In [1]:

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

from concurrent.futures import ThreadPoolExecutor
import psutil
```

In [2]:

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

# Data transformation to return the 84785 urls that have GDI topics.
GDI_links = pd.read_csv(filepath)
GDI_links = GDI_links[['uid', 'url', 'keywords', 'classifiers']]
GDI_links = GDI_links.dropna(axis = 0, how = 'any')
GDI_links = GDI_links[0:50]
GDI_links.head()
```

Out[2]:

| | uid | url | keywords | classifiers |
|---|---|---|---|---|
| 0 | 380117 | https://www.westernjournal.com/trump-right-jud... | coronavirus\|covid\|qanon\|biden\|voterfraud\|antii... | voterfraud |
| 1 | 448712 | https://pjmedia.com/news-and-politics/victoria... | antilatinx\|whitesupremacy\|antiblack | whitesupremacy |
| 2 | 256646 | https://www.breitbart.com/politics/2020/12/23/... | biden\|antilatinx\|antiimmigrant\|coronavirus | antilatinx\|biden |
| 3 | 406930 | https://www.theepochtimes.com/supreme-court-ju... | bigtech\|qanon\|qanon | bigtech |
| 4 | 281134 | https://thefederalist.com/2021/01/09/twitter-i... | pseudoscience\|biden\|disinformation | biden |

In [3]:

```
print(len(GDI_links["uid"]))
```

50

In [4]:

```
def scrapeArticle(id, url):
    try:
        article = Article(url)
        article.download()
        article.parse()
        results = {id: article.text}
    except:
        article = ''
        results = {id: article}
    print(id)
    return results

def set_up_threads(urls):
    threads_count = psutil.cpu_count()
    with ThreadPoolExecutor(max_workers = threads_count) as executor:
        return tqdm(executor.map(scrapeArticle,
                                 urls["uid"],
                                 urls["url"],
                                 timeout = 60))
```

In [5]:

```
urls = GDI_links[['uid','url']]

scrape = set_up_threads(urls)

results = {}

for s in scrape:
      results[list(s.keys())[0]] = list(s.values())[0]

final = pd.DataFrame.from_records([results]).T
```

```
0it [00:00, ?it/s]
```

```
448712
805032
196599
281134380117

256646
147565
217524
276039406930

202163
157800
307097
226088
208364
376591
404672
822213
300549
513266
170089
411610
213072
818583
573539
358374
310587
171258
248910
243361
209514262610

134816
381184
190655
412523
274810
442075
248826
668936
173287
472746
339429
798944
446044
897877
167779
725267
299503
213488
```

```
50it [00:23,  2.13it/s]
```

In [6]:

```
GDI_links.index = final.index
GDI_links["full_text"] = final
GDI_links.head()
```

Out[6]:

Out[6]:

| | uid | url | keywords | classifier |
|---|---|---|---|---|
| 380117 | 380117 | https://www.westernjournal.com/trump-right-jud... | coronavirus\|covid\|qanon\|biden\|voterfraud\|antii... | voterfrau |
| 448712 | 448712 | https://pjmedia.com/news-and-politics/victoria... | antilatinx\|whitesupremacy\|antiblack | whitesupremac |
| 256646 | 256646 | https://www.breitbart.com/politics/2020/12/23/... | biden\|antilatinx\|antiimmigrant\|coronavirus | antilatinx\|bide |
| 406930 | 406930 | https://www.theepochtimes.com/supreme-court-ju... | bigtech\|qanon\|qanon | bigtec |
| 281134 | 281134 | https://thefederalist.com/2021/01/09/twitter-i... | pseudoscience\|biden\|disinformation | bide |

In [7]:

```
GDI_links.to_csv("new_test2.csv")
```

In [ ]:

```
```