Import Modules

!pip install -r require.txt

In [ ]:

```
import pandas as pd
import requests as r
from bs4 import BeautifulSoup as bs
from selenium import webdriver
#driver = webdriver.Firefox(executable_path="D://geckodriver.exe")
import re
import time
from pymongo import MongoClient as client
from newspaper import Article
```

Times of india Scarper

Key Points

1. SVC of Mongo Db must be give as parameter, default is local MongoDB

2. Driver is needed in this model of Scraper

download geckodriver from https://github.com/mozilla/geckodriver/releases (https://github.com/mozilla/geckodriver/releases)

3.Provide path of geckodriver in model as paameter driver = "path of geckodriver"

4.Page from and Page to is integer value which scrape details from 1 - 3 (Ex: 1 - 3) is default

5.Content is a parameter takes values such as sports and business

6.Stores Data in DB name internrndd and collections is timesofindia

Sample Code :

```
d = times_of_india(content='sports',svc="mongodb://localhost:27017/",driver="D://geckodriver.exe",pagefrom=1,paget
d.load_to_database()
```

In [75]:

```
class times_of_india():
    def __init__(self,content,svc="mongodb://localhost:27017/",driver="D://geckodriver.exe"
        self.svc=svc
        self.content=content
        self.driver = webdriver.Firefox(executable_path=driver)
        self.driver.set_window_position(0, 0)
        self.pagefrom = int(pagefrom)
        self.pageto = int(pageto)
    ##GET SUBLINKS of Category
    def get_sub_links(self):
        print("Init-Sub-category")
        l=set()
        page = r.get("https://timesofindia.indiatimes.com/"+self.content)
        soup = bs(page.content,"html.parser")
        l=set()
        for i in soup.find("nav").find_all("a",href=True):
            if re.search("/"+self.content+"/*",i['href']):
                l.add("https://timesofindia.indiatimes.com"+i['href'])
        return l
    ### Scarpes Inner Link of SubLinks
    def get_full_links(self):
        print("Init-Get-Links")
        l=set()
        links = self.get_sub_links()
        for j in links:
            page = r.get(j)
            soup=bs(page.content,'html.parser')
            try:
                for i in soup.find("div",{"class":re.compile('main-content*')}).find_all("a

                    try:
                        if i['href'].split('.')[-1]=='cms':
                            if i['href'].split("/")[0]=="https:":
                                l.add(i['href'])
                            else:
                                l.add("https://timesofindia.indiatimes.com"+i['href'])
                    except:
                        pass
            except:
                pass
        return l
    ### FOR full links it starts fetch content and stored in DB
    def get_content(self):
        link = list(self.get_full_links())
        print("Started Fetching")
        l=[]
        for i in link[self.pagefrom:self.pageto]:
            try:
                print(i)
                d={}
                d['Link']=i
                page = r.get(i)
                d['Content'] = bs(page.content,'html.parser').find("div",{"class":re.compil
                self.driver.get(i)
                time.sleep(2)
                try:
                    d['Title']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div
                except:
                    d['Title']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div
```

```
                try:
                    d['Time']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div[
                except:
                    d['Time']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div[
                try:
                    d['Image']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div
                except:
                    d['Image']=self.driver.find_element_by_xpath("/html/body/div[2]/div/div
                l.append(d)
            except:
                pass

        return l
    ##Connector to DB
    def load_to_database(self):
        try:
            l=self.get_content()
            connect = client(self.svc)
            ##database name
            db=connect.internrndd
            ##table name
            col = db['timesofindia']
            col.insert_many(l)
            self.close_drive()
            return "Success"
        except:
            self.close_drive()
            return "Error Raised"

    def close_drive(self):
        self.driver.close()
```

