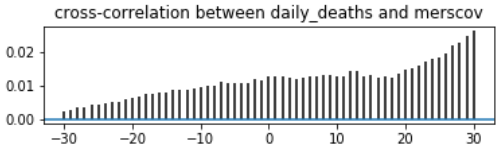
The above terms show higher correlation at higher lags, which means that when time series plots of the above are shifted right, they correlate well with “infection” and “deaths”. The lagged effect of these variables on the response is better than contemporaneous effect.

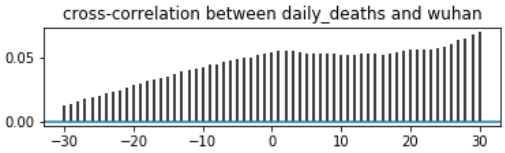
Thus, to predict infection or death at time ‘t’, counts of the above words at time ‘t-v’, (‘v’ is a suitable lag) can be better predictor.

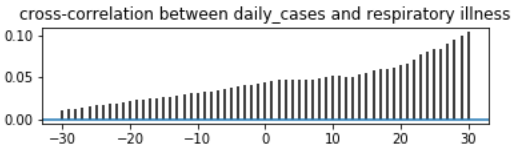
4) Term 4

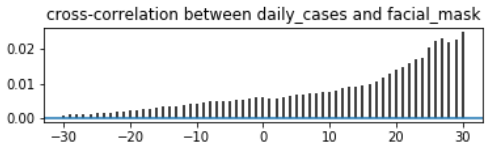


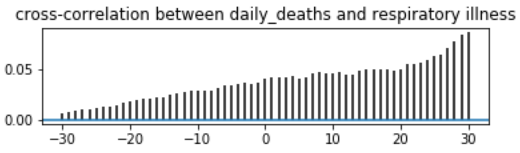


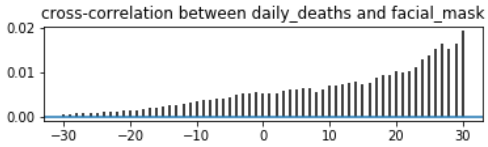




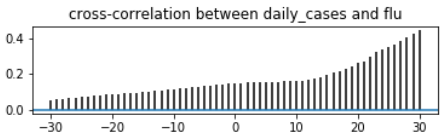


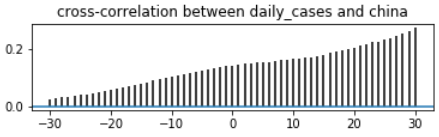


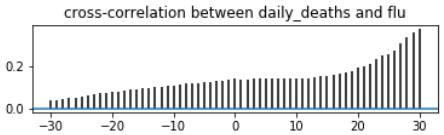


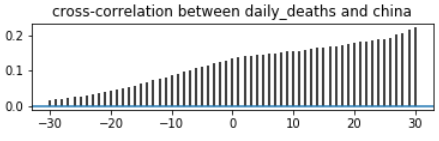


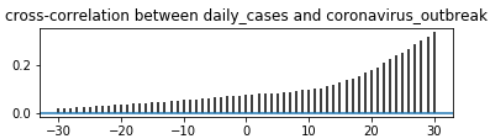
5) Term 5

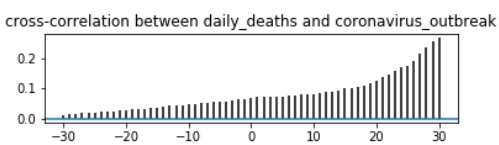




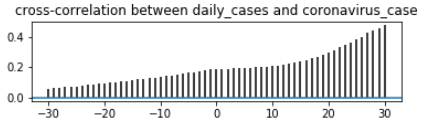


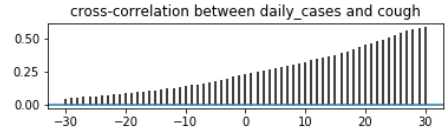


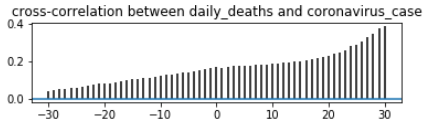


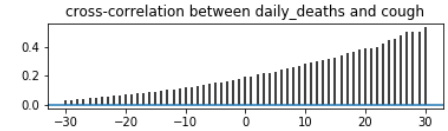


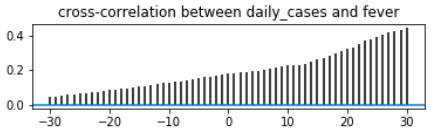
Term 6

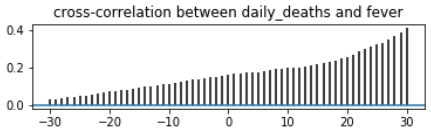




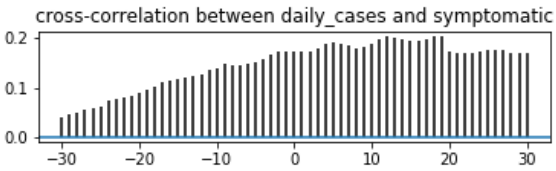


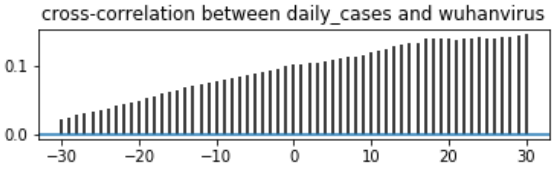


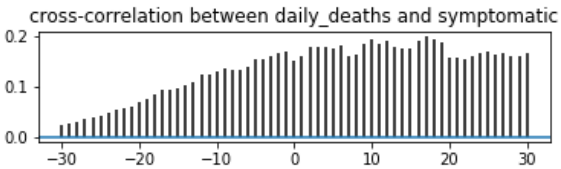


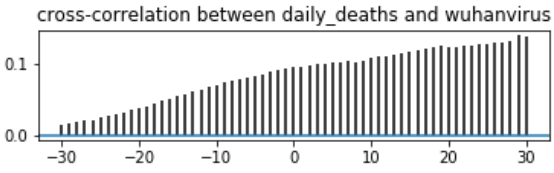


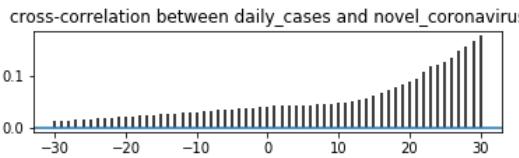
7) Term 7

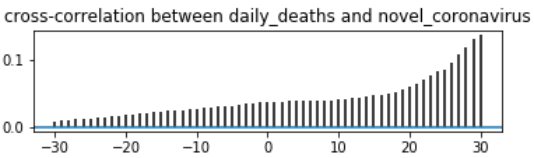




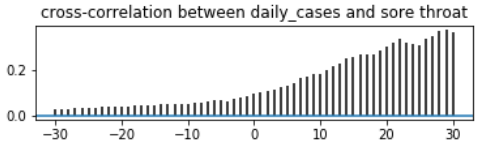


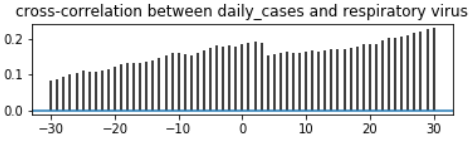


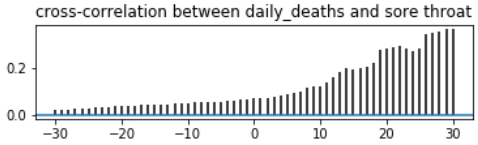


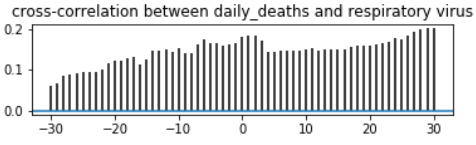


