IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

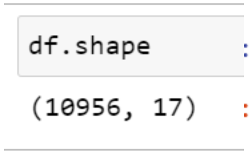
Email: ckoppara@syr.edu

Corpus Statistics and Python Programming

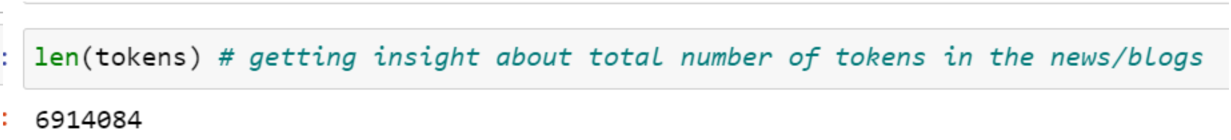
DATA

Author of the data set is Ran Geva. It is a free dataset available in the portal https://ieee-dataport.org/authors/ran-geva. It has four months of data from **Dec 2019 till March 2020**. This is the period when virus was first detected in Wuhan, China and has set off a global pandemic. So, we can get to analyze initial apprehension among people from different parts of the world.

Dataset has 5.2M posts from news and the blogs about corona virus but we are considering only a subset of this dataset for the analysis. Dataset is in JSON format with each json object referring to a news message/blog. JSON data is imported into python application as a data frame with each JSON object treated as a row and fields of the JSON object are treated as columns. In total, the data frame has **10,956 rows** with **17 columns** and there are **6,914,084** tokens in the corpus.



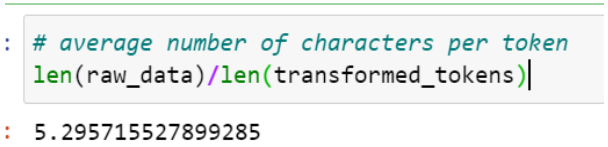
```
df.shape

(10956, 17)
```



```
: len(tokens) # getting insight about total number of tokens in the news/blogs

: 6914084
```

From initial observation, below information can be found in the data:

- Average length of each token in the corpus is around **5 characters**



```
: # average number of characters per token
  len(raw_data)/len(transformed_tokens)

: 5.295715527899285
```

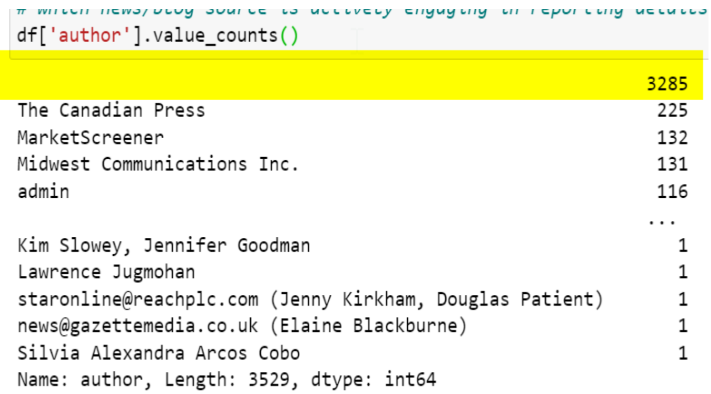
IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

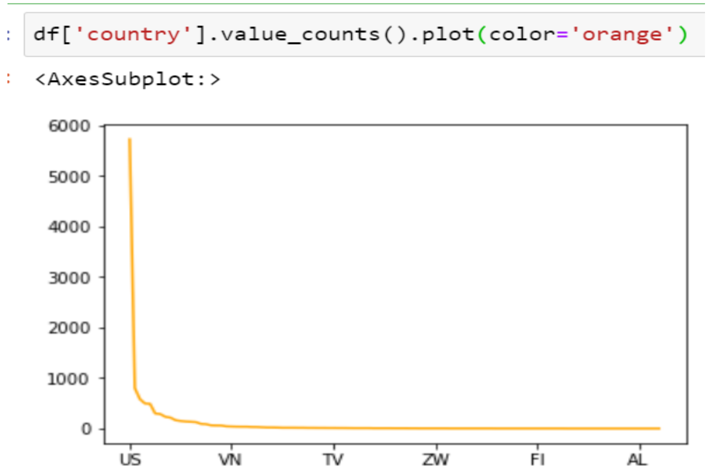
Email: ckoppara@syr.edu

- Most of the posts have **unknown** data source



```
# which news/blog source is actively engaging in reporting details
df['author'].value_counts()

                                                          3285
The Canadian Press                                         225
MarketScreener                                             132
Midwest Communications Inc.                                131
admin                                                      116
                                                          ...
Kim Slowey, Jennifer Goodman                                 1
Lawrence Jugmohan                                            1
staronline@reachplc.com (Jenny Kirkham, Douglas Patient)     1
news@gazettemedia.co.uk (Elaine Blackburne)                  1
Silvia Alexandra Arcos Cobo                                  1
Name: author, Length: 3529, dtype: int64
```

- Most of the data in the corpus belong to **USA**



```
df['country'].value_counts().plot(color='orange')

<AxesSubplot:>
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu

- Popular Facebook posts with highest likes, shares, and comments

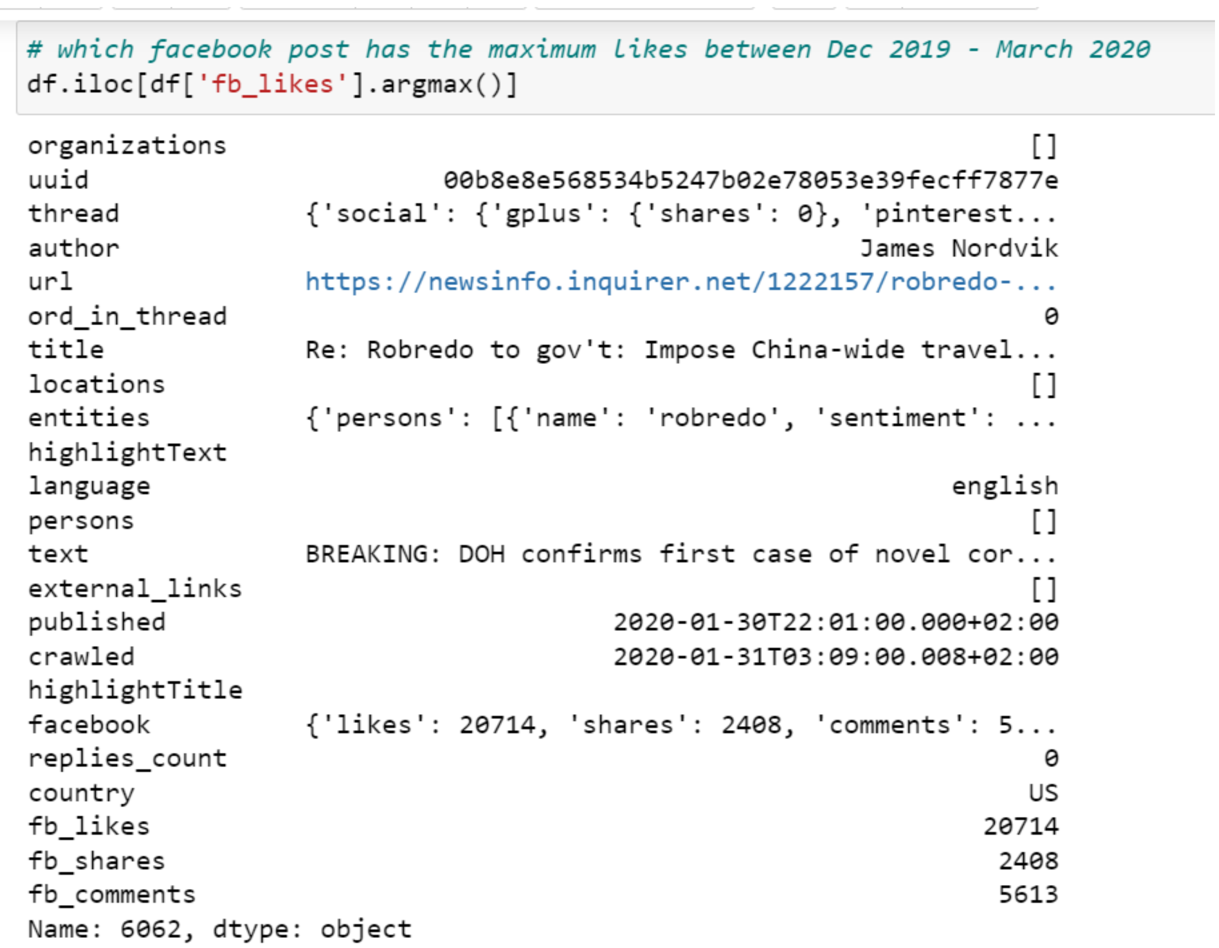
Faceboook	**Likes**	**Shares**	**Comments**
Count	20714	6324	6749
Country	US	Philippines	Great Britain
Title of the Post	Re: Robredo to gov't: Impose China-wide travel ban now | Inquirer News	Employees of Negros Oriental hotel, resort where Chinese with nCoV stayed now on quarantine	Brexit: Boris Johnson to hail 'dawn of a new era' - BBC News

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu



```
# which facebook post has the maximum likes between Dec 2019 - March 2020
df.iloc[df['fb_likes'].argmax()]
```

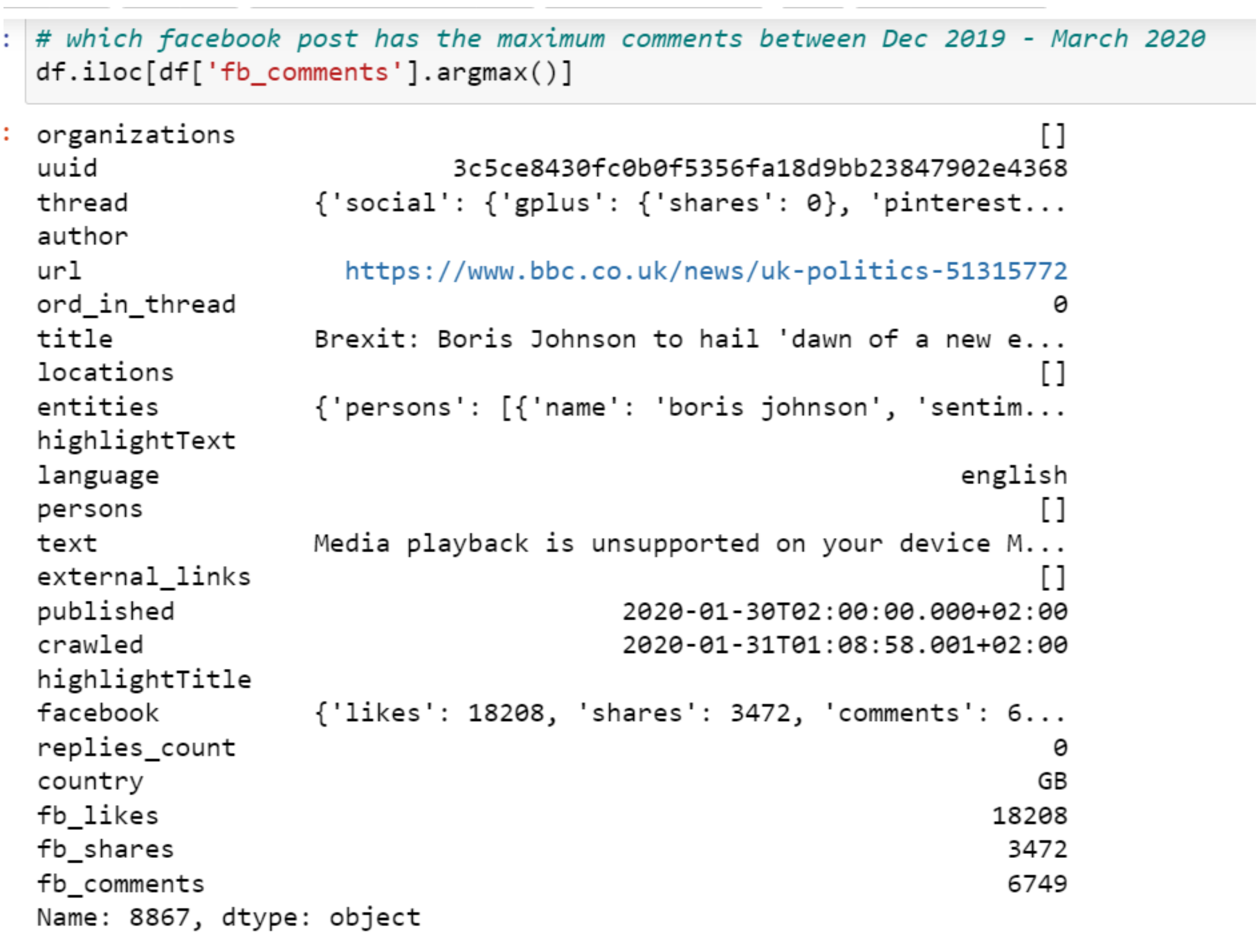
```
organizations                                                       []
uuid                          00b8e8e568534b5247b02e78053e39fecff7877e
thread               {'social': {'gplus': {'shares': 0}, 'pinterest...
author                                                   James Nordvik
url                  https://newsinfo.inquirer.net/1222157/robredo-...
ord_in_thread                                                        0
title                Re: Robredo to gov't: Impose China-wide travel...
locations                                                           []
entities             {'persons': [{'name': 'robredo', 'sentiment': ...
highlightText
language                                                       english
persons                                                             []
text                 BREAKING: DOH confirms first case of novel cor...
external_links                                                      []
published                                2020-01-30T22:01:00.000+02:00
crawled                                  2020-01-31T03:09:00.008+02:00
highlightTitle
facebook             {'likes': 20714, 'shares': 2408, 'comments': 5...
replies_count                                                        0
country                                                             US
fb_likes                                                         20714
fb_shares                                                         2408
fb_comments                                                       5613
Name: 6062, dtype: object
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu



```
# which facebook post has the maximum comments between Dec 2019 - March 2020
df.iloc[df['fb_comments'].argmax()]
```

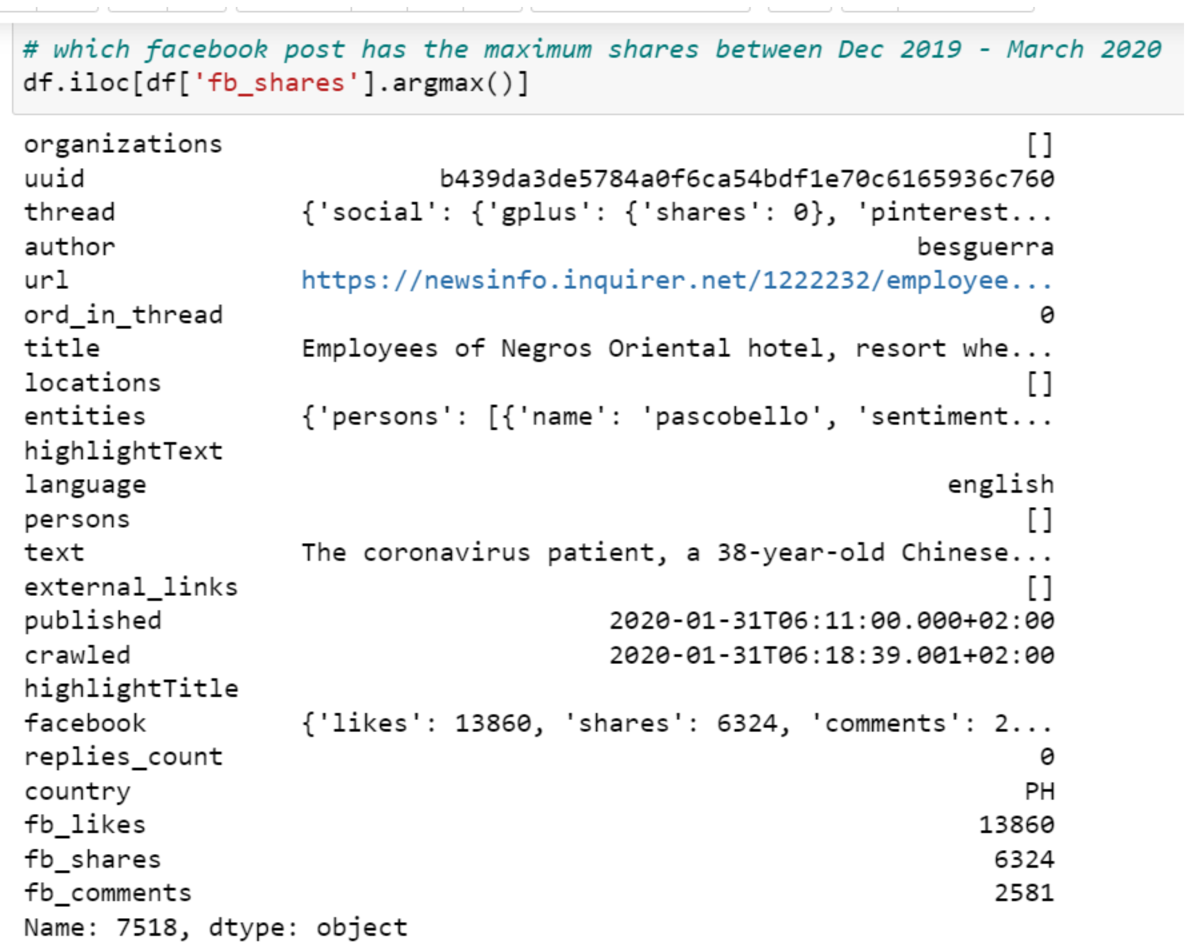
```
organizations                                                    []
uuid                       3c5ce8430fc0b0f5356fa18d9bb23847902e4368
thread            {'social': {'gplus': {'shares': 0}, 'pinterest...
author
url                 https://www.bbc.co.uk/news/uk-politics-51315772
ord_in_thread                                                     0
title             Brexit: Boris Johnson to hail 'dawn of a new e...
locations                                                        []
entities          {'persons': [{'name': 'boris johnson', 'sentim...
highlightText
language                                                    english
persons                                                          []
text              Media playback is unsupported on your device M...
external_links                                                   []
published                             2020-01-30T02:00:00.000+02:00
crawled                               2020-01-31T01:08:58.001+02:00
highlightTitle
facebook          {'likes': 18208, 'shares': 3472, 'comments': 6...
replies_count                                                     0
country                                                          GB
fb_likes                                                      18208
fb_shares                                                      3472
fb_comments                                                    6749
Name: 8867, dtype: object
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu



```python
# which facebook post has the maximum shares between Dec 2019 - March 2020
df.iloc[df['fb_shares'].argmax()]
```

```
organizations                                                    []
uuid                       b439da3de5784a0f6ca54bdf1e70c6165936c760
thread                  {'social': {'gplus': {'shares': 0}, 'pinterest...
author                                                    besguerra
url                     https://newsinfo.inquirer.net/1222232/employee...
ord_in_thread                                                     0
title                   Employees of Negros Oriental hotel, resort whe...
locations                                                        []
entities                {'persons': [{'name': 'pascobello', 'sentiment...
highlightText
language                                                    english
persons                                                          []
text                    The coronavirus patient, a 38-year-old Chinese...
external_links                                                   []
published                             2020-01-31T06:11:00.000+02:00
crawled                               2020-01-31T06:18:39.001+02:00
highlightTitle
facebook                {'likes': 13860, 'shares': 6324, 'comments': 2...
replies_count                                                     0
country                                                          PH
fb_likes                                                      13860
fb_shares                                                      6324
fb_comments                                                    2581
Name: 7518, dtype: object
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

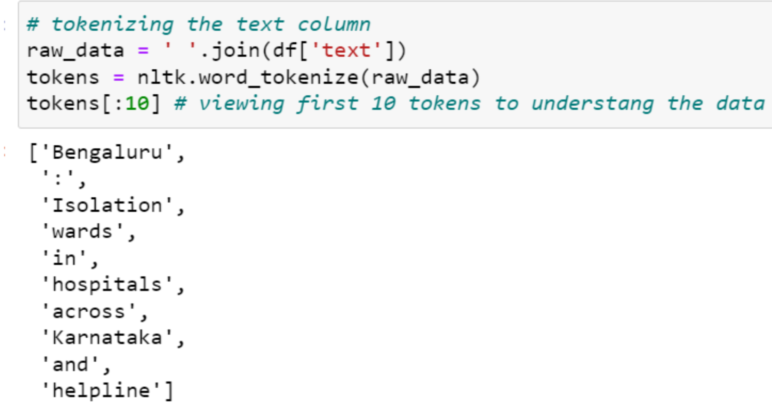
SUID: 326926205

Email: ckoppara@syr.edu

DATA PRE-PROCESSING

In data pre-processing, performed below tasks:

