Crawling the websites

Approach:

To Crawl the URLs given in the dataset and look for categories in the websites

Building the web crawler was pretty straightforward.
Python has really good opensource support for crawling the webpages online.
Packages used :

- BeautifulSoup - An opensource solution to easily locate information in a html page given an html tag
- requests - this is required for requesting the webpage and to download it locally to perform information search
- pandas - dataset is imported to pandas dataframe which makes it easy to access the data columnwise
- openpyxl - for writing each output to an excel file
- datefinder - to properly parsed a given string into a datetime object

Issues encountered:

- Inconsistent Web Page Practices:most of the websites doesnt include any support fot metadata, This makes crawling really difficult since there is no consistent practice to store the information on the web page,This is probably due to unprofessional/ unexperienced developers
- Slow Process:Crawling is really slow and tiring process if you dont have the sufficient resources for eg:Multiple Online Servers And since your machine has to constantly request for pages online this also creates congestion in the network
- Network Dependent:if your network becomes unresponsive this will stop the process of crawling
- repeated requests:if you request a server again and again within a short period of time, it is possible that your ip will get blacklisted and you wont be able to crawl again
- Crawling support: most of the webpages on the internet has Robots.txt file which provides rules as to which content you are allowed to crawl and which are not.Some dont allow any crawling whatsoever
- Constantly writing to file: crawling is not suitable for saving the records into ram,you have to keep writing to a file to make sure you dont lose any progress if an unexpected event happens

Insights:

Out of Total 87 Unique websites from the datasets, 62 websites are perfectly crawlable (some needs effort to look for information).the remaining 25 websites are either not crawlable or they dont give you the access to crawl,some dont have proper tags to look for information ,some dont even have the categories to look for,some has inconsistent web meta data structure

Since every website has a different information structure it becomes unavoidable to write unique conditions to look for information at each websites which is a really tiring process

Some websites have 'Category tags',some has 'tags' which gives only related topics.My initial approach is to look for Category tag first, if it is not there then look for related tags and if there are multiple tags select the first one. since this is not consistent in websites , it is possible to get the names of personalities instead of categories

In [1]:

```
import urllib3
import datefinder
from openpyxl import Workbook
import openpyxl
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

In [10]:

```
df1=pd.read_excel(r'C:\Users\windows\Desktop\Bigdata dataset\1.xlsx',usecols=[0],se
```

In [11]:

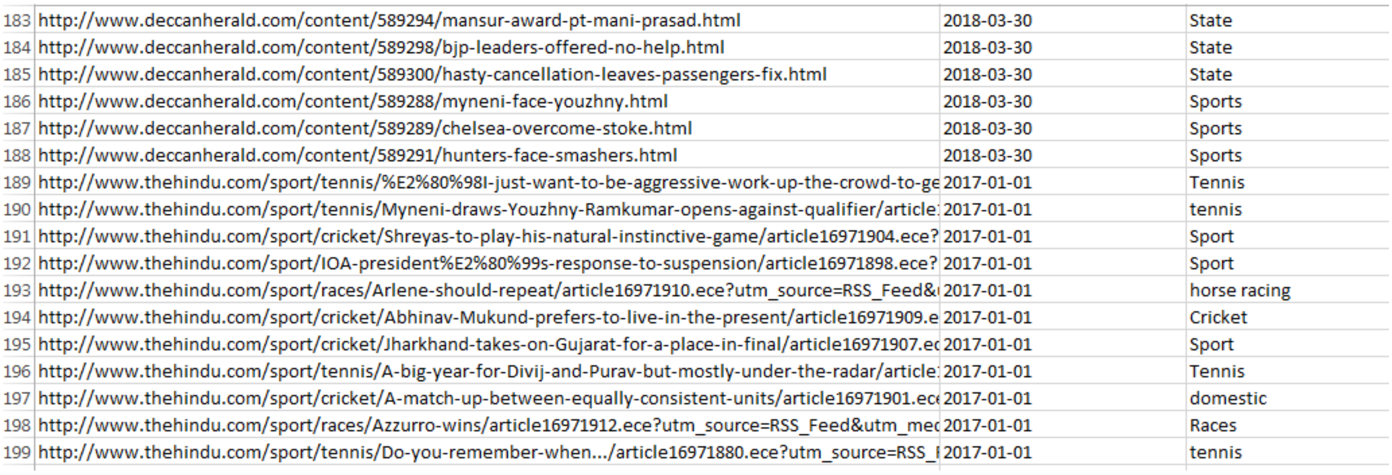
```
urls=df1.url.values
```

In [ ]:

```
book = openpyxl.load_workbook(r'C:\Users\windows\Desktop\Bigdata dataset\book1.xlsx
sheet=book.get_sheet_by_name("Sheet1")
for pg in urls[sheet.max_row-1:]:
    cls=""
    datetime=""
    try:
        r=requests.get(pg)
    except requests.exceptions.RequestException as e:
        print (e)
        continue
    r.encoding='utf-8'
    soup = BeautifulSoup(r.content,'lxml')
    if "thehindu.com" in pg:
        publishbox=soup.find('none')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        data = soup.findAll('div',attrs={'class':'tag-button'})
        if data is not None:
            for div in data:
                a=div.findAll('a')
                for c in a:
                    cls=c.text.strip()
        if cls=="":
            maincat=soup.find('a',attrs={'class':'tag-button section-button'})
            if maincat is not None:
                cls=maincat.text.strip()

    elif "accommodationtimes.com" in pg:
        publishbox=soup.find('div',attrs={'id':'datemeta_1'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls='Real Estate'

    elif "afaqs.com" in pg:
        publishbox=soup.find('div',attrs={'class':'repo_name'})
        if publishbox is not None:
            str=publishbox.text.strip()
            str=str.split('|')
            cls=str[1]
            cls=cls.split(' ')[2].strip()
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "betterphotography.in" in pg:
        publishbox=soup.find('time')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls="photography"

    elif "bollywoodhungama.com" in pg:
        publishbox=soup.find('time')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())

            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'tag-links'})
        if tags is not None:
            tags=tags.text.strip()
            cls=tags.split(',')[0].split(':')[1].strip()

    elif "bollywoodhungama.com/movie" in pg:
        publishbox=soup.find('time')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls='movie'

    elif "breakingnews.ie" in pg:
        publishbox=soup.find('span',attrs={'class':'date-time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'keywords'})
        if tags is not None:
            tags=tags.text.strip()
            cls=tags.split(',')[0].split(':')[1].strip()

    elif "thehindubusinessline.com" in pg:
        publishbox=soup.find('none')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'tag-button'})
        if tags is not None:
            cls=tags.text.strip()

    elif "business-standard.com" in pg:
        publishbox=soup.find('span',attrs={'itemprop':'datePublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'related-keyword'})
        if tags is not None:
            cls=tags.text.split('|')[-1].strip()

    elif "catchnews.com" in pg:
        publishbox=soup.findAll('span',attrs={'class':'artical_news_time'})
        if publishbox is not None:
            for publish in publishbox:
                date=datefinder.find_dates(publish.text.strip())
                for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'ins_keyword'})
        if tags is not None:
            cls=tags.text.strip().split('\n')[0]

    elif "constructionworld.in" in pg:
        publishbox=soup.find('span',attrs={'class':'date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'tags'})
        if tags is not None:
            for link in tags:
                a=link.find('a')
            a=a.text.strip()

    elif "daijiworld.com" in pg:
        publishbox=soup.find('ul',attrs={'class':'post-tags'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "daily.bhaskar.com" in pg:
        publishbox=soup.find('p',attrs={'class':'dba_pdate'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "bhaskar.com" in pg:
        publishbox=soup.find('p',attrs={'class':'ba_date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0]

    elif "deccanherald.com" in pg:
        publishbox=soup.find('div',attrs={'class':'postedBy'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'breadcrumb'})
        if tags is not None:
            cls=tags.text.strip().split('»')[1].strip()

    elif "dnaindia.com" in pg:
        publishbox=soup.find('date')
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

        tags=soup.find('ol',attrs={'class':'breadcrumbx'})
        if tags is not None:
            cls=tags.text.strip().split('\n')[1].strip()

    elif "equitymaster" in pg:
        publishbox=soup.find('div',attrs={'class':'closeDate'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
    elif "cricinfo.com" in pg:
        publishbox=soup.find('span',attrs={'data-dateformat':'date1'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls="sports"

    elif "daily-express-weird" in pg:
        publishbox=soup.find('div',attrs={'class':'dates'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls="weird"

    elif "firstpost.com" in pg:
        publishbox=soup.find('span',attrs={'class':'article-date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

        tags=soup.find('a',attrs={'class':'section-btn'})
        if tags is not None:
            cls=tags.text.strip()

    elif "fonearena" in pg:
        publishbox=soup.find('time',attrs={'class':'entry-date published updated'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content').split(' ')[0]

    elif "forbesindia.com" in pg:
        publishbox=soup.find('div',attrs={'class':'update-date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
    elif "glamsham.com" in pg:
        publishbox=soup.find('meta',attrs={'name':'Last-Modified'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'news_keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]


