Statistical/Content Analysis of Coronavirus tweets

**Abstract:**

The detail analysis of Tweets is provided in this paper. The statistical properties of tweets are analyzed in this work and its follwes zips law. The regionwise analysis of the tweets generested are also explored in this work. The frequently discussed terms are identified in the different lockdown phases in India and the USA.

Task: SSS

**Introduction**:

Dataset description:

Task ; SSS

**Literature work**

**(Aboli, Aditya,Shrushti)**

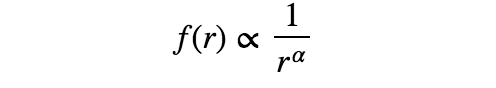
**Methodology:**

The methodology is divided in the following section.

**(Aboli, Aditya,Shrushti)**

1. **Statistical properties of terms in Tweets**

ZIPF ANALYSIS Zip’s Law or Power Law states that frequency of a word is inversely proportional to rank of the word. One thing to note is that the actual observations in most cases does not strictly follow Zipf’s distribution, but rather follow a trend of “near-Zipfian” distribution.

Zipf’s Law can be written as follows: the rth most frequent word has a frequency f(r) that scales according to 

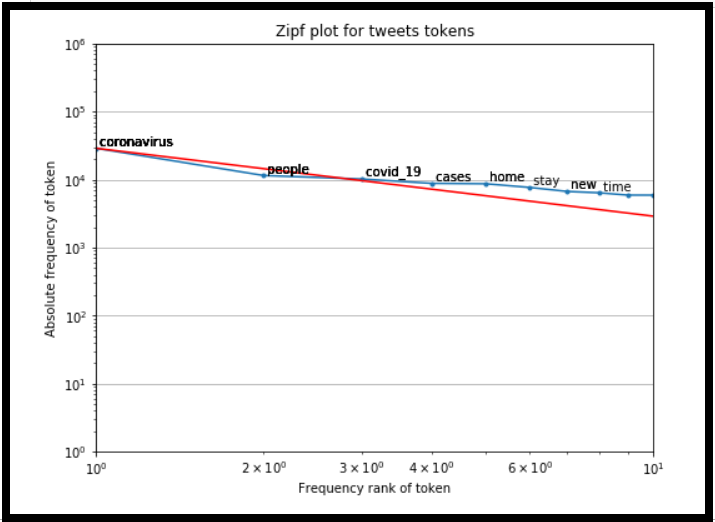
for

Let’s see how the tweet tokens and their frequencies look like on a plot.

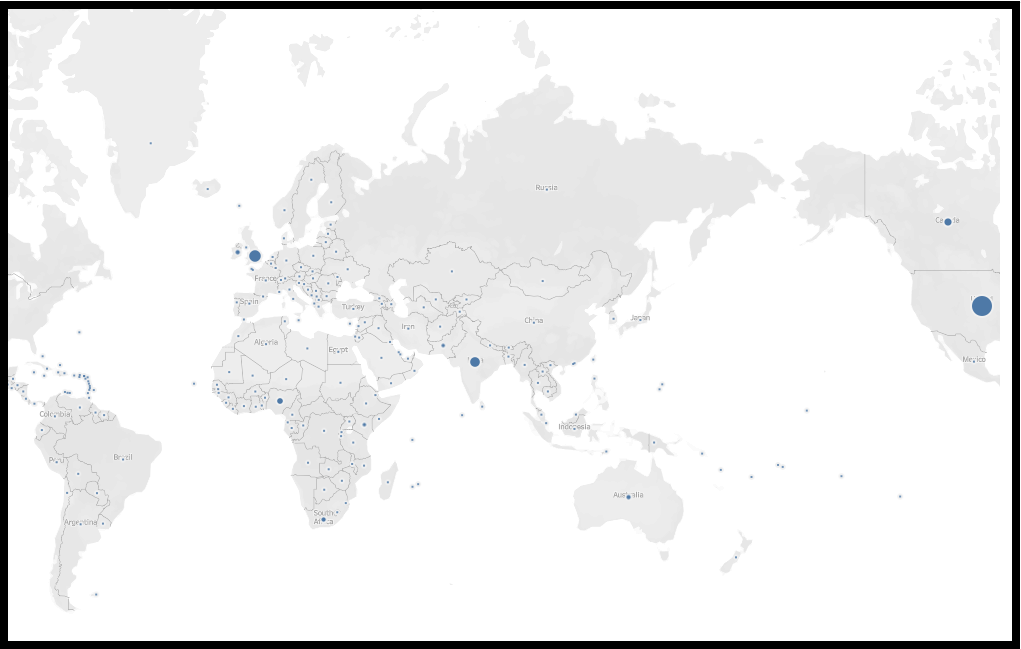
We will plot this on a log-log graph, with X-axis being log(rank), Y-axis being log(frequency). By plotting on a log-log scale the result will yield roughly linear line on the graph.