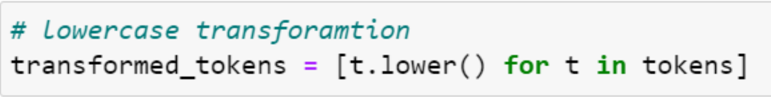
1. **Tokenization**: In all the news articles and blogs, JSON field **'text'** has all the information about the virus**. 'text'** is very crucial for analysis. Raw data in the **'text'** field across all the news and blogs is tokenized.



```
# tokenizing the text column
raw_data = ' '.join(df['text'])
tokens = nltk.word_tokenize(raw_data)
tokens[:10] # viewing first 10 tokens to understang the data
```

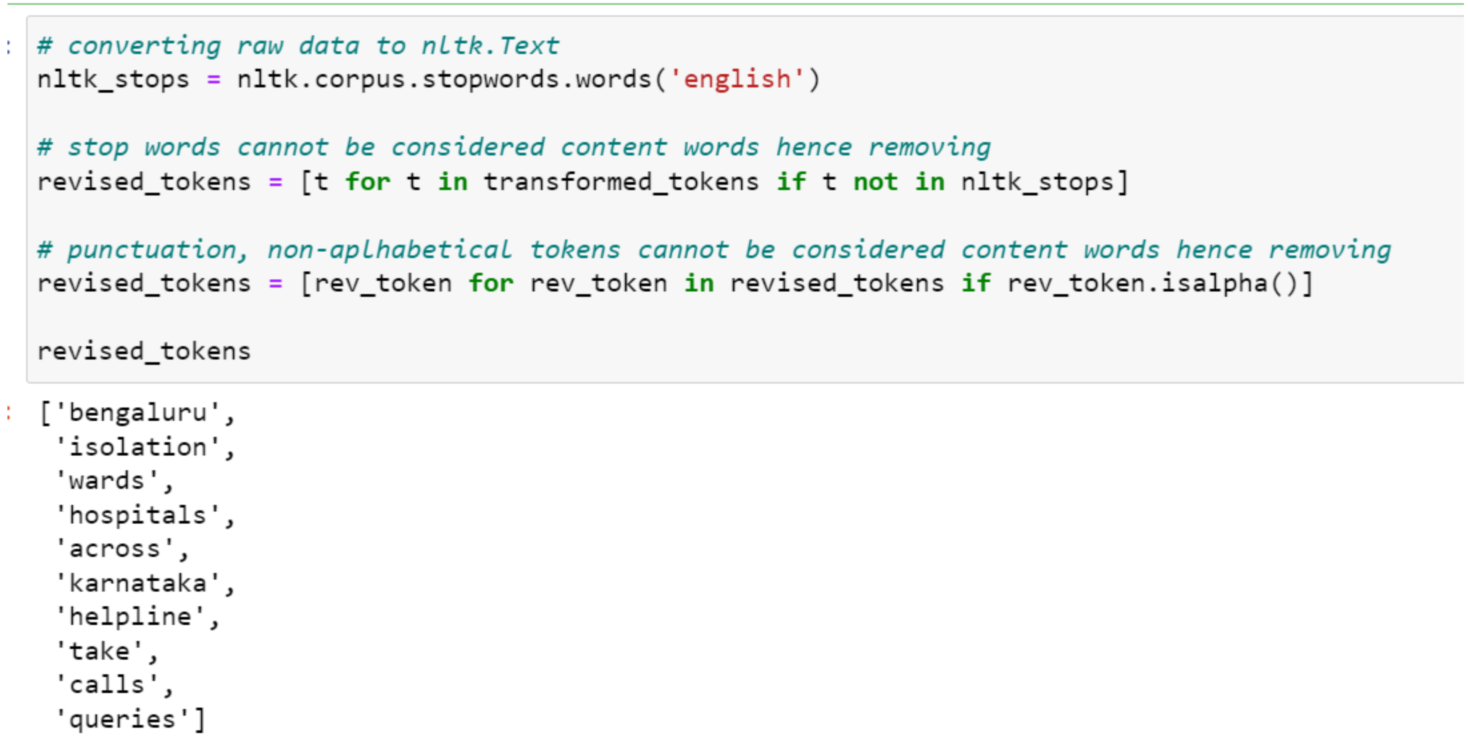
```
['Bengaluru',
 ':',
 'Isolation',
 'wards',
 'in',
 'hospitals',
 'across',
 'Karnataka',
 'and',
 'helpline']
```

2. **Lowercase Transformation**: Tokens such as 'corona' and 'Corona' are being considered as different words. This will create inconsistency while performing corpus statistics. hence, all the tokens are transformed into lowercase



```
# Lowercase transforamtion
transformed_tokens = [t.lower() for t in tokens]
```

3. **Stop words and punctuation removal**: Made use of stop words provided by python NLTK library to remove all the stop words from the corpus**.** Stop words does not add much value in corpus statistics though it might be required in understanding the context around a particular token. Also, removed all the punctuations from the tokenized words. It does not contribute much to the corpus statistics.



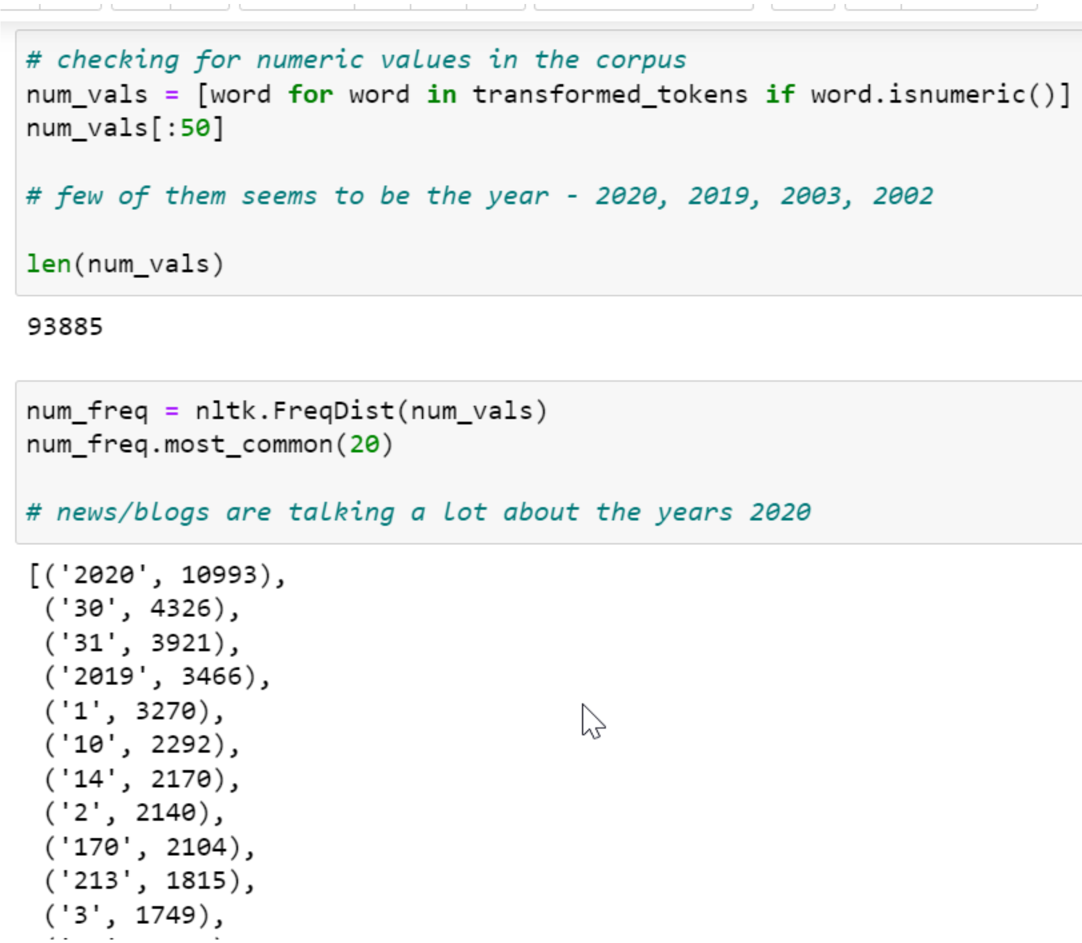
```
# converting raw data to nltk.Text
nltk_stops = nltk.corpus.stopwords.words('english')

# stop words cannot be considered content words hence removing
revised_tokens = [t for t in transformed_tokens if t not in nltk_stops]

# punctuation, non-aplhabetical tokens cannot be considered content words hence removing
revised_tokens = [rev_token for rev_token in revised_tokens if rev_token.isalpha()]

revised_tokens
```

```
['bengaluru',
 'isolation',
 'wards',
 'hospitals',
 'across',
 'karnataka',
 'helpline',
 'take',
 'calls',
 'queries']
```

4. **Numeric data extraction:** Extracted all the numeric tokens present in the corpus to understand

    a. number of covid cases reported by the news/in the blogs

b. around which year/month, there was more media-hype



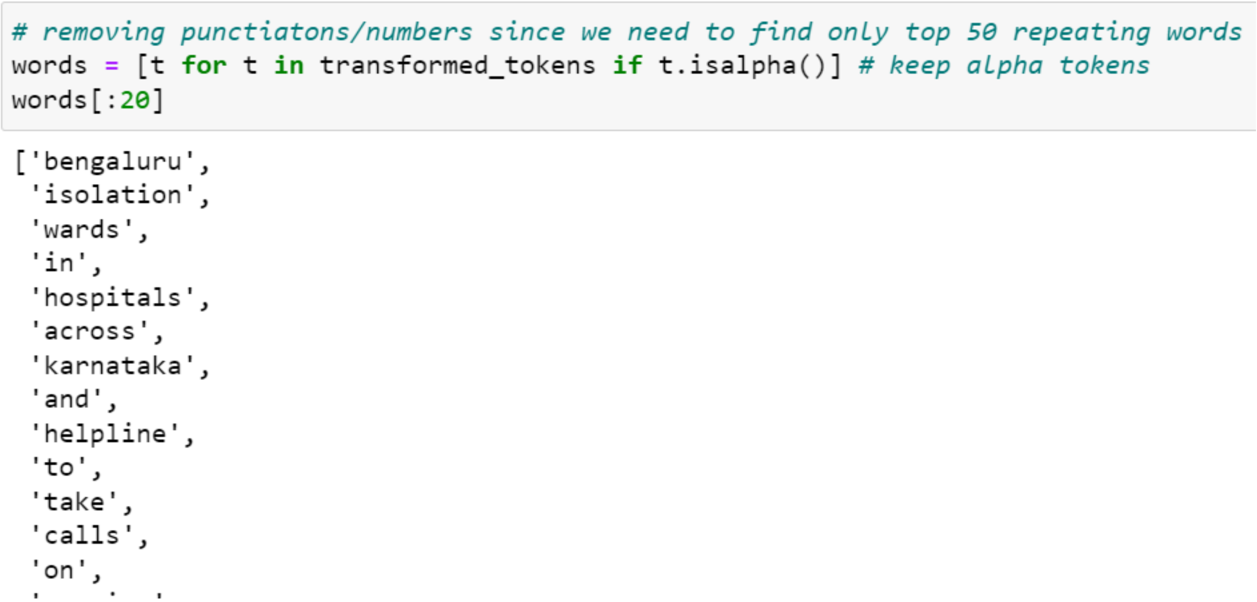
```
# checking for numeric values in the corpus
num_vals = [word for word in transformed_tokens if word.isnumeric()]
num_vals[:50]

# few of them seems to be the year - 2020, 2019, 2003, 2002

len(num_vals)
```

```
93885
```

```
num_freq = nltk.FreqDist(num_vals)
num_freq.most_common(20)

# news/blogs are talking a lot about the years 2020
```

```
[('2020', 10993),
 ('30', 4326),
 ('31', 3921),
 ('2019', 3466),
 ('1', 3270),
 ('10', 2292),
 ('14', 2170),
 ('2', 2140),
 ('170', 2104),
 ('213', 1815),
 ('3', 1749),
```

5. **Alphabetical data extraction:** Extracted all the words to understand the corpus better from bigram analysis, trigram analysis, frequently used words etc.,



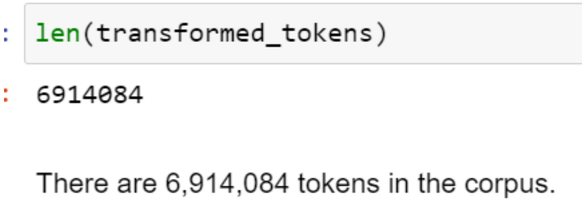
```
# removing punctiatons/numbers since we need to find only top 50 repeating words
words = [t for t in transformed_tokens if t.isalpha()] # keep alpha tokens
words[:20]
```

```
['bengaluru',
 'isolation',
 'wards',
 'in',
 'hospitals',
 'across',
 'karnataka',
 'and',
 'helpline',
 'to',
 'take',
 'calls',
 'on',
```

DATA ANALYSIS

Understanding Lexical Diversity

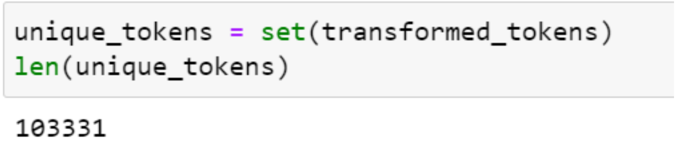
Total number of **tokens** in the corpus = 6914084.



```
: len(transformed_tokens)
: 6914084
```

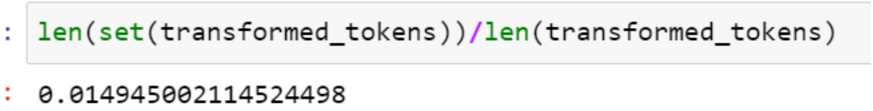
There are 6,914,084 tokens in the corpus.

Size of the **vocabulary** (unique tokens) = 103331.



```
unique_tokens = set(transformed_tokens)
len(unique_tokens)
103331
```

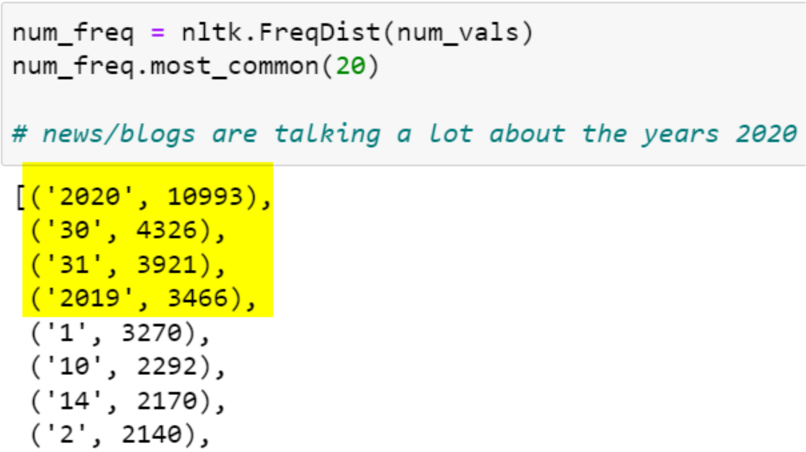
TTR - Type to Token ratio (number of unique tokens/total tokens) = 0.015. closer the TTR ration to 1, greater the lexical richness of the corpus. 0.015 is on the lower end. **Corpus is not lexically diverse**.



```
: len(set(transformed_tokens))/len(transformed_tokens)
: 0.014945002114524498
```

Numeric Data Analysis

From the numeric data frequency distribution, dates January 30, 2020, January 31, 2020, and the year 2019 are occurring quiet often in the corpus.



```
num_freq = nltk.FreqDist(num_vals)
num_freq.most_common(20)

# news/blogs are talking a lot about the years 2020
```

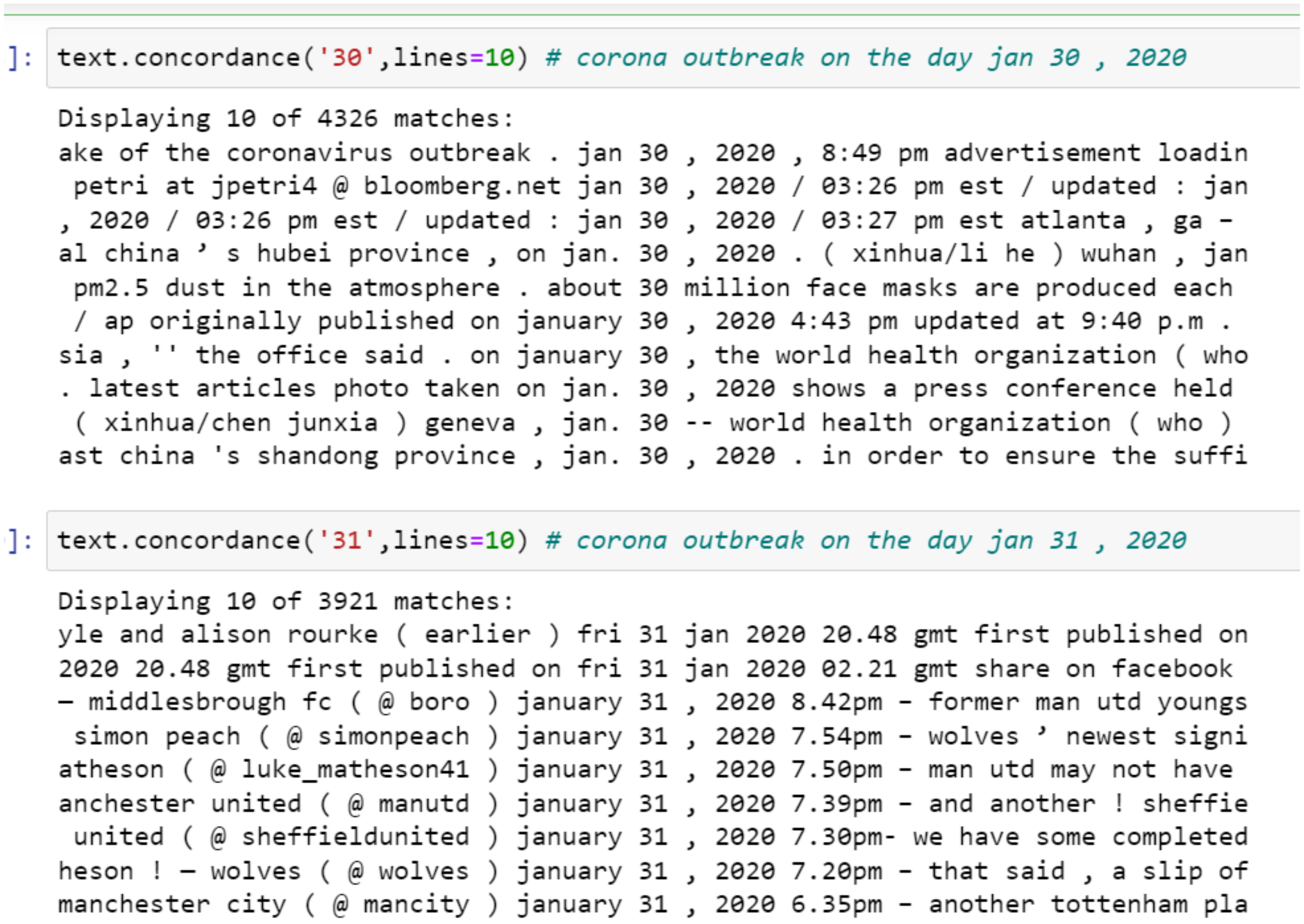
```
[('2020', 10993),
 ('30', 4326),
 ('31', 3921),
 ('2019', 3466),
 ('1', 3270),
 ('10', 2292),
 ('14', 2170),
 ('2', 2140),
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu



```
]: text.concordance('30',lines=10) # corona outbreak on the day jan 30 , 2020
```

```
Displaying 10 of 4326 matches:
ake of the coronavirus outbreak . jan 30 , 2020 , 8:49 pm advertisement loadin
 petri at jpetri4 @ bloomberg.net jan 30 , 2020 / 03:26 pm est / updated : jan
, 2020 / 03:26 pm est / updated : jan 30 , 2020 / 03:27 pm est atlanta , ga –
al china ’ s hubei province , on jan. 30 , 2020 . ( xinhua/li he ) wuhan , jan
 pm2.5 dust in the atmosphere . about 30 million face masks are produced each
 / ap originally published on january 30 , 2020 4:43 pm updated at 9:40 p.m .
sia , '' the office said . on january 30 , the world health organization ( who
. latest articles photo taken on jan. 30 , 2020 shows a press conference held
 ( xinhua/chen junxia ) geneva , jan. 30 -- world health organization ( who )
ast china 's shandong province , jan. 30 , 2020 . in order to ensure the suffi
```