    elif "huffingtonpost.in" in pg:
        publishbox=soup.find('span',attrs={'class':'timestamp__date timestamp__date
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('span',attrs={'class':'entry-eyebrow'})
        if tags is not None:
            cls=tags.text.strip()

    elif "indiaglitz" in pg:
        publishbox=soup.find('meta',attrs={'itemprop':'datePublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'itemprop':'articleSection'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "indiainfoline" in pg:
        publishbox=soup.find('p',attrs={'class':'source fs14e mt5 mb5'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "indiatoday"  in pg:
        publishbox=soup.find('li',attrs={'class':'pubdata'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'news_keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]


    elif "indiatvnews" in pg:
        publishbox=soup.find('meta',attrs={'http-equiv':'Last-Modified'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'news_keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0].split(' ')[0]

    elif "itnewsonline" in pg:
        datetime=""
        cls='technology'

    elif "kanglaonline" in pg:
        publishbox=soup.find('time',attrs={'class':'entry-date updated td-module-da
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('ul',attrs={'class':'td-tags td-post-small-box clearfix'})
        if tags is not None:
            a=tags.find('a')
            cls=a.text.strip()
    elif "feedproxy.google.com/~r/LinuxForYou/" in pg:
        publishbox=soup.find('time',attrs={'class':'entry-date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'meta-tags clearfix'})
        if tags is not None:
            a=tags.find('a')
            cls=a.text.strip()

    elif "mainstreamweekly" in pg:
        publishbox=soup.find('p',attrs={'class':'surtitre'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "http://feedproxy.google.com/~r/MathrubhumiEnglish/" in pg:
        publishbox=soup.find('div',attrs={'class':'date_outer'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "http://feedproxy.google.com/~r/allhealthnews/" in pg:
        publishbox=soup.find('meta',attrs={'name':'DC.date.issued'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "http://feedproxy.google.com/~r/medianama/" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'news_keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "moneycontrol" in pg:
        publishbox=soup.find('meta',attrs={'http-equiv':'Last-Modified'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'news_keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]



    elif "businesstoday" in pg:
        publishbox=soup.find('meta',attrs={'itemprop':'datePublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content',None).split(',')[0]

    elif "msn.com" in pg:
        publishbox=soup.find('span',attrs={'class':'time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls=pg.split('/')[5]

    elif "http://feedproxy.google.com/~r/feedburner/HlRy/"  in pg:
        publishbox=soup.find('span',attrs={'class':'source'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "newsonair.nic.in" in pg:
        publishbox=pg.split('=')[1].split('&')[0].strip()
        if publishbox is not None:
            cls=publishbox

    elif "http://feedproxy.google.com/~r/newscomauwtfndm/" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0].strip()

    elif "nowrunning.com" in pg:
        publishbox=soup.find('meta',attrs={'name':'pubdate'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=pg.split('/')[5]
        if tags is not None:
            cls=tags.strip()

    elif "odditycentral.com" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('span',attrs={'class':'post-info'})
        if tags is not None:
            a=tags.find('a')
            cls=a.text.strip()

    elif "feedproxy.google.com/~r/oneindia" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=pg.split('/')[4].split('-')[-1]
        if tags is not None:
            cls=tags.strip()

    elif "feedproxy.google.com/~r/openthemagazine/" in pg:
        publishbox=soup.find('time',attrs={'class':'pub_date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0].strip()

    elif "pib.nic.in" in pg:
        publishbox=soup.find('span',attrs={'style':'float:right;font-size:80%;font-
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'id':'thd1'})
        if tags is not None:
            cls='Government of India'

    elif "feedproxy.google.com/~r/pluggd/" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "prokerala.com" in pg:
        publishbox=soup.find('meta',attrs={'itemprop':'datePublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)

    elif "www.rediff.com" in pg:
        publishbox=soup.find('meta',attrs={'itemprop':'datePublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=pg.split('/')[3].strip()
        if tags is not None:
            cls=tags

    elif "http://feeds.reuters.com/~r/reuters/" in pg:
        publishbox=soup.find('div',attrs={'class':'date_V9eGk'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('div',attrs={'class':'channel_4KD-f'})
        if tags is not None:
            a=tags.find('a')
            cls=a.text.strip()

    elif "saharasamay" in pg:
        publishbox=soup.find('td',attrs={'id':'printDate'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'Keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0]

    elif "www.santabanta.com" in pg:
        publishbox=soup.find('div',attrs={'id':'pubdate'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=pg.split('/')[3]
        if tags is not None:
            cls=tags.strip()

