In [2]:

```
import pandas as pd
import numpy as np
import requests as rq
from bs4 import BeautifulSoup
from tqdm import tqdm

import os
from google.colab import drive
drive.mount('/content/drive')
os.chdir("/content/drive/MyDrive/Colab_Notebooks/")
```

```
Mounted at /content/drive
```

In [30]:

```
# Enter file path for the GDI million links csv. The file is in GDI folder under partner
data on SharePoint.
filepath = 'scraped_fulltext.csv'

# Enter the search term list you'd like to find in texts
topics = ['voter fraud','white supremacy','anti-latinx','biden','big tech', \
          'anti-muslim','abortion','president biden','coronavirus',\
          'anti-lgbt','misogyny','anti-immigrant','anti-black',\
          'antivaxx','pseudoscience','qanon','5g','critical race theory','aliens']

# Data transformation to return the 84785 urls that have GDI topics.
GDI_links = pd.read_csv(filepath)
# GDI_links = GDI_links[GDI_links['error'] != 'Error']
GDI_links = GDI_links[['uid', 'url', 'keywords', 'classifiers', 'full_text']]
GDI_links = GDI_links.dropna(axis = 0, how = 'any')
```

Out[30]:

| | uid | url | keywords | classifier |
|---|---|---|---|---|
| 0 | 380117 | https://www.westernjournal.com/trump-right-jud... | coronavirus\|covid\|qanon\|biden\|voterfraud\|antii... | voterfrau |
| 1 | 448712 | https://pjmedia.com/news-and-politics/victoria... | antilatinx\|whitesupremacy\|antiblack | whitesupremac |
| 2 | 256646 | https://www.breitbart.com/politics/2020/12/23/... | biden\|antilatinx\|antiimmigrant\|coronavirus | antilatinx\|bide |
| 3 | 406930 | https://www.theepochtimes.com/supreme-court-ju... | bigtech\|qanon\|qanon | bigtec |
| 5 | 196599 | https://www.zerohedge.com/political/leftists-s... | biden | bide |
| 6 | 805032 | https://www.commondreams.org/news/2021/01/25/b... | antiimmigrant\|biden\|biden\|voterfraud | bide |
| 7 | 276039 | https://www.thegatewaypundit.com/2020/10/break... | biden | bide |
| 8 | 202163 | https://www.wnd.com/2020/11/twj-exclusive-vira... | biden\|voterfraud\|antiimmigrant | voterfraud\|bide |

In [ ]:

```
# Topics from the topic list found in the 'full_text' column, stored in the'topic_found'
column.
texts = GDI_links['full_text']
final = []

for i in texts:
    try:
      result = ''
      for t in topics:
        if i.lower().find(t.lower()) != -1:
          result = result + t + "|"
      result = result[:-1]
      final.append(result)
    except:
      final.append(np.nan)
```

```
GDI_links['topic_found'] = final
```

In [ ]:

```
# Enter filepath/name of output csv.
destination = 'TopicFoundInText.csv'

GDI_links.to_csv(destination)
```