```
]: text.concordance('31',lines=10) # corona outbreak on the day jan 31 , 2020
```

```
Displaying 10 of 3921 matches:
yle and alison rourke ( earlier ) fri 31 jan 2020 20.48 gmt first published on
2020 20.48 gmt first published on fri 31 jan 2020 02.21 gmt share on facebook
– middlesbrough fc ( @ boro ) january 31 , 2020 8.42pm – former man utd youngs
 simon peach ( @ simonpeach ) january 31 , 2020 7.54pm – wolves ’ newest signi
atheson ( @ luke_matheson41 ) january 31 , 2020 7.50pm – man utd may not have
anchester united ( @ manutd ) january 31 , 2020 7.39pm – and another ! sheffie
 united ( @ sheffieldunited ) january 31 , 2020 7.30pm- we have some completed
heson ! – wolves ( @ wolves ) january 31 , 2020 7.20pm – that said , a slip of
manchester city ( @ mancity ) january 31 , 2020 6.35pm – another tottenham pla
```

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu

Upon digging up the news articles around January 30th and 31st 2020, there were many news articles about “WHO declaring COVID-19 a Public Health Emergency of International Concern” and “International sporting events canceled in China as coronavirus spreads”.

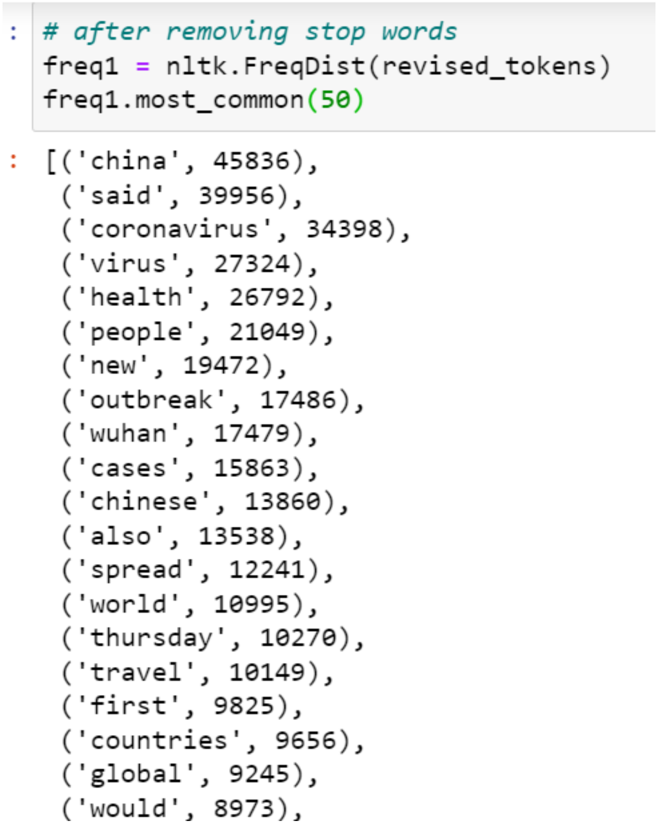
https://www.onthisday.com/date/2020/january/30

https://theweek.com/10things/884779/10-things-need-know-today-january-30-2020

Frequency Distribution

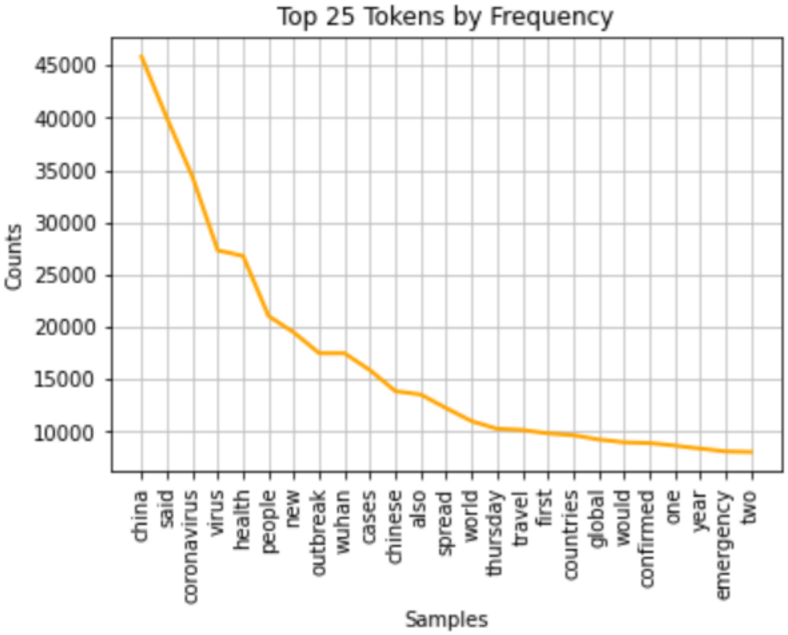
Before computing frequency distribution, all the stop words, punctuations, and numeric tokens are removed. Most repeated words found in the corpus are china, said, coronavirus, virus, health, people, new, outbreak, wuhan, cases, Chinese etc., Looking at the top 50 words, we can assume that people are talking about the origin of covid, declaring global health emergency, travel restrictions etc.,

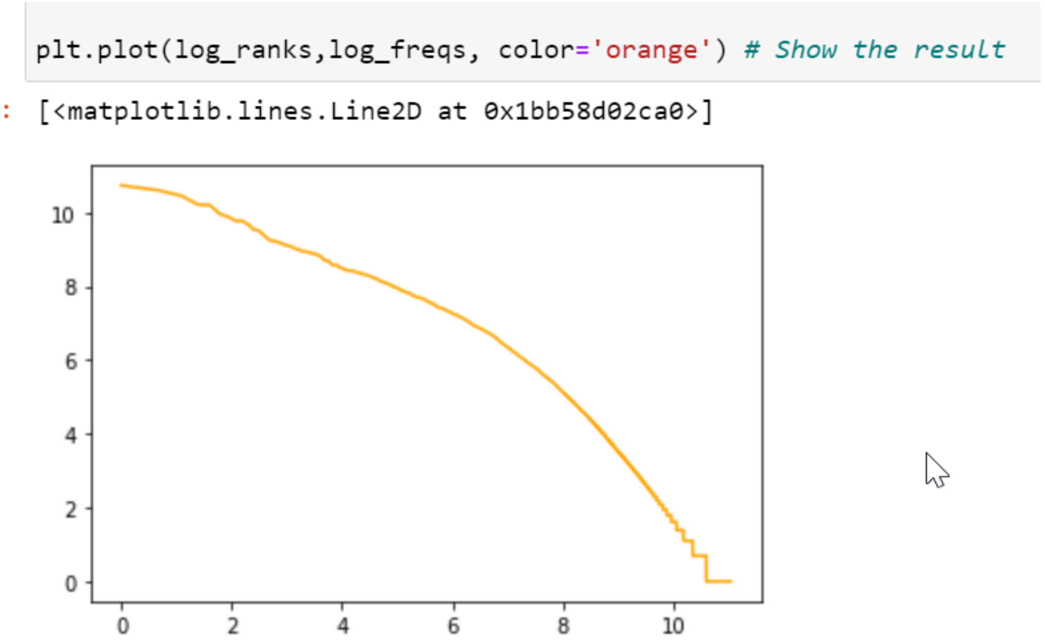
Top 50 words by their frequency: From the below chart, we can notice that frequency of the token is inversely proportional to the rank hence confirming the Zipf’s law. Word frequency and ranking on a log scale follows a nice straight line with negative slope.



```
: # after removing stop words
  freq1 = nltk.FreqDist(revised_tokens)
  freq1.most_common(50)
```

```
: [('china', 45836),
   ('said', 39956),
   ('coronavirus', 34398),
   ('virus', 27324),
   ('health', 26792),
   ('people', 21049),
   ('new', 19472),
   ('outbreak', 17486),
   ('wuhan', 17479),
   ('cases', 15863),
   ('chinese', 13860),
   ('also', 13538),
   ('spread', 12241),
   ('world', 10995),
   ('thursday', 10270),
   ('travel', 10149),
   ('first', 9825),
   ('countries', 9656),
   ('global', 9245),
   ('would', 8973),
```


Top 25 Tokens by Frequency
Counts
45000
40000
35000
30000
25000
20000
15000
10000
china
said
coronavirus
virus
health
people
new
outbreak
wuhan
cases
chinese
also
spread
world
thursday
travel
first
countries
global
would
confirmed
one
year
emergency
two
Samples

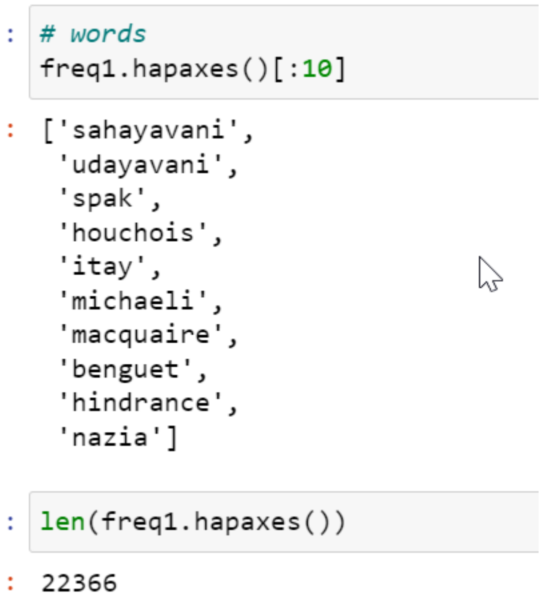


```
plt.plot(log_ranks,log_freqs, color='orange') # Show the result
```

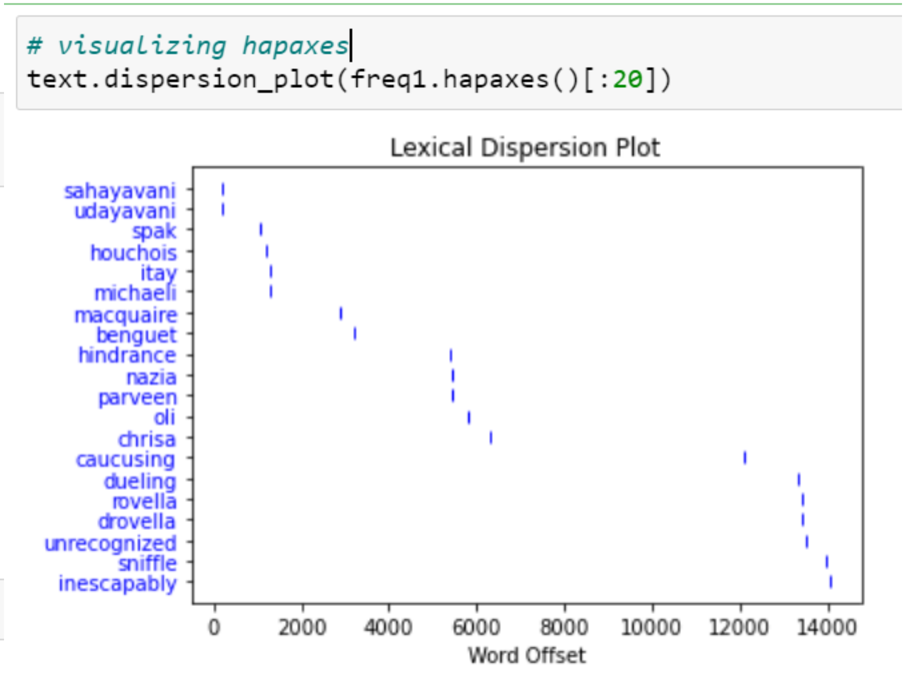
```
: [<matplotlib.lines.Line2D at 0x1bb58d02ca0>]
```

Hapax legomenon

There are 22,366 hapax tokens. We can just ignore these hapaxes from the corpus for further analysis.