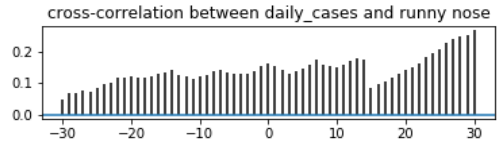
8) Term 8

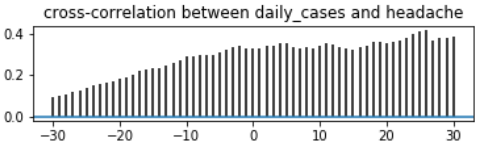

cross-correlation between daily_cases and sore throat
0.2
0.0
−30
−20
−10
0
10
20
30

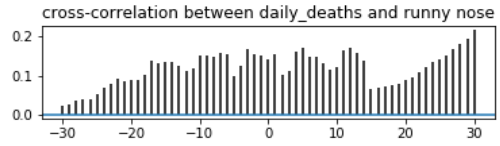

cross-correlation between daily_cases and respiratory virus
0.2
0.1
0.0
−30
−20
−10
0
10
20
30

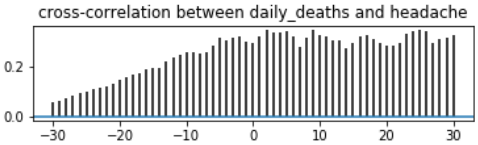

cross-correlation between daily_deaths and sore throat
0.2
0.0
−30
−20
−10
0
10
20
30


cross-correlation between daily_deaths and respiratory virus
0.2
0.1
0.0
−30
−20
−10
0
10
20
30

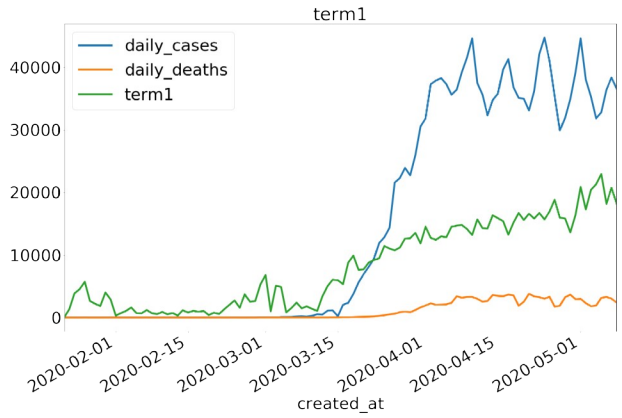

cross-correlation between daily_cases and runny nose
0.2
0.1
0.0
−30
−20
−10
0
10
20
30

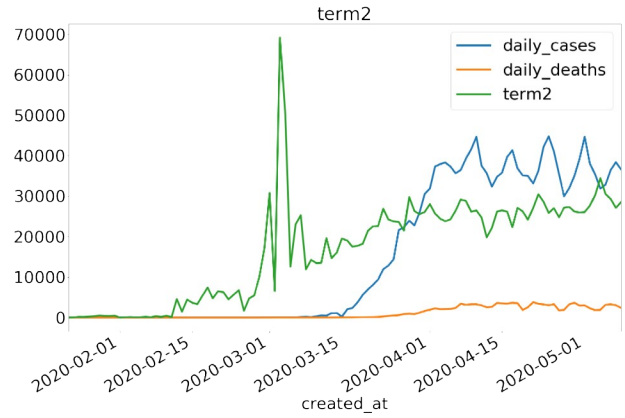

cross-correlation between daily_cases and headache
0.4
0.2
0.0
−30
−20
−10
0
10
20
30