In [73]:

```
d = times_of_india('sports')
d.load_to_database()
```

```
Init-Get-Links
Init-Sub-category
Started Fetching
https://timesofindia.indiatimes.com/sports/golf/top-stories/aditi-ashok-fini
shes-t-30-in-andalucia-open/articleshow/79478985.cms (https://timesofindia.i
ndiatimes.com/sports/golf/top-stories/aditi-ashok-finishes-t-30-in-andalucia
-open/articleshow/79478985.cms)
https://timesofindia.indiatimes.com/sports/wwe/top-stories/brock-lesnar-re-s
igns-with-wwe/articleshow/63704948.cms (https://timesofindia.indiatimes.com/
sports/wwe/top-stories/brock-lesnar-re-signs-with-wwe/articleshow/63704948.c
ms)
https://timesofindia.indiatimes.com/sports/wwe/top-stories/wwe-confirms-wres
tler-has-tested-positive-for-covid-19/articleshow/76407629.cms (https://time
sofindia.indiatimes.com/sports/wwe/top-stories/wwe-confirms-wrestler-has-tes
ted-positive-for-covid-19/articleshow/76407629.cms)
```

Out[73]:

```
'Success'
```

First Post Scraper

First Post Scraper

Key Points

1. SVC of Mongo Db must be give as parameter, default is local MongoDB

2. Driver is not needed in this model of Scraper

download geckodriver from https://github.com/mozilla/geckodriver/releases (https://github.com/mozilla/geckodriver/releases)

3.Provide path of geckodriver in model as paameter driver = "path of geckodriver"

4.Page from and Page to is integer value which scrape details from 1 - 3 (Ex: 1 - 3) is default

5.Content is a parameter takes values such as sports and business

6.Stores Data in DB name internrndd and collections is firstpost

Sample Code :

```
d = Firstpost(content='sports',svc="mongodb://localhost:27017/",pagefrom=1,pageto=8)
d.fetch()
```

In [3]:

```
class Firstpost():
    def __init__(self,content,limit=5,svc="mongodb://localhost:27017/",page_from=1,page_to=
        self.svc = svc
        self.cate=content
        self.limit=limit
        self.page_from = page_from
        self.page_to= page_to
    def fetch(self):
        j = self.page_from
        while(j<=self.page_to):
            if j==0:
                url="https://www.firstpost.com/category/{}/".format(self.cate)
            else:
                url="https://www.firstpost.com/category/"+self.cate+"/page/"+str(j)+"/"
            print(url)
            print(r.get(url).status_code)
            if r.get(url).status_code==200:

                page = r.get(url)
                soup = bs(page.content,'html.parser')
                l=[]
                for i in soup.find_all("div",{"class":"big-thumb"}):
                    d={}
                    try:
                        ##Get Sublinks
                        sub_url = i.find("a",href=True)['href']
                        if sub_url.split(":")[0]=="https" and sub_url.split(".")[-1]=="html
                            print(sub_url)
                            article = Article(sub_url)
                            article.download()
                            article.parse()
                            d['Link']=sub_url
                            #print(article.publish_date)
                            d['Date']=article.publish_date
                            #print(article.text)
                            d["Content"]=article.text
                            #print(article.title)
                            d["Title"]=article.title
                            ##print(article.top_image)
                            d['Image']=article.top_image
                            #print("\n"*3)
                            #print("#"*77)

                            l.append(d)


                        except:
                            pass
                    if self.load_to_database(l)==True:
                        print("Inserted")

            else:
                break
            print("&&"*60)
            j+=1
    def load_to_database(self,l):

            connect = client(self.svc)
            if connect:
```

```
            print("Connected")
        #DB name
        db=connect.internrndd
        #table firstpost
        col = db['firstpost']
        col.insert_many(l)

        return True
```

In [4]:

```
fifetch = Firstpost(content="sports",limit=1,page_from=8,page_to=16)
```

In [6]:

```
fifetch.fetch()
```

```
mpics-2022-called-off-due-to-covid-19-9063131.html (https://www.firstpost.
com/sports/another-test-event-for-beijing-winter-olympics-2022-called-off-
due-to-covid-19-9063131.html)
https://www.firstpost.com/sports/isl-2020-21-hyderabad-fcs-watertight-defe
nce-holds-bengaluru-fc-to-goalless-draw-9063051.html (https://www.firstpos
t.com/sports/isl-2020-21-hyderabad-fcs-watertight-defence-holds-bengaluru-
fc-to-goalless-draw-9063051.html)
https://www.firstpost.com/sports/bundesliga-borussia-dortmund-beaten-at-ho
me-by-cologne-bayern-munich-extend-lead-9063011.html (https://www.firstpos
t.com/sports/bundesliga-borussia-dortmund-beaten-at-home-by-cologne-bayern
-munich-extend-lead-9063011.html)
https://www.firstpost.com/sports/serie-a-cristiano-ronaldo-less-juventus-h
eld-by-benevento-inter-milan-score-three-past-sassuolo-9063001.html (http
s://www.firstpost.com/sports/serie-a-cristiano-ronaldo-less-juventus-held-
by-benevento-inter-milan-score-three-past-sassuolo-9063001.html)
https://www.firstpost.com/sports/formula-1-2020-dominant-lewis-hamilton-se
ts-track-record-at-bahrain-gp-to-clinch-98th-career-pole-9062961.html (htt
ps://www.firstpost.com/sports/formula-1-2020-dominant-lewis-hamilton-sets-
track-record-at-bahrain-gp-to-clinch-98th-career-pole-9062961.html)
https://www.firstpost.com/sports/premier-league-liverpool-frustrated-by-va
```

IndianExpress

Key Points

1. SVC of Mongo Db must be give as parameter, default is local MongoDB

2. Driver is not needed in this model of Scraper

download geckodriver from https://github.com/mozilla/geckodriver/releases (https://github.com/mozilla/geckodriver/releases)

3.Provide path of geckodriver in model as paameter driver = "path of geckodriver"

4.Page from and Page to is integer value which scrape details from 1 - 3 (Ex: 1 - 3) is default

5.Content is a parameter takes values such as sports and business

6.Stores Data in DB name internrndd and collections is indianexpress

Sample Code :

```
d = indianexpress(content='sports',svc="mongodb://localhost:27017/",pagefrom=1,pageto=8)
d.fetch()
```

In [35]:

```
class indianexpress():
    def __init__(self,content,limit=5,svc="mongodb://localhost:27017/",page_from=1,page_to=
        self.svc = svc
        self.cate=content
        self.limit=limit
        self.page_to=page_to
        self.page_from = page_from
    def fetch(self):
        j=self.page_from
        while(j<=self.page_to):
            if j==0:

                url="https://indianexpress.com/section/{}/".format(self.cate)
            else:
                url="https://indianexpress.com/section/"+self.cate+"/page/"+str(j)+"/"
            print(url)

            if r.get(url).status_code==200:
                print(r.get(url).status_code)
                page = r.get(url)
                soup = bs(page.content,'html.parser')
                l=[]
                for i in soup.find_all("div",{"class":re.compile("articles")}):
                    d={}
                    try:
                        sub_url = i.find("a",href=True)['href']
                        print(sub_url)
                        if sub_url.split(":")[0]=="https":

                            article = Article(sub_url)
                            article.download()
                            article.parse()
                            d['Link']=sub_url
                            #print(article.publish_date)
                            d['Date']=article.publish_date
                            #print(article.text)
                            d["Content"]=article.text
                            #print(article.title)
                            d["Title"]=article.title
                            ##print(article.top_image)
                            d['Image']=article.top_image
                            #print("\n"*3)
                            #print("#"*77)
                            l.append(d)
                    except:
                        pass

                self.load_to_database(l)
                print("Inserted")

            else:
                break
            print("&&"*66)
            j+=1
    def load_to_database(self,l):
        try:

            connect = client(self.svc)
            db=connect.internrndd
```

```
            col = db['indianexpress']
            col.insert_many(l)
            return "Success"
        except:

            return "Error Raised"
```

In [36]:

```
indian = indianexpress("sports",limit=5)
```

In [37]:

```
print(indian.fetch())
```

```
https://indianexpress.com/section/sports/page/1/ (https://indianexpress.co
m/section/sports/page/1/)
200
https://indianexpress.com/article/sports/tennis/wta-2021-season-schedule-o
utside-australia-venue-7091453/ (https://indianexpress.com/article/sports/
tennis/wta-2021-season-schedule-outside-australia-venue-7091453/)
https://indianexpress.com/article/sports/cricket/will-pucovski-india-vs-au
stralia-test-series-comments-7091438/ (https://indianexpress.com/article/s
ports/cricket/will-pucovski-india-vs-australia-test-series-comments-709143
8/)
https://indianexpress.com/article/sports/cricket/pakistan-covid-19-positiv
e-tests-new-zealand-training-denied-7091420/ (https://indianexpress.com/ar
ticle/sports/cricket/pakistan-covid-19-positive-tests-new-zealand-training
-denied-7091420/)
https://indianexpress.com/article/sports/cricket/west-indies-mcc-spirit-of
-cricket-award-win-7091409/ (https://indianexpress.com/article/sports/cric
ket/west-indies-mcc-spirit-of-cricket-award-win-7091409/)
https://indianexpress.com/article/sports/cricket/india-vs-australia-1st-t2
0i-live-cricket-score-online-7091290/ (https://indianexpress.com/article/s
```

MoneyControl

Key Points

1. SVC of Mongo Db must be give as parameter, default is local MongoDB ¶

2. Driver is not needed in this model of Scraper

download geckodriver from https://github.com/mozilla/geckodriver/releases (https://github.com/mozilla/geckodriver/releases)

3.Provide path of geckodriver in model as paameter driver = "path of geckodriver"

4.Page from and Page to is integer value which scrape details from 1 - 3 (Ex: 1 - 3) is default

5.Content is a parameter takes values such as market and mutual-funds

6.Stores Data in DB name internrndd and collections is firstpost

Sample Code :

```
d = moneycontrol(content='sports',svc="mongodb://localhost:27017/",pagefrom=1,pageto=8)
d.load_to_database()
```

In [65]:

```
class moneycontrol():
    def __init__(self,content,limit=5,svc="mongodb://localhost:27017/",page_from=1,page_to=
        self.svc = svc
        self.cate=content
        self.limit=limit
        self.page_to=page_to
        self.page_from = page_from

    def fetch(self):
        j=self.page_from
        while(j<=self.page_to):
            if j==0:

                url="https://www.moneycontrol.com/news/business/{}/".format(self.cate)
            else:
                url="https://www.moneycontrol.com/news/business/"+self.cate+"/page-"+str(j)
            print(url)

            if r.get(url).status_code==200:
                print(r.get(url).status_code)
                page = r.get(url)
                soup = bs(page.content,'html.parser')
                l=[]
                for i in soup.find_all("li",{"class":"clearfix"}):
                    try:
                        if re.search("https://www.moneycontrol.com/news/",i.find("a",href=T
                            d={}
                            sub_url=i.find("a",href=True)['href']
                            art=Article(sub_url)
                            art.download()
                            art.parse()
                            d["Link"]=sub_url
                            d['Title']=art.title
                            d['Content']=art.text
                            d['Image']=art.top_image
                            try:
                                d['Date']=bs(art.html,'html.parser').find("div","article_sc
                            except:
                                pass
                            l.append(d)


                    except:
                        pass


                self.load_to_database(l)
                print("Inserted")

            else:
                pass
            print("&&"*66)
            j+=1
    def load_to_database(self,l):
        try:

            connect = client(self.svc)
            db=connect.internrndd
            col = db['moneycontrol']
```

```
            col.insert_many(l)
            return "Success"
        except:

            return "Error Raised"
```

In [66]:

```
money = moneycontrol("mutual-funds")
money.fetch()
```

```
https://www.moneycontrol.com/news/business/mutual-funds/page-1/ (https://ww
w.moneycontrol.com/news/business/mutual-funds/page-1/)
200
Inserted
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
https://www.moneycontrol.com/news/business/mutual-funds/page-2/ (https://ww
w.moneycontrol.com/news/business/mutual-funds/page-2/)
200
Inserted
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&
```