```
: # words
  freq1.hapaxes()[:10]

: ['sahayavani',
   'udayavani',
   'spak',
   'houchois',
   'itay',
   'michaeli',
   'macquaire',
   'benguet',
   'hindrance',
   'nazia']

: len(freq1.hapaxes())

: 22366
```

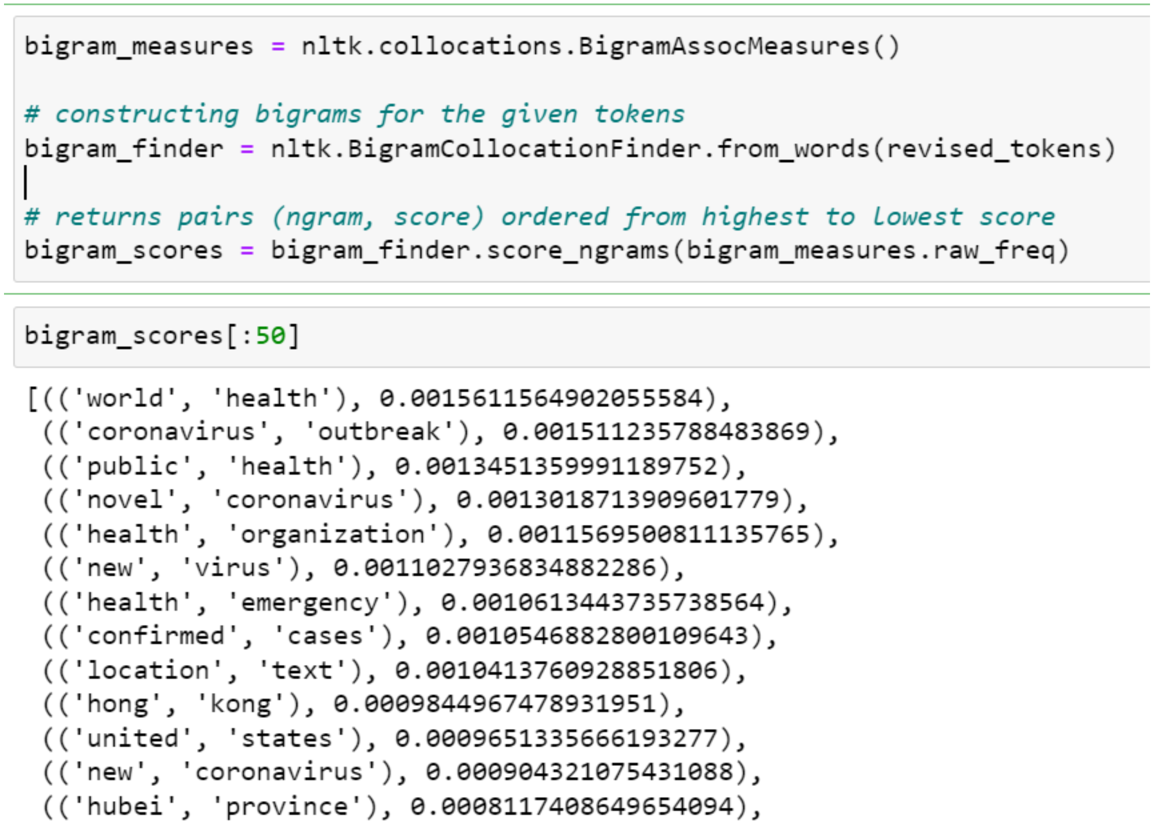


```
# visualizing hapaxes
text.dispersion_plot(freq1.hapaxes()[:20])
```

Bigram Analysis

Before computing frequency distribution, all the stop words, punctuations, and numeric tokens are removed. With bigrams, we can find more insight about the data than with unigrams/tokens. It helps to understand the context of the corpus in much better way.

Top bigrams by frequencies found in the corpus are “world health”, “coronavirus outbreak”, “public health”, “health organization”, “new virus”, “health emergency” “confirmed cases”, “location text”, “hong kong” etc., bigrams corroborate to the assumption that we made while analyzing most frequent tokens. Corpus is all about origin of covid, declaring global health emergency, travel restrictions etc.,

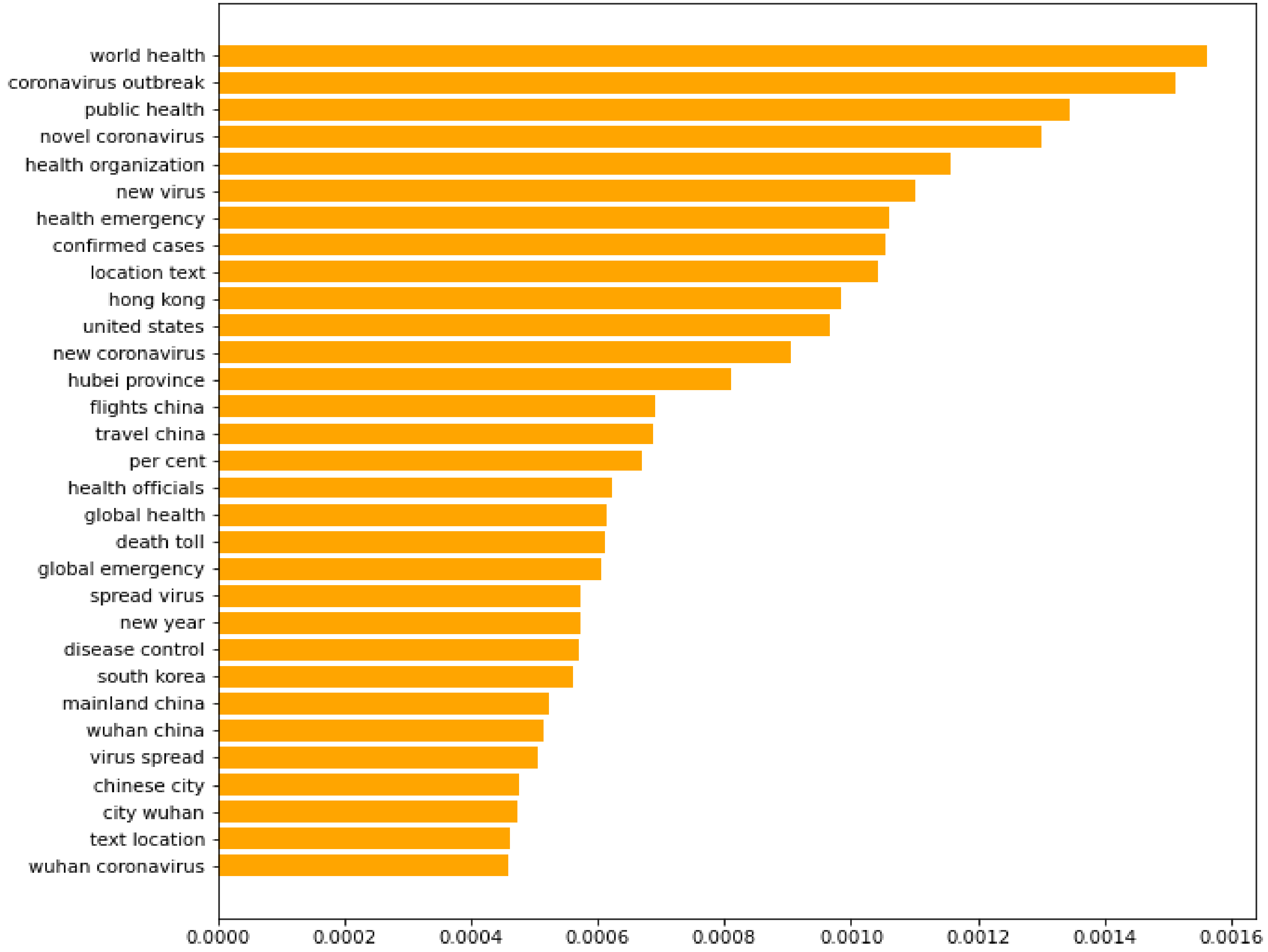


```
bigram_measures = nltk.collocations.BigramAssocMeasures()

# constructing bigrams for the given tokens
bigram_finder = nltk.BigramCollocationFinder.from_words(revised_tokens)

# returns pairs (ngram, score) ordered from highest to lowest score
bigram_scores = bigram_finder.score_ngrams(bigram_measures.raw_freq)
```

```
bigram_scores[:50]
```

```
[(('world', 'health'), 0.0015611564902055584),
 (('coronavirus', 'outbreak'), 0.001511235788483869),
 (('public', 'health'), 0.0013451359991189752),
 (('novel', 'coronavirus'), 0.001301871390960l779),
 (('health', 'organization'), 0.0011569500811135765),
 (('new', 'virus'), 0.0011027936834882286),
 (('health', 'emergency'), 0.0010613443735738564),
 (('confirmed', 'cases'), 0.0010546882800109643),
 (('location', 'text'), 0.0010413760928851806),
 (('hong', 'kong'), 0.0009844967478931951),
 (('united', 'states'), 0.0009651335666193277),
 (('new', 'coronavirus'), 0.000904321075431088),
 (('hubei', 'province'), 0.0008117408649654094),
```

Visualizing Top Bigrams found in the corpus:


world health
coronavirus outbreak
public health
novel coronavirus
health organization
new virus
health emergency
confirmed cases
location text
hong kong
united states
new coronavirus
hubei province
flights china
travel china
per cent
health officials
global health
death toll
global emergency
spread virus
new year
disease control
south korea
mainland china
wuhan china
virus spread
chinese city
city wuhan
text location
wuhan coronavirus
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
0.0016

Visualizing Top Bigrams by PMI score:

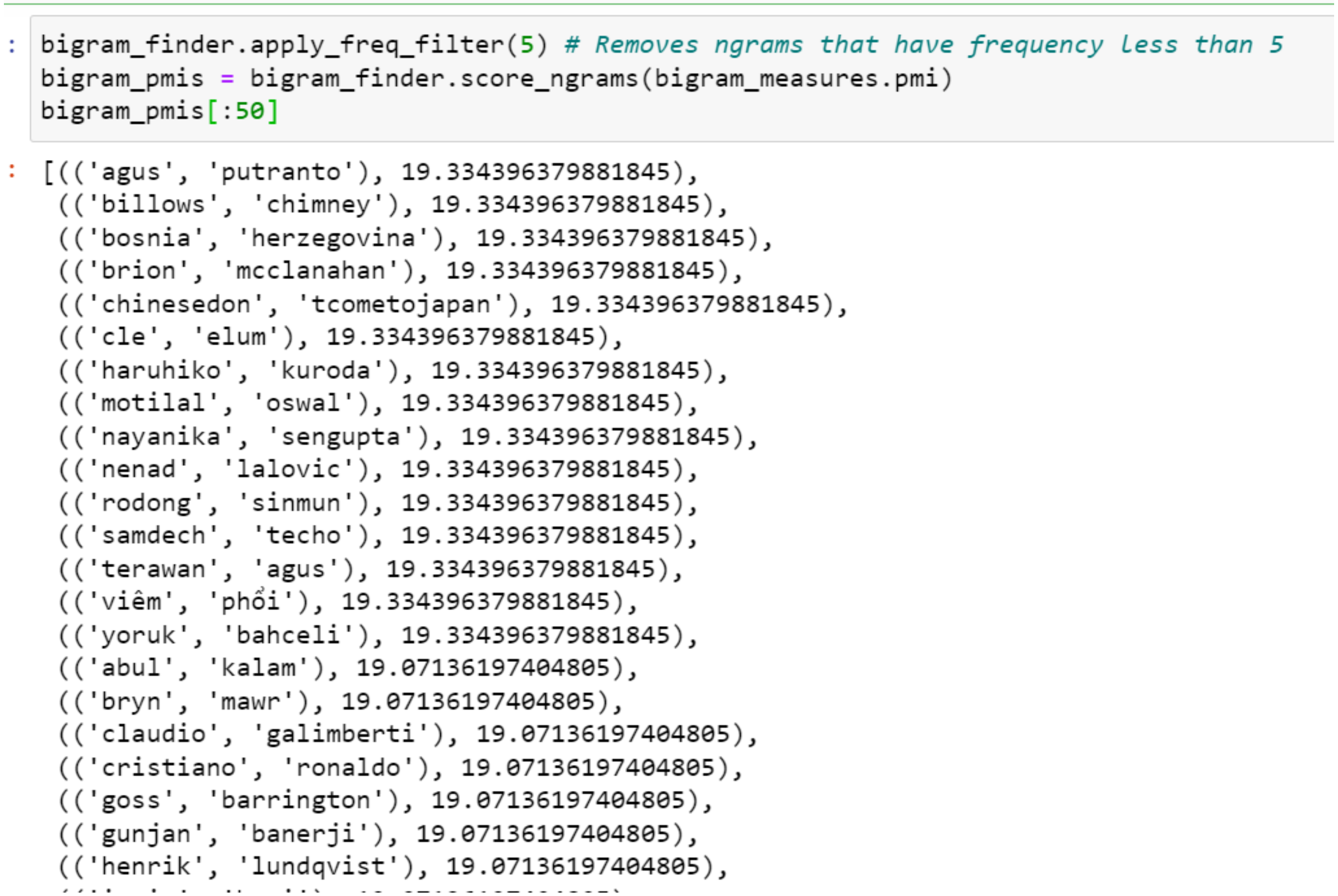
Higher the PMI score, more strongly connected bigram tokens. Looking at top bigrams scored by their PMI measure, several of these are:

a. leaders name from different country - “Bosnia and Herzegovina”, “Haruhiko Kuroda”

b. country names – “Cle Elum”

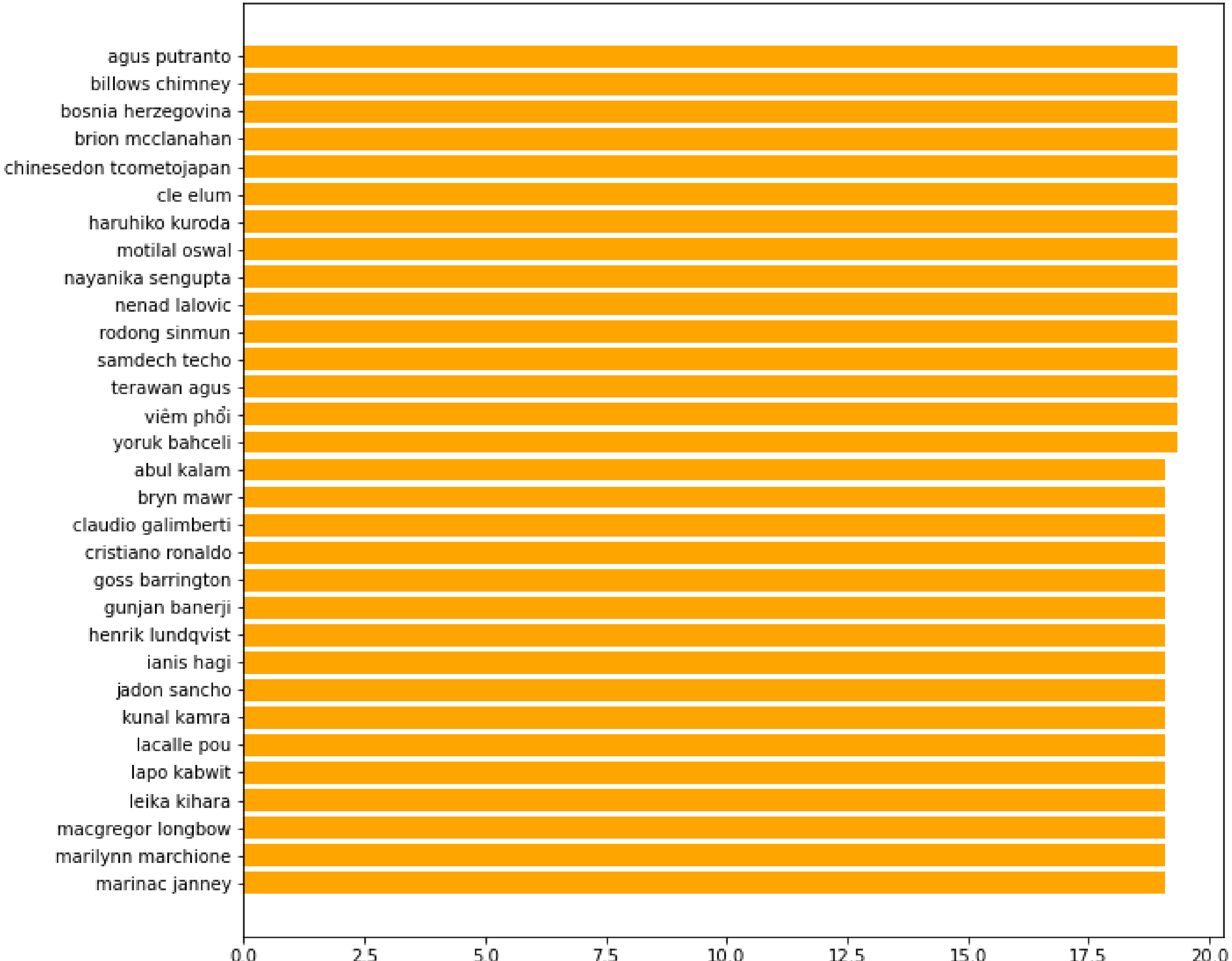
c. historian/reporter names – “Brion McClanahan”,” Nayanika Sengupta”

These words tend to reveal location, people, countries etc.,



```
: bigram_finder.apply_freq_filter(5) # Removes ngrams that have frequency less than 5
  bigram_pmis = bigram_finder.score_ngrams(bigram_measures.pmi)
  bigram_pmis[:50]
```