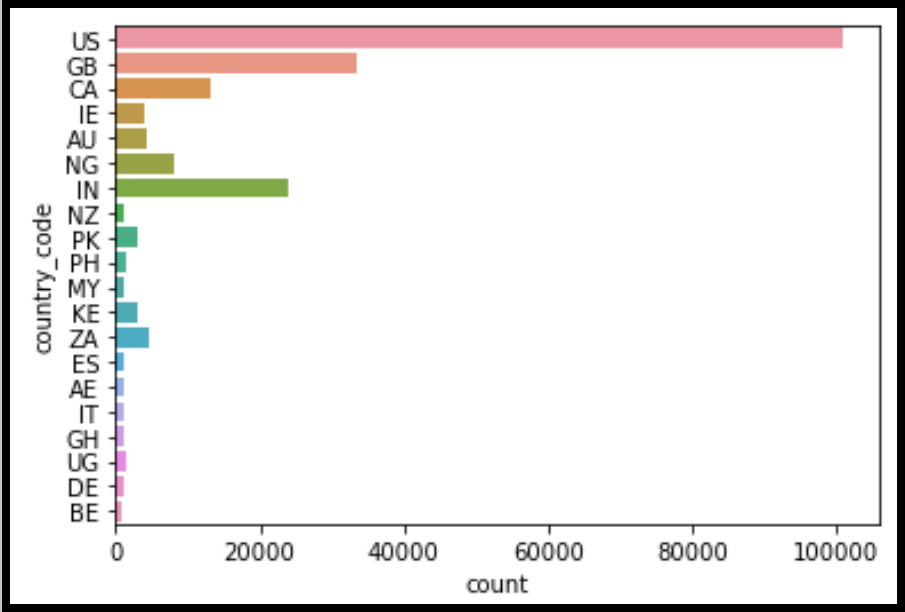
Again we see a roughly linear curve, but deviating below the expected line on higher ranked words, and at the lower ranks we see the actual observation line lies above the expected linear line.

We proved that even the tweet tokens follow “near-Zipfian” distribution.



1. **Region wise analysis**





The above bar graph and the maps shows that from a distribution of randomly sampled tweetst throughout various countries, USA, Great Britain, India and Canada have the most tweets.

1. TOPIC MODELLING

Finding the most common words from NMF and Bag of Words model, we shortlisted these 294 words/ phrases that occurred the most, excluding the common stop words, pronouns.

['people', 'cases', 'stay', 'coronavirus', 'covid\_19', 'deaths',

'home', 'covid', 'stayathomeandstaysafedied', 'confirmed', 'safe',

'death', 'borisjohnson', 'dying', 'healthy', 'coronaupdate',

'toll', 'corona', 'die', 'number', 'latest',

'lockdowneffecttested', 'reported', 'save', 'news', 'boris',

'virus', 'positive', 'staying', 'authorities', 'covid2019', 'need',

'tweets', 'coronavirustruthpositive', 'recovered', 'lives',

'coronalockdownthink', 'nigeria', 'covid19pandemicvirus', 'april',

'stayhomesavelivesmasks', 'socialdistancinginfected', 'today',

'wash', 'canadian', 'hope', 'government', 'million', 'strong',

'how', 'why', 'hands', 'spain', 'whencoronavirusisoverthere',

'state', 'working', 'what', 'stayathome', 'discharged', 'outbreak',

'recovery', 'stop', 'worldwide', 'inside', 'understand', 'update',

'work', 'china', 'thelockdown', 'county', 'protect', 'nhs',

'cancelstudentdebtkeep', 'support', 'studentdebtstimuluspetition',

'family', 'together', 'stimulate', 'debt', 'sign', 'community',

'coronavirusoutbreaksupport', 'student', 'ukchange', 'food',

'fight', 'package', 'uk', 'patients', 'must', 'urge', 'ppe',

'live', 'friends', 'frontline', 'test', 'help', 'economy', 'great',

'stand', 'next', 'govt', 'staff', 'times', 'provide', 'others',

'coronavirusupdatescommunity', 'needed', 'team', 'tested', 'beat',

'via', 'first', 'continue', 'speakerpelosi', 'youtube', 'many',

'disease', 'medical', 'died', 'win', 'moneyforthepeoplein',

'change', 'updates', 'overcome', 'response', 'alerted', 'crisis',

'trump', 'new', 'realdonaldtrumpto', 'york', 'who', 'city',

'spread', 'president', 'nyc', 'god', 'normal', 'feel', 'symptoms',

'americans', 'newyork', 'app', 'potus', 'jersey', 'self', 'join',

'america', 'daily', 'reports', 'slow', 'donald', 'best', 'total',

'risk', 'states', 'brooklyn', 'download', 'newyorkcity', 'allah',

'cnn', 'yorkers', 'identify', 'media', 'zealand', 'reporting',

'funding', 'record', 'orleans', 'press', 'country', 'manhattan',

'surpasses', 'hoax', 'isolation', 'pray', 'trumppressconfreal',

'prevent', 'lord', 'lockdown', 'stayhome', 'health',

'coronaviruspandemicstayathome', 'staysafe', 'care', 'extended',

'workers', 'india', 'real', 'socialdistancinghealthcare',

'coronavirusoutbreakapril', 'stayhomesavelivespublic',

'covid19pandemictill', 'stayhealthy', 'stayhomesavelivesdays',

'flattenthecurvedoctors', 'washyourhands', 'hospital',

'coronavirususadue', 'savelives', 'nurses', 'weeks', 'love',

'coronavirusupdatecorona', 'life', 'quarantineactivitiesextend',

'sad', 'easter', 'mental', 'coronavirusupdates2020', 'staystrong',

'masks', 'london', 'besafe', 'living', 'lockdown2020', 'essential',

'coronalockdownpolice', 'needs', 'everyone', 'end', 'minister',

'coronavirustruthday', 'world', 'time', 'social', 'distancing',

'socialdistancingone', 'difficult', 'distance', 'practice',

'physical', 'countries', 'would', 'long', 'practicing', 'week',

'chinese', 'maintain', 'morning', 'war', 'usa', 'happy',

'coronavirusoutbreakthink', 'last', 'wuhan', 'measures', 'night',

'rules', 'rest', 'spending', 'guidelines', 'beautiful', 'global',

'right', 'mask', 'make', 'face', 'good', 'important',

'stayhomestaysafeplease', 'quarantine', 'pandemic', 'narendramodi',

'quarantinelifethank', 'share', 'socialdistancingknow', 'days',

'leaders', 'indiafightscoronasomeone', 'oneworld', 'pmoindia',

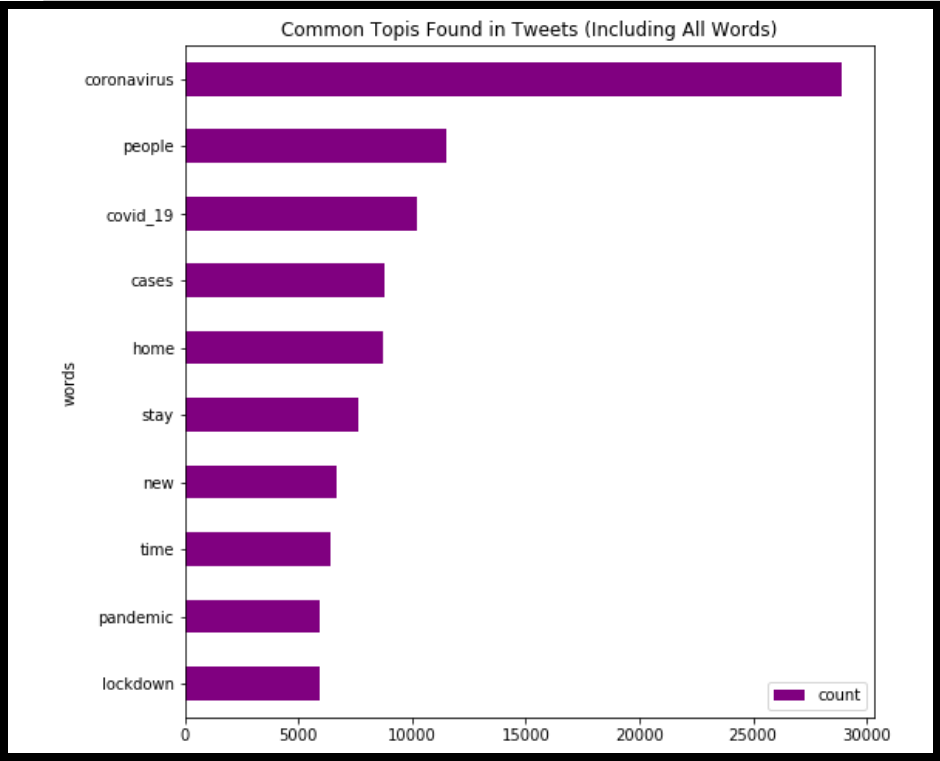
'quarantineandchillre', 'plan', 'defeat', 'pm', 'donate', 'heroes',

'follow', 'doctors', 'retweet', 'california', 'read', 'tiktok',

'bored', 'service', 'amid', 'dear', 'realdonaldtrumpin',

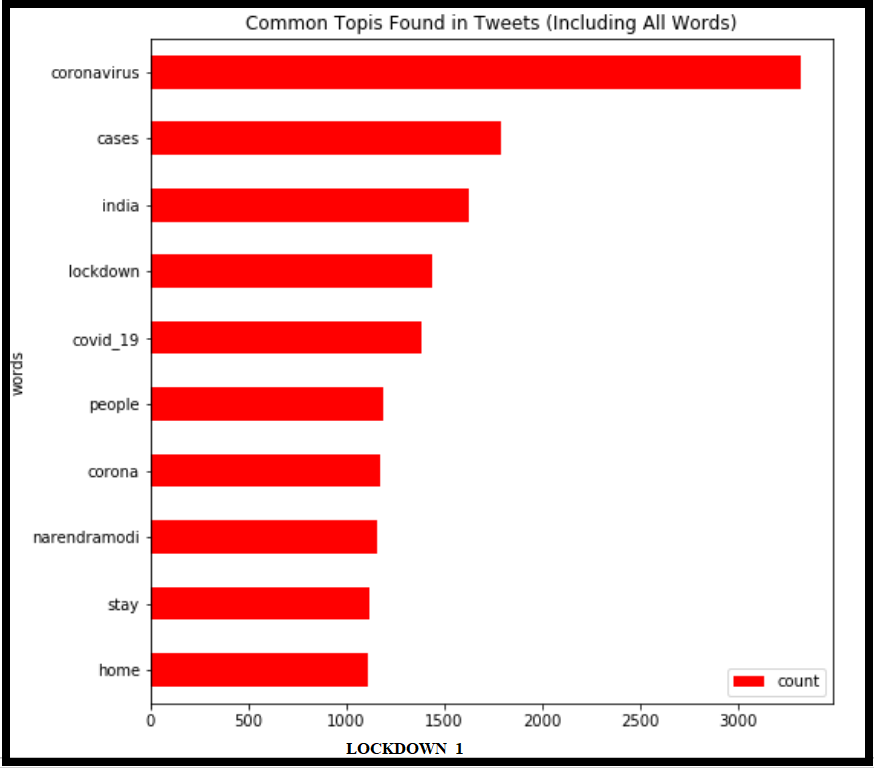
'stayhomeindia', 'stayhomesaveliveswhen', 'stayhomesavelivesmake',

