In [1]:
```
#Loading Metadata
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
plt.style.use('ggplot')
```

In [2]:
```
root_path = 'data/CORD_19_research_challenge'
```

In [3]:
```
metadata_path = f'{root_path}/Covid_19_Dataset.csv'
```

In [13]:
```
df_covid = pd.read_csv(metadata_path, dtype={
    'pubmed_id': str,
    'Microsoft Academic Paper ID': str,
    'doi': str
})
```

In [14]:
```
df_covid.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 256684 entries, 0 to 256683
Data columns (total 9 columns):
 #   Column            Non-Null Count   Dtype 
---  ------            --------------   ----- 
 0   Unnamed: 0        256684 non-null  int64 
 1   paper_id          256684 non-null  object
 2   doi               247073 non-null  object
 3   abstract          181579 non-null  object
 4   body_text         256684 non-null  object
 5   authors           254329 non-null  object
 6   title             256680 non-null  object
 7   journal           231637 non-null  object
 8   abstract_summary  256684 non-null  object
dtypes: int64(1), object(8)
memory usage: 17.6+ MB
```

In [15]:
```
#Handle Possible Duplicates
```

In [16]:
```
df_covid.drop_duplicates(['abstract', 'body_text'], inplace=True)
df_covid['abstract'].describe(include='all')
```

Out[16]:
```
count                                                181488
unique                                               180200
top       Publisher's Note Springer Nature remains neutr...
freq                                                    211
```