Similiar approach for Times of India without Selenium

In [67]:

```
class times_of_india1():
    def __init__(self,content,limit=5,svc="mongodb://localhost:27017/",page_from=1,page_to=
        self.svc = svc
        self.cate=content
        self.limit=limit
        self.page_to=page_to
        self.page_from = page_from

    def fetch(self):
        j=self.page_from
        while(j<=self.page_to):
            if j==0 or j==1:

                url="https://timesofindia.indiatimes.com/sports/"
            else:
                url="https://timesofindia.indiatimes.com/business/india-business/"+str(j)
            print(url)

            if r.get(url).status_code==200:
                print(r.get(url).status_code)
                page = r.get(url)
                soup = bs(page.content,'html.parser')
                l=[]
                for i in soup.find("div",{"id":"c_articlelist_stories_2"}).find_all("a"):
                    try:
                        if(i['href'].split(".")[-1]=="cms"):
                                d={}
                                sub_url="https://timesofindia.indiatimes.com"+i['href']
                                art=Article(sub_url)
                                art.download()
                                art.parse()
                                d["Link"]=sub_url
                                d['Title']=art.title
                                d['Content']=art.text
                                d['Image']=art.top_image
                                d['Date']=art.publish_date
                                print(art.publish_date)
                                l.append(d)


                    except:
                        pass


                self.load_to_database(l)
                print("Inserted")

            else:
                pass
            print("&"*66)
            j+=1
    def load_to_database(self,l):
        try:

            connect = client(self.svc)
            db=connect.internrndd
            col = db['timesofindia']
            col.insert_many(l)
            return "Success"
```

```
        except:

            return "Error Raised"
```

In [68]:

```
tr = times_of_india1("business",page_from=1,page_to=3)
```

GUI

In [ ]:

```
import tkinter
from PIL import Image, ImageTk
from tkinter import ttk
top = tkinter.Tk()
top.geometry('1200x750')
top.configure(background="black")
def printf():
    if content.get()=="Money Control":
        print(rvar.get())
    if path.get()=="":
        tkinter.messagebox.showerror("showerror", "Path Not Found")
    print(content.get())
    print(path.get())
    print(rvar.get())
    print(svc.get())
    print(pagefrom.get())
    print(pageto.get())
class Example(tkinter.Frame):
    def __init__(self, master, *pargs):
        tkinter.Frame.__init__(self, master, *pargs)



        self.image = Image.open("bg2.png")
        self.img_copy= self.image.copy()


        self.background_image = ImageTk.PhotoImage(self.image)

        self.background = tkinter.Label(self, image=self.background_image)
        self.background.pack(fill="both", expand=True)
        self.background.bind('<Configure>', self._resize_image)

    def _resize_image(self,event):

        new_width = event.width
        new_height = event.height

        self.image = self.img_copy.resize((new_width, new_height))

        self.background_image = ImageTk.PhotoImage(self.image)
        self.background.configure(image =  self.background_image)



e = Example(top)
e.pack()


uname = tkinter.Label(top, text = "Websites",bg="#0a0a29",fg="#3b57ab",font=('courier', 15,

#creating label
password = tkinter.Label(top, text = "Driver Path ",bg="#0a0a29",fg="#3b57ab",font=('courie

radiovalue = tkinter.Label(top, text = "Content ",bg="#0a0a29",fg="#3b57ab",font=('courier'
path = tkinter.StringVar()
path.set("Provide Driver Path")
svc = tkinter.StringVar()
```

```
svc.set("mongodb://localhost:27017/")
rvar = tkinter.StringVar()

pagefrom = tkinter.StringVar()
pagefrom.set("Ex : 1")
pageto = tkinter.StringVar()
pageto.set("Ex : 3")
r1 = tkinter.Radiobutton(top, text='Sports', variable=rvar, value='sports',bg="#0a0a29",fg=
r2 = tkinter.Radiobutton(top, text='Business', variable=rvar, value='business',bg="#0a0a29"
mongosvc = tkinter.Label(top,text="MongoDB SVC",bg="#0a0a29",fg="#3b57ab",font=('courier',
e2 = tkinter.Entry(top, width = 29,textvariable=path,font = ('courier', 15)).place(x =350,
e3 = tkinter.Entry(top, width = 29,textvariable=svc,font = ('courier', 15)).place(x =350, y
page = tkinter.Label(top,text="Page",bg="#0a0a29",fg="#3b57ab",font=('courier', 15,"bold"))
page1 = tkinter.Entry(top, width = 12,textvariable=pagefrom,font = ('courier', 15)).place(x
page2 = tkinter.Entry(top,width=12, textvariable=pageto,font = ('courier', 15)).place(x=550
sbmitbtn = tkinter.Button(top, text = "Submit",activebackground = "pink", activeforeground

n = tkinter.StringVar()
content = ttk.Combobox(top, width = 27,
                           textvariable = n,font = ('courier', 15))


content['values'] = ("Times of India","First Post","Indian Express","Money Control")


content.place(x = 350, y = 250)

content.current(0)

top.mainloop()
```