'Music']

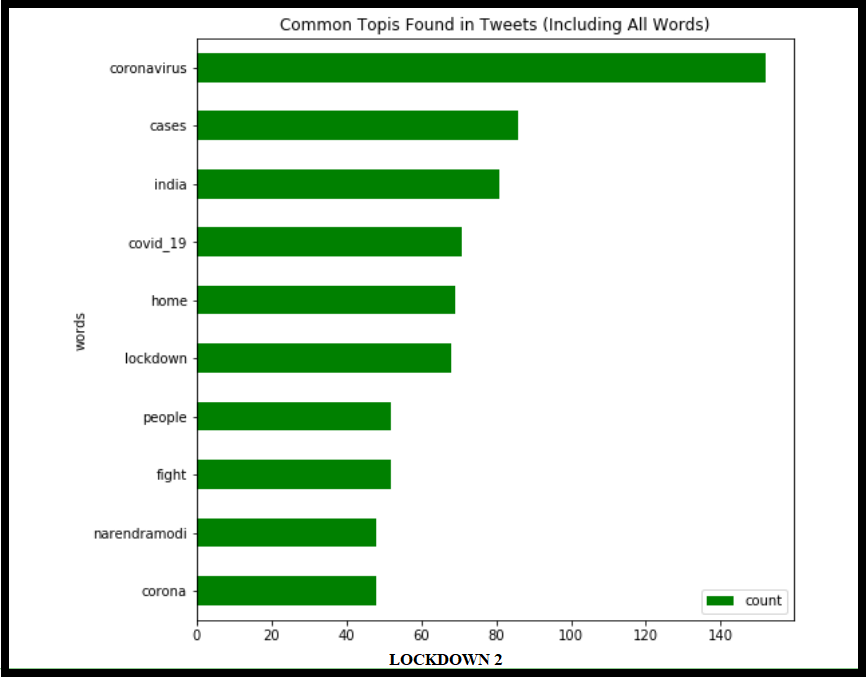


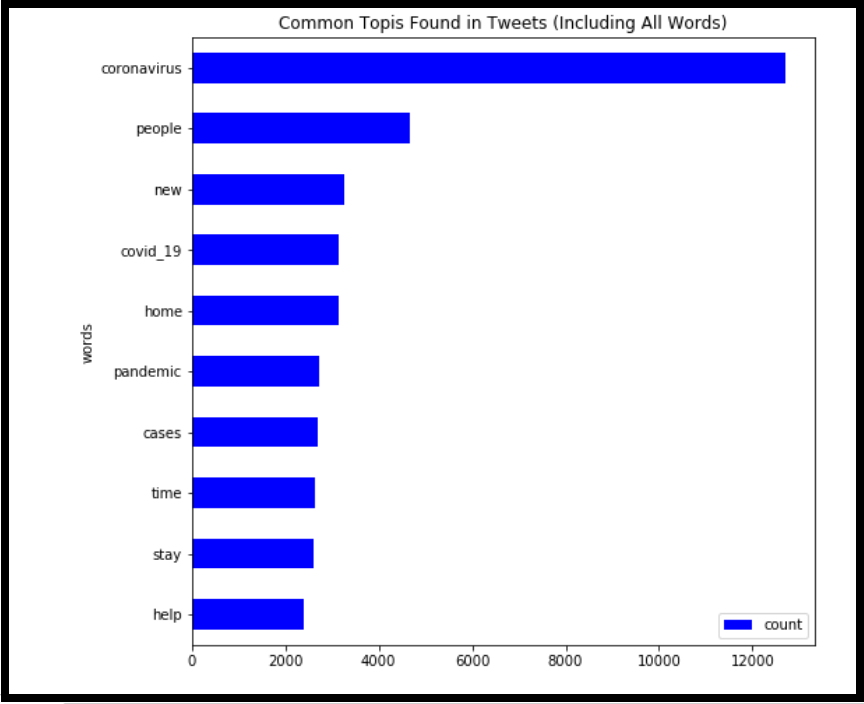
The above graph shows the frequency of occurrence of topics in all the tweets. Coronavirus is the most talked about term .