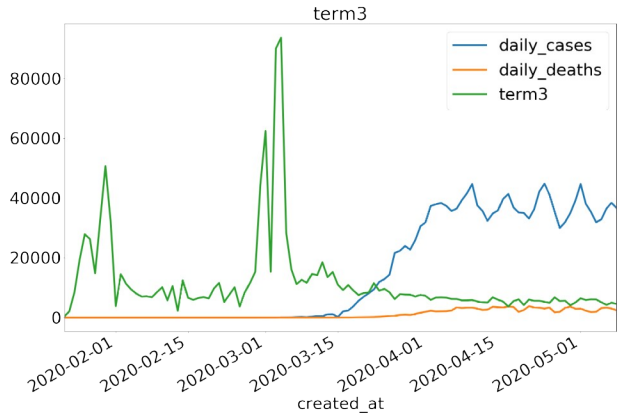

cross-correlation between daily_deaths and runny nose
0.2
0.1
0.0
−30
−20
−10
0
10
20
30

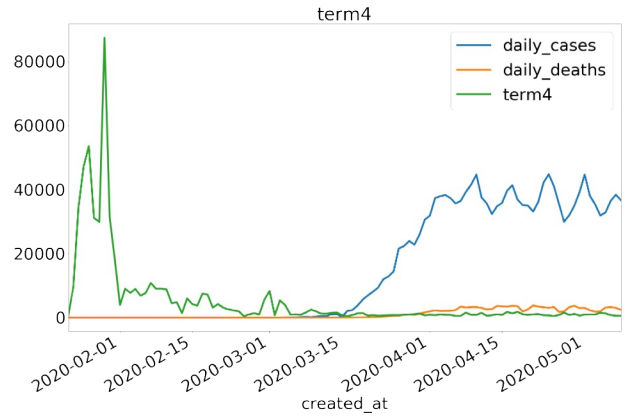

cross-correlation between daily_deaths and headache
0.2
0.0
−30
−20
−10
0
10
20
30