```
: [(('agus', 'putranto'), 19.334396379881845),
   (('billows', 'chimney'), 19.334396379881845),
   (('bosnia', 'herzegovina'), 19.334396379881845),
   (('brion', 'mcclanahan'), 19.334396379881845),
   (('chinesedon', 'tcometojapan'), 19.334396379881845),
   (('cle', 'elum'), 19.334396379881845),
   (('haruhiko', 'kuroda'), 19.334396379881845),
   (('motilal', 'oswal'), 19.334396379881845),
   (('nayanika', 'sengupta'), 19.334396379881845),
   (('nenad', 'lalovic'), 19.334396379881845),
   (('rodong', 'sinmun'), 19.334396379881845),
   (('samdech', 'techo'), 19.334396379881845),
   (('terawan', 'agus'), 19.334396379881845),
   (('viêm', 'phổi'), 19.334396379881845),
   (('yoruk', 'bahceli'), 19.334396379881845),
   (('abul', 'kalam'), 19.07136197404805),
   (('bryn', 'mawr'), 19.07136197404805),
   (('claudio', 'galimberti'), 19.07136197404805),
   (('cristiano', 'ronaldo'), 19.07136197404805),
   (('goss', 'barrington'), 19.07136197404805),
   (('gunjan', 'banerji'), 19.07136197404805),
   (('henrik', 'lundqvist'), 19.07136197404805),
```


agus putranto
billows chimney
bosnia herzegovina
brion mcclanahan
chinesedon tcometojapan
cle elum
haruhiko kuroda
motilal oswal
nayanika sengupta
nenad lalovic
rodong sinmun
samdech techo
terawan agus
viêm phổi
yoruk bahceli
abul kalam
bryn mawr
claudio galimberti
cristiano ronaldo
goss barrington
gunjan banerji
henrik lundqvist
ianis hagi
jadon sancho
kunal kamra
lacalle pou
lapo kabwit
leika kihara
macgregor longbow
marilynn marchione
marinac janney
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

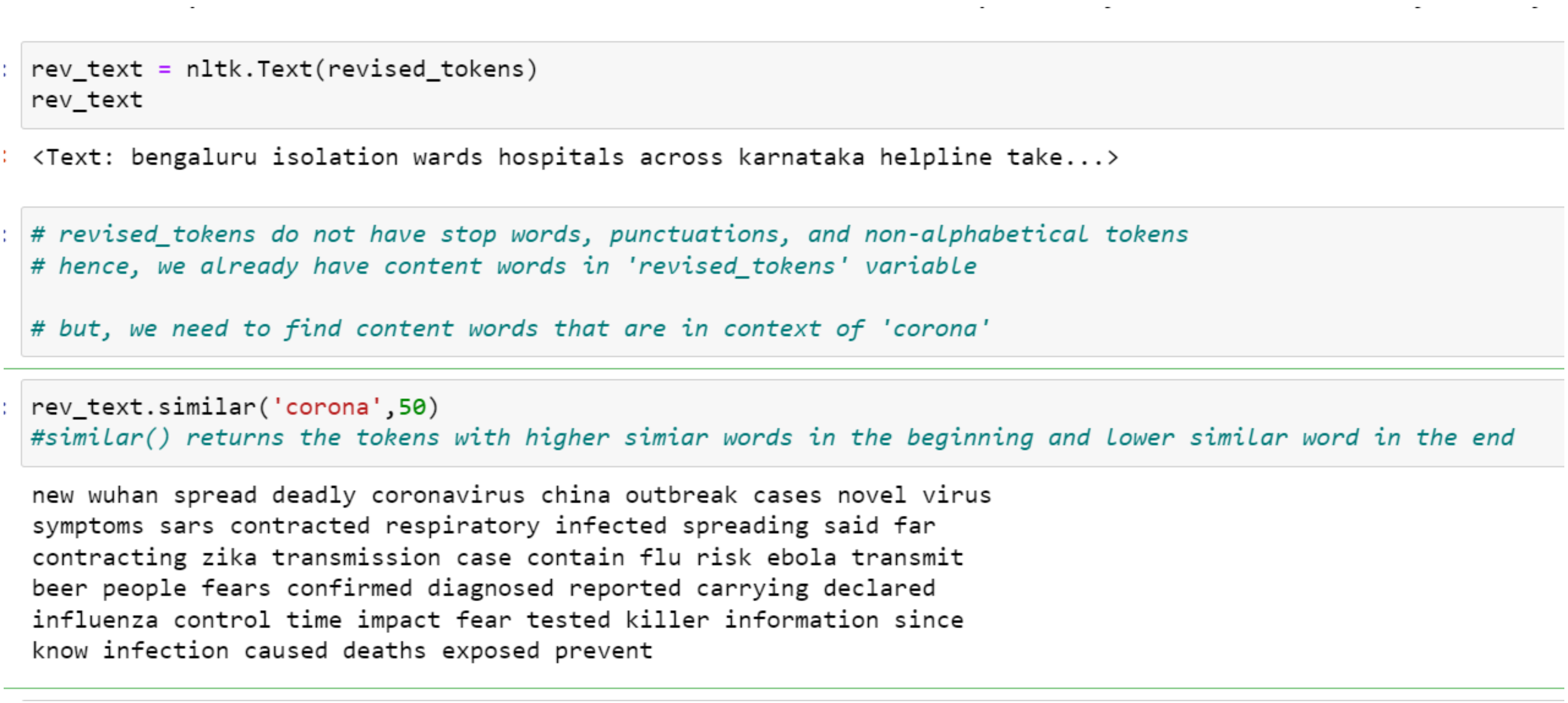
Email: ckoppara@syr.edu

Top 50 content words in the context of the word “Covid” Please explain how you define the context in your study
By removing all the stop words, punctuations, and numeric tokens from the corpus, we will be left with only content words. From the content words, we can find all the tokens that have similar meaning/context by using **similar (“covid”, 50)** NLTK library API to find words in the context of “covid”. **similar ()** returns the tokens with higher similarity words in the beginning and lower similarity word in the end.

Top 50 words in the context of the word “covid” that was found in the corpus are
"new","wuhan","spread","deadly","coronavirus","china","outbreak","cases","novel","virus","symptoms",

"sars","contracted","respiratory","infected","spreading","said","far","contracting","zika","transmission","case","contain","flu","risk","ebola","transmit","beer","people","fears","confirmed","diagnosed",

"reported","carrying","declared","influenza","control","time","impact","fear","tested","killer",
"information","since","know","infection","caused","deaths","exposed","prevent".



```
rev_text = nltk.Text(revised_tokens)
rev_text
```

```
<Text: bengaluru isolation wards hospitals across karnataka helpline take...>
```

```
# revised_tokens do not have stop words, punctuations, and non-alphabetical tokens
# hence, we already have content words in 'revised_tokens' variable

# but, we need to find content words that are in context of 'corona'
```

```
rev_text.similar('corona',50)
#similar() returns the tokens with higher simiar words in the beginning and lower similar word in the end
```

```
new wuhan spread deadly coronavirus china outbreak cases novel virus
symptoms sars contracted respiratory infected spreading said far
contracting zika transmission case contain flu risk ebola transmit
beer people fears confirmed diagnosed reported carrying declared
influenza control time impact fear tested killer information since
know infection caused deaths exposed prevent
```

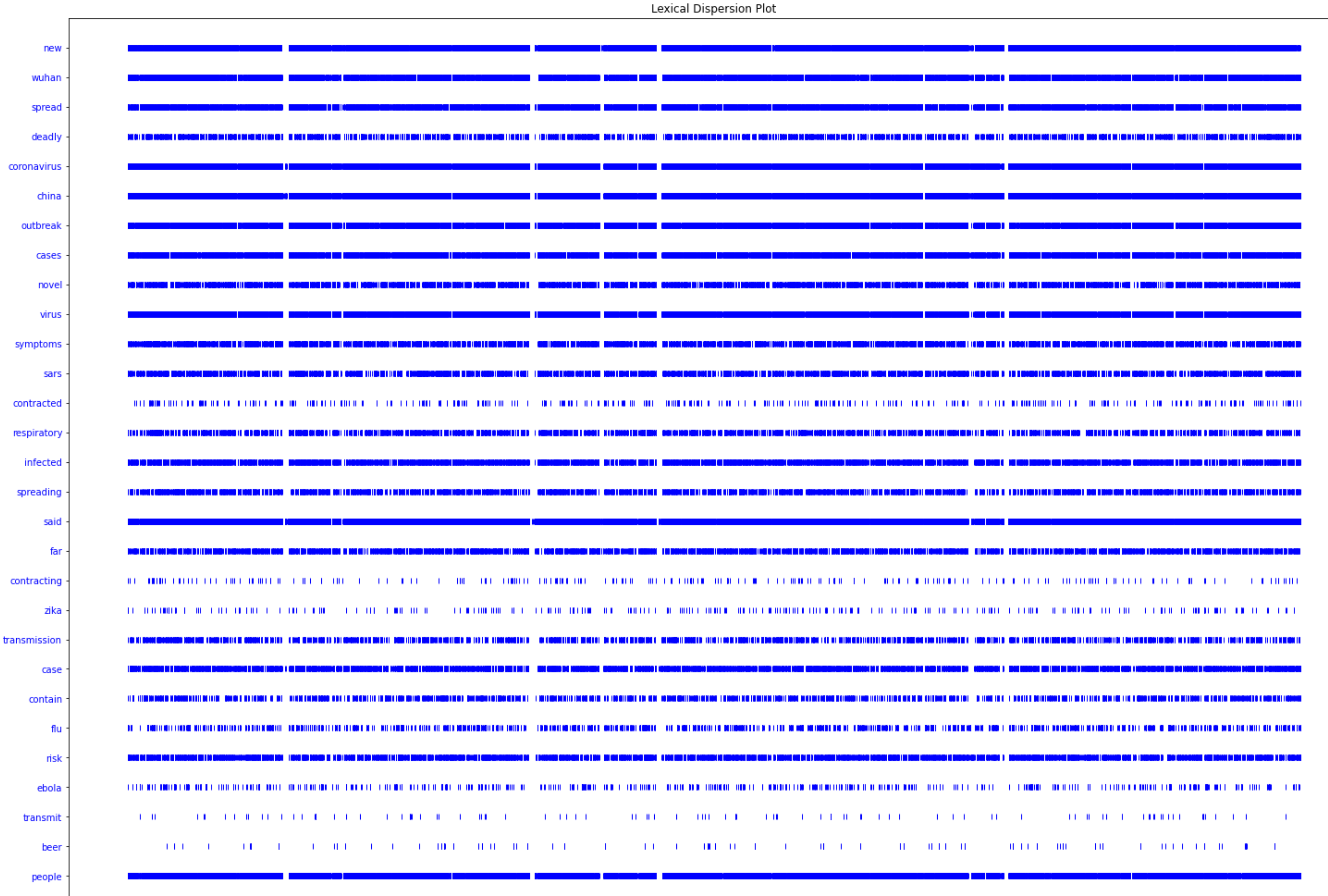
Dispersion plot of the above tokens found in the corpus:

IST 664: Natural Language Processing

Name: Chaithra Kopparam Cheluvaiah

SUID: 326926205

Email: ckoppara@syr.edu



INTERPRETATION OF THE RESULTS

Below are the inferences that can be drawn from the corpus statistics:

- Since the data (news/blogs) are from **Dec 2019 till March 2020**, covid outbreak had just begun. So, news articles and the blogs are mostly talking about origin of covid, declaring global emergency, travel restrictions etc.,
- Most of the news/blogs belong to USA
- January 30, 2020, and January 31,2020 dated articles are high in number due to WHO declaring covid as global pandemic
- From looking at the sentimental words from high frequent tokens – fear, deadly, emergency, we might say that people are paranoid about the new virus being spread
- Lexical richness of the corpus is low. This is obvious because news articles/blogs are curated in such a way that they are more understandable

Additional analysis that can be performed in the future are:

- Sentiment analysis of the news messages/blogs would have helped in understanding whether news/blogs are alarming/neutral etc.,
- Identifying fake news by classification/clustering technique
- Temporal Reasoning to identify how people felt about covid over the time