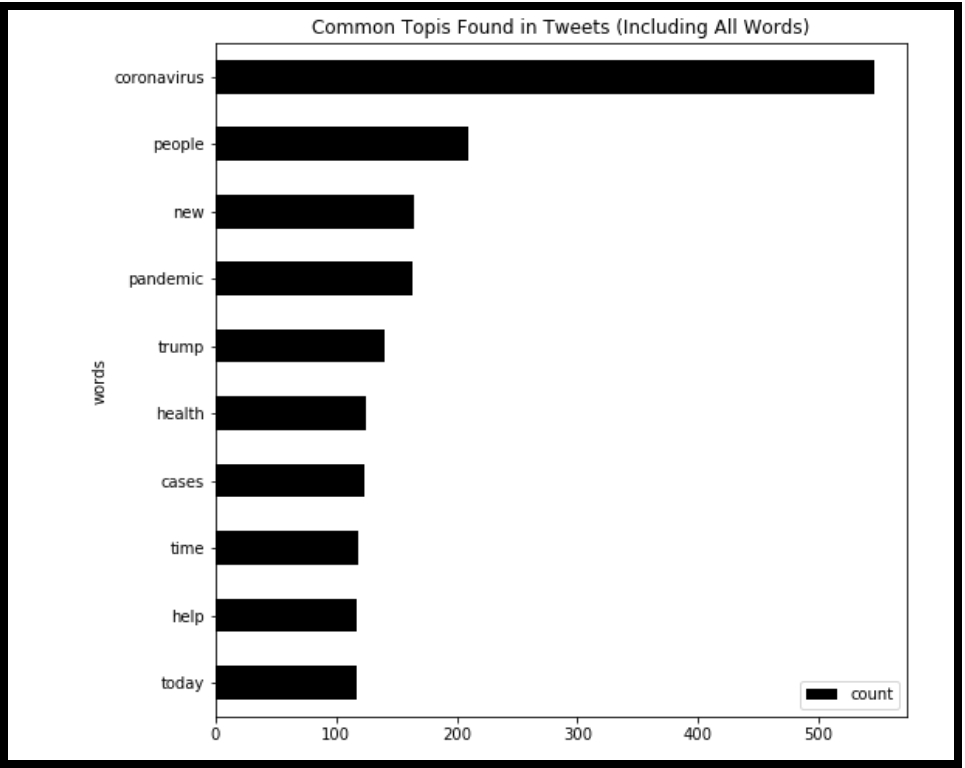
INDIA LOCKDOWN 1 [(25 March – 14 April)](https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdown_in_India#Phase_1_(25_March_%E2%80%93_14_April))

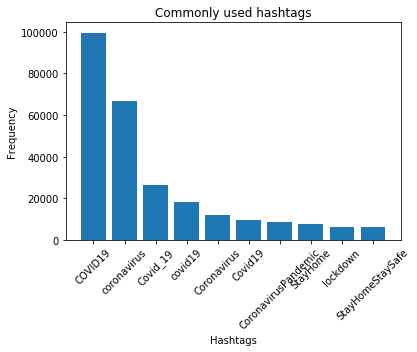


INDIA LOCKDOWN 2[(15 April – 3 May)](https://en.wikipedia.org/wiki/COVID-19_pandemic_lockdown_in_India#Phase_2_(15_April_%E2%80%93_3_May))



USA LOCKDOWN 1

USA LOCKDOWN 2



The above bar graph shows the frequency of each hashtag in the entire tweets dataset. It can be seen that #COVID19, #coronavirus and #Covid\_19 are the most common.

**Analysis / Findings:**

Perform the comparison of tweets from lockdown 1.0 to 4.0

**Conclusion and future scope:**

**(SSS)**

**References:**