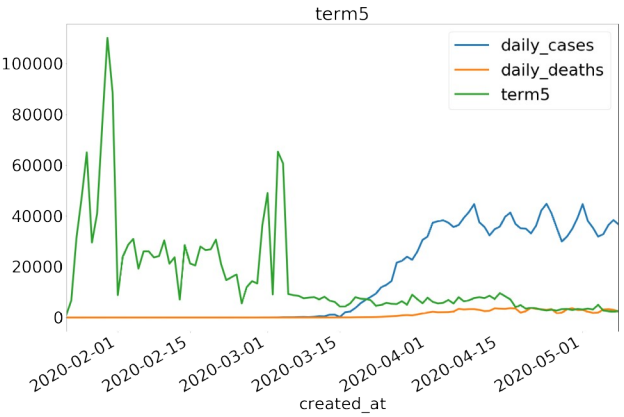
Time series plots of clubbed terms

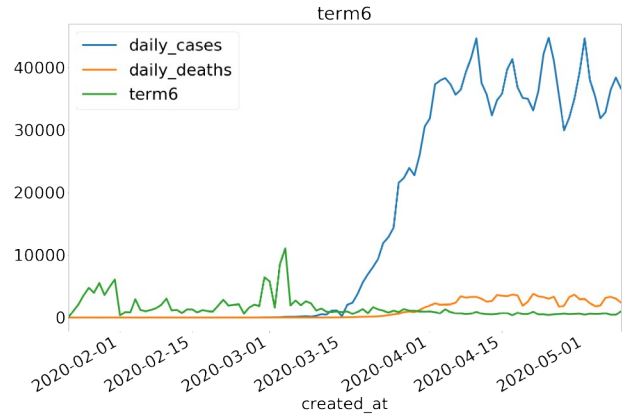


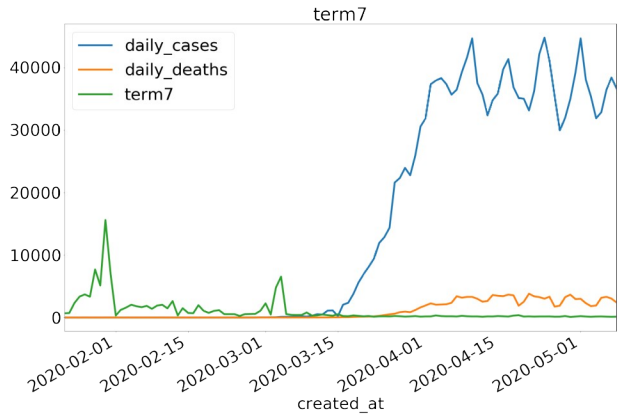


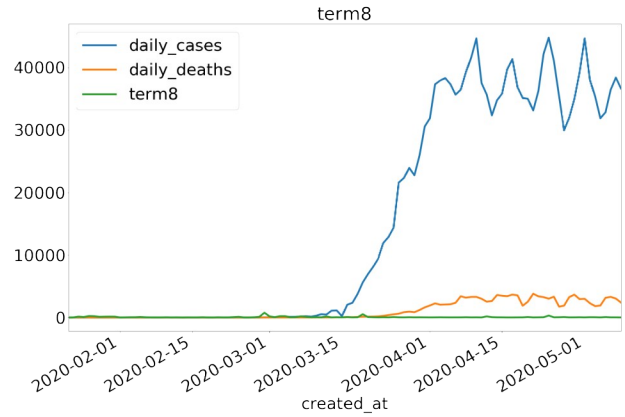












Prediction Model results

Prediction model was obtained using the whole tweet data from January-May.
First 100 observations were reserved for training and the remaining data of seven days were used as test data.
The term counts of 20 most frequently occuring terms were used predictors to model future covid19 cases. To test efficacy of models, Pearson's correlation(r) was used.

Taking cue from existing research papers on modelling outbreaks using SNS data like twitter, first a linear model was fit. The linear model has been shown to be quite successful in predicting Influenza outbreaks using counts of "outbreak" and "symptoms" in tweets.

An Elasic Net model was first fit with parameters **alpha** = 1 and **l1_ratio** = 0.4. Different values of the hyperparameters were used for which the results were almost same.
The results are -,
training, r= 0.9659843042760209
test, r= -0.5348689325140943
On training set, the fitted values have high correlation with observed values, but the model performs poorly on test set with the predicted values having negative correlation.

Next, ARIMA model was considered. The ACF and PACF plots suggested non-stationarity series. Thus, appropriate differencing was done after which the ARMA model AR=4, MA=0 with lowest AIC was chosen.
Training, r=0.9941087909171533
test, r= 0.10874414589085242
The fitted values have very high correlation with observed values. On the test set model performs better than ElasticNet, but the correlation with actual values is still low.

Now, using the earlier obtained term counts as exogenous features, an ARIMAX model was fit.
The results are – training, r= 0.9957496179583817
test ,r = 0.6674229546801306
The model shows marked improvement over the previous ARIMA model on the test set, thus proving the effectiveness of "term counts" as regressors.

ARIMAX with clubbed_terms : The clubbed terms were obtained as shown previously by binning all words with similar cross correlation plot as a single term as summing their frequencies.
Training, r=0.9946939071902215 ; test, r= 0.6179583924305259
The clubbed terms are less effective in predicting future values than previous model.

ARIMAX model with lagged features : Exogenous variable with the appropriate lags obtained from cross-correlation plots was used. The results are - training, r = 0.9956100430890172
test, r = 0.7995231113079285
Using lagged variables as regressors offers much more improvement than ARIMAX with contemporaneous features.

ARIMAX model with lagged symptom features: In this case, only lagged counts of symptom terms were used as regressors. On the training set, r = 0.9942564155111253 ; test , r=0.39159130. This was observed to be slightly better than using symptom terms without lags whose 'r' on test set was obtained as 0.24182964537249.

From the above results, ARIMAX model with lagged word counts seem to be appropriate prediction models. However, it comes with certain disadvantages like interpretability and they tend to converge to mean in the long run. Over time new features could be discovered and used in the model. LSTM models could not be used as neural networks require huge amount of data to train.