    elif "www.sify.com" in pg:
        publishbox=soup.find('meta',attrs={'name':'modified-date'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'Keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0].split(' ')[0].strip()

    elif "sikhsiyasat.net" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'property':'article:section'})
        if tags is not None:
            cls='politics'

    elif "www.sportskeeda.com" in pg:
        publishbox=soup.find('span',attrs={'class':'keeda-time-since'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
            print(datetime)
        cls='sports'

    elif "tech.firstpost.com" in pg:
        publishbox=soup.find('meta',attrs={'property':'og:updated_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls='technology'

    elif "www.mydigitalfc.com" in pg:
        publishbox=soup.find('span',attrs={'datatype':'xsd:dateTime'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        cls='finance'

    elif "www.telegraph.co.uk" in pg:
        publishbox=soup.find('meta',attrs={'name':'DCSext.articleFirstPublished'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'name':'keywords'})
        if tags is not None:
            cls=tags.get('content').split(',')[0].strip()

    elif "www.topnews.in" in pg:
        publishbox=soup.find('span',attrs={'class':'submitted'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('span',attrs={'class':'taxonomy'})
        if tags is not None:
            a=tags.find('a')
            cls=a.text.strip().split(' ')[0]

    elif "trak.in" in pg:
        publishbox=soup.find('meta',attrs={'property':'article:published_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'property':'article:section'})
        if tags is not None:
            cls=tags.get('content').strip()

    elif "valueresearchonline.com" in pg:
        publishbox=soup.find('p',attrs={'class':'dateline'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.text.strip())
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
            print(datetime)

    elif "zeenews.india.com" in pg:
        publishbox=soup.find('meta',attrs={'property':'og:updated_time'})
        if publishbox is not None:
            date=datefinder.find_dates(publishbox.get('content',None))
            for i in date:
                datetime='{:%Y-%m-%d}'.format(i)
        tags=soup.find('meta',attrs={'property':'article:tag'})
        if tags is not None:
            cls=tags.get('content').strip()

    tupl=(pg,datetime,cls)
    sheet.append(tupl)
    book.save(r'C:\Users\windows\Desktop\Bigdata dataset\book1.xlsx')
    print('done')
```

...

Output:



183	http://www.deccanherald.com/content/589294/mansur-award-pt-mani-prasad.html	2018-03-30	State
184	http://www.deccanherald.com/content/589298/bjp-leaders-offered-no-help.html	2018-03-30	State
185	http://www.deccanherald.com/content/589300/hasty-cancellation-leaves-passengers-fix.html	2018-03-30	State
186	http://www.deccanherald.com/content/589288/myneni-face-youzhny.html	2018-03-30	Sports
187	http://www.deccanherald.com/content/589289/chelsea-overcome-stoke.html	2018-03-30	Sports
188	http://www.deccanherald.com/content/589291/hunters-face-smashers.html	2018-03-30	Sports
189	http://www.thehindu.com/sport/tennis/%E2%80%98I-just-want-to-be-aggressive-work-up-the-crowd-to-ge	2017-01-01	Tennis
190	http://www.thehindu.com/sport/tennis/Myneni-draws-Youzhny-Ramkumar-opens-against-qualifier/article	2017-01-01	tennis
191	http://www.thehindu.com/sport/cricket/Shreyas-to-play-his-natural-instinctive-game/article16971904.ece?	2017-01-01	Sport
192	http://www.thehindu.com/sport/IOA-president%E2%80%99s-response-to-suspension/article16971898.ece?	2017-01-01	Sport
193	http://www.thehindu.com/sport/races/Arlene-should-repeat/article16971910.ece?utm_source=RSS_Feed&	2017-01-01	horse racing
194	http://www.thehindu.com/sport/cricket/Abhinav-Mukund-prefers-to-live-in-the-present/article16971909.e	2017-01-01	Cricket
195	http://www.thehindu.com/sport/cricket/Jharkhand-takes-on-Gujarat-for-a-place-in-final/article16971907.ec	2017-01-01	Sport
196	http://www.thehindu.com/sport/tennis/A-big-year-for-Divij-and-Purav-but-mostly-under-the-radar/article	2017-01-01	Tennis
197	http://www.thehindu.com/sport/cricket/A-match-up-between-equally-consistent-units/article16971901.ec	2017-01-01	domestic
198	http://www.thehindu.com/sport/races/Azzurro-wins/article16971912.ece?utm_source=RSS_Feed&utm_mec	2017-01-01	Races
199	http://www.thehindu.com/sport/tennis/Do-you-remember-when.../article16971880.ece?utm_source=RSS_	2017-01-01	tennis

In [ ]: