▾ IST 462 Homework #5

Instructions

- Complete all 5 questions in this assignment.
- It is always better to attempt a problem as partial credit may be granted.

▾ Questions

▾ Web Scraping: HTML

▾ (1a) Choose a web site and retrieve the HTML data using urllib.

- If it doesn't return a page (perhaps because it uses cookies or other interactions to actually return content), try a different site.
  - **NOTE:** Some sites (e.g. Google) blocks webscrapers so your code will run once or twice, but then fail. Please be conscious of this.

(1b) Use BeautifulSoup to parse the object (like we did in class).

NOTE: The BeatifulSoup object will be used in (2) and (3).

[20 points]

```
from urllib import request
from bs4 import BeautifulSoup
from pprint import pprint


# Enter your code here, printing relevant answers to console:

cnn = "https://www.cnn.com/us"
#url im using is the cnn home page for the us
response = request.urlopen(cnn)
#using url open to call the cnn url and then return a response on if the url was found



request.urlopen(cnn)
#looking the HTTP response

    <http.client.HTTPResponse at 0x7a293765e5c0>


response.code
#looking at the response code, 200 means a successful http request

    200
```

```
cnn_html = response.read().decode('utf8')
cnn_html
#reads the content of the HTTP response and decodes it as a utf-8 string.
```

```
'  <!DOCTYPE html>\n<html lang="en" data-uri="cms.cnn.com/_pages/cl9yme9db000026nh7j9ix9et@published" data-l
ayout-uri="cms.cnn.com/_layouts/layout-no-rail/instances/us-v1@published">\n  <head><style>body,h1,h2,h3,h4,
h5{font-family:cnn_sans_display,helveticaneue,Helvetica,Arial,Utkal,sans-serif}h1,h2,h3,h4,h5{font-weight:70
0}:root{--theme-primary:#cc0000;--theme-background:#0c0c0c;--theme-divider:#404040;--theme-copy:#404040;--th
eme-copy-accent:#e6e6e6;--theme-copy-accent-hover:#ffffff;--theme-icon-color:#e6e6e6;--theme-icon-color-hove
r:#ffffff;--theme-ad-slot-background-color:#0c0c0c;--theme-ad-slot-text-color:#b1b1b1;--theme-ad-slot-text-h
over:#ffffff;--theme-font-family:cnn_sans_display,helveticaneue,Helvetica,Arial,Utkal,sans-serif;--theme-sea
rchbox-border:#b1b1b1;--theme-copy-follow:#ffffff;--theme-article-spacing-top:0px;--theme-link-color-hover:#
```

```
htmlsoup = BeautifulSoup(cnn_html, 'html.parser')
#using beutifulsoup to parse down the HTML document, allows us to navigate and manipulate the HTML content more easily
type(htmlsoup)
#checking the type of html soup to ensure that the html doc has been parse downed
```

```
bs4.BeautifulSoup
```

```
htmlsoup.title.string
#looking at the title string of the html document
```

```
'US | CNN'
```

```
firsttitle = htmlsoup.title
firsttitle.string
```

```
'US | CNN'
```

```
type(firsttitle)
#type of the title is element
```

```
bs4.element.Tag
```

```
firsttitle.name
#retrieve tags name
```

```
'title'
```

```
firsttitle.get_text()
#retrives tags text
```

```
'US | CNN'
```

▼ (2): Count the number of hyperlink URLs (which begin with http) in the <u>head</u> & <u>body</u> tags.

[20 points]

```
head_links = htmlsoup.head.find_all('a', href=True)
#finding all the anchor tags (a) within the head section of the html document
body_links = htmlsoup.body.find_all('a', href=True)
#finding all of the anchor tags (a) within the body section of the html document
```

```
#finding all of the anchor tags (a) within the body section of the html document


http_head_links_count = 0
#setting head links count to 0
http_body_links_count = 0
#setting body links count to 0

for link in head_links:
  if link['href'].startswith('http'):
#href is a tag within html so looking for where the link tag starts with http, basically where there is a link within the head
    http_head_links_count += 1
#adding one to the head_links_count where this is applicable


for link in body_links:
  if link['href'].startswith('http'):
    http_body_links_count += 1
#doing the same thing as above but for the body tags


print("Number of 'http' links in <head>: ", http_head_links_count)
print("Number of 'http' links in <body>: ", http_body_links_count)


    Number of 'http' links in <head>:  0
    Number of 'http' links in <body>:  328
```

(3a): Generate a unique list of hyperlink URLs (which begin with http) from the <u>body</u> tag and convert it into a pandas <u>series</u> object.

- If your website body has no hyperlinks, please select a new website and repeat question (1).

(3b): Create a function which parses the domain name (parent website) from a hyperlink URL.

- for example:
  - **URL:** http://www.espn.com/mens-college-basketball/story/_/id/26281956/know-every-team-bracket
  - **domain name:** espn.com

    - or www.espn.com

(3c): Create a new pandas series composed of domain names.

- Consider using the **series.map()** method, applying your function from (3b) over the series in (3a)

[20 points]

```
import pandas as pd


#3a
links = []
#creating an empty list for links
for tag in body_links:
#iterate over the html anchor elements in body_links
```

```
    href = tag.get("href")
#find the value of the href attribute which contains a url
    if href and href.startswith('http'):
#if the url starts with http and does not contain an empty value
        links.append(href)
#add that link to the links list

pandas_series = pd.Series(links)
#make a pandas series of the links and store it as pandas_series
```

```
type(pandas_series)
#checking the type to make sure it stored as a pandas series and not an numpy array
```

```
    pandas.core.series.Series
```

```
print("Unique Hyperlinks URLs in the body tag: ")
print(pandas_series.unique())
#printing the series, making sure to only print unique values (no duplicate links)
```

⤓

```
https://abc13.com/sh-288-crash-3-year-old-child-killed-deadly-rollover-right-between-driver-and-passenger/13991081/
'https://www.kshb.com/news/local-news/kearney-mans-skeleton-crew-creates-spooky-halloween-delight-for-neighbors'
'https://www.cbsnews.com/minnesota/news/one-minnesota-family-has-fun-with-taylor-swift-themed-halloween-decor/'
'https://www.cbsnews.com/detroit/news/video-man-arrested-after-pointing-laser-at-police-helicopter-in-detroit/'
'https://www.kptv.com/2023/10/30/officers-campus-after-explosion-osu/'
'https://www.fox13now.com/news/local-news/man-stuck-for-12-hours-after-knee-gets-stuck-in-crack-while-climbing'
'https://www.kcra.com/article/plane-diverted-sacramento-airport-following-possible-engine-failure/45684379'
'https://www.ksl.com/article/50763322/go-forward-when-you-can-101-year-old-woman-enrolls-at-byu-idaho'
'https://www.azfamily.com/2023/10/30/phoenix-neighbors-hope-reunite-family-with-mystery-urn-found-their-gated-community/'
'https://abc7.com/anaheim-officer-involved-shooting-woman-killed-bb-gun-pellet/13990407/'
'https://www.cbsnews.com/sacramento/news/sacramento-police-introduce-mobile-station-to-foster-community-engagement/'
'https://www.kitv.com/news/they-managed-to-stab-me-ten-times-in-total-waikiki-stabbing-victim-recovering-from-attack/article_1304ead6-7643-11ee-b048-7fe26afbe004.html
'https://www.koat.com/article/atrisco-heritage-football-player-recovering/45684682'
'https://www.eastidahonews.com/2023/10/why-a-former-deputy-cut-off-an-outlaws-leg-and-the-controversy-surrounding-a-notorious-gunfight/'
'https://www.cnn.com/travel/news' 'https://www.cnn.com'
'https://www.cnn.com/terms' 'https://www.cnn.com/privacy'
'https://warnermediaprivacy.com/policycenter/b2c/WMNS'
'https://www.cnn.com/accessibility' 'https://www.cnn.com/newsletters'
'https://www.cnn.com/transcripts']
```

```python
#3b: creating a function to extract the domain

def extract_domain(url):
#creating the function with the input of a url
  if "://" in url:
#looking for :// within the url so then we can split based on that
    url = url.split('://')[1]
# using :// as the separator to split the url and taking the [1] which would be the actual domain
#thats stored as the whole url. basically is ifts http://google.com/news that changes it to google.com/news
  sections = url.split('/')
#next we want to take anything after the slash and get rid of it
#split the url from the first slash
  domain = sections[0]
#take the value on the left store as domain
#basicially splitting google.com/news and only taking google.com

  if domain.startswith('www.'):
#getting rid of the www.; looking for it within the domain and if its there
    domain = domain[4:]
#take the 4th value in the string to the end

  return domain
```

```python
#3c
domains = pandas_series.map(extract_domain)
#calling the extract domain function on the pandas series
#storing this as domains

domainNames = pd.DataFrame(domains.unique(), columns=["Domains"])
#then taking the domains (only the unique ones) and making it into a pandas data frame with the column name Domains
domainNames.index += 1
#setting the index (count) to start at 1

domainNames
#printing the pandas dataframe
```

11	max.com
12	store.cnn.com
13	coupons.cnn.com
14	careers.wbd.com
15	wlos.com
16	wral.com
17	wsvn.com
18	turnto10.com
19	wfsb.com
20	wmtw.com
21	cbsnews.com
22	wsmv.com
23	auburnpub.com
24	albanyherald.com
25	kctv5.com
26	kmov.com
27	wlwt.com
28	wcfcourier.com
29	stltoday.com
30	wnem.com
31	abc13.com
32	kshb.com
33	kptv.com
34	fox13now.com

35	kcra.com
36	ksl.com
37	azfamily.com
38	abc7.com
39	kitv.com
40	koat.com

| 42 | warnermediaprivacy.com |

Web Scraping: XML

(4a) Choose an RSS feed site and retrieve the XML data using urllib.

- If it doesn't return a page (perhaps because it uses cookies or other interactions to actually return content), try a different site.
- Here is a list of RSS Feeds to get you started:
  - **https://rss.com/popular-rss-feeds/**

(4b) Use ElementTree to parse the object (like we did in class).

(4c): Count the number of Top-level <u>parent tags</u> & their number of <u>child</u> tags.

- Top-level means that a tag has **no** parent tag
- Only count direct children to these top level tags
  - Do not count grandchildren, etc

[20 points]

```
import urllib.request
import xml.etree.ElementTree as etree
import io


# Enter your code here, printing relevant answers to console:

xml = "https://moxie.foxnews.com/google-publisher/world.xml"
#this is my xml link

xmlstring = urllib.request.urlopen(xml).read().decode('utf8')
#reading in the url as utf8 (similar to how we did w the html doc)
len(xmlstring)
#finding the length of the xml string

    161798


xmlfile = io.StringIO(xmlstring)
#string io to open it up
tree = etree.parse(xmlfile)
#using etree.parse to parse down the xml file
```

```
type(tree)
#checking the type
```

```
    xml.etree.ElementTree.ElementTree
```

```
root = tree.getroot()
#ask the tree (variable) to get the root element of the XML tree
type(root)
print(root)
#root is an instance of the element class
```

```
    <Element 'rss' at 0x7a2925209b20>
```

```
root.tag
#shows that the XML document follows the format of an RSS feed.
```

```
    'rss'
```

```
root.attrib
#root attribute, shows that the RSS feed adheres to the RSS 2.0 specification
```

```
    {'version': '2.0'}
```

```
root.text
#text is empty
```

```
    '\n  '
```

```
for child in root:
    print(child)
#for the child it prints the element channell for the root tag
```

```
    <Element 'channel' at 0x7a29252085e0>
```

```
firstchild = root[0]
#first child is the first element at the 0th or 1st position under the root element
firstchild
```

```
    <Element 'channel' at 0x7a29252085e0>
```

```
firstgrandchild = firstchild[0]
#following the same logic as above but for grandchildren, first grandchild is element
firstgrandchild
```

```
    <Element 'title' at 0x7a292560b7e0>
```

```
firstgrandchild.attrib
```

```
    {}
```

```
firstgrandchild.text
#text of the girstgrandchild
```

```
'Latest World News on Fox News'
```

```
firstchild.findall('item')
#find all XML elements with the tag name 'item' within the 'firstchild' element
```

```
[<Element 'item' at 0x7a292560b290>,
 <Element 'item' at 0x7a29256142c0>,
 <Element 'item' at 0x7a2925614bd0>,
 <Element 'item' at 0x7a2925615760>,
 <Element 'item' at 0x7a2925616110>,
 <Element 'item' at 0x7a2925616b10>,
 <Element 'item' at 0x7a29256176f0>,
 <Element 'item' at 0x7a29254340e0>,
 <Element 'item' at 0x7a2925434860>,
 <Element 'item' at 0x7a29254351c0>,
 <Element 'item' at 0x7a2925435b70>,
 <Element 'item' at 0x7a29254367f0>,
 <Element 'item' at 0x7a2925437240>,
 <Element 'item' at 0x7a2925437d30>,
 <Element 'item' at 0x7a292545c860>,
 <Element 'item' at 0x7a292545d350>,
 <Element 'item' at 0x7a292545dc60>,
 <Element 'item' at 0x7a292545e7a0>,
 <Element 'item' at 0x7a292545f0b0>,
 <Element 'item' at 0x7a292545f920>,
 <Element 'item' at 0x7a292546c360>,
 <Element 'item' at 0x7a292546ce00>,
 <Element 'item' at 0x7a292546d710>,
 <Element 'item' at 0x7a292546e110>,
 <Element 'item' at 0x7a292546eb10>]
```

```
itemlist = firstchild.findall('item')
#storing this as itemlist
len(itemlist)
#finding the length of itemlist which is the count of how many element items there are within the first child
```

```
25
```

```
[i.tag for i in itemlist[0]]
#goes through the item list and extracts the tag names of these elements. creates a list of the tag names
```

```
['link',
 'guid',
 'title',
 'description',
 '{http://purl.org/rss/1.0/modules/content/}encoded',
 '{http://search.yahoo.com/mrss/}content',
 'category',
 'category',
 'category',
 'category',
 'category',
 'category',
 'category',
 'category',
 'category',
 'pubDate']
```

```python
for indx, element in enumerate(firstitem):
#looping through the first item and finding the element while enumerating them (giving them indexes)
    print(indx, element.tag, element.attrib, element.text)
#printing the index, element tag (link, guid, etc), the attribute, and then the text for the first item
```

```
0 link {} https://www.foxnews.com/world/un-tells-world-governments-not-repatriate-escaped-north-korean-refugees
1 guid {'isPermaLink': 'true'} https://www.foxnews.com/world/un-tells-world-governments-not-repatriate-escaped-north-korean-refugees
2 title {} UN tells world governments not to repatriate escaped North Korean refugees
3 description {} The United Nations is calling on member countries to refrain from repatriating escaped North Korean refugees.
4 {http://purl.org/rss/1.0/modules/content/}encoded {} <p>The United Nations is calling on world governments to refrain from repatriating <a href="https://www.foxnews.
5 {http://search.yahoo.com/mrss/}content {'url': 'https://a57.foxnews.com/static.foxnews.com/foxnews.com/content/uploads/2023/01/931/523/North-korea-covid-restrictions
6 category {'domain': 'foxnews.com/metadata/dc.identifier'} 57d1c085-cb9a-5921-993d-b0b1666025ce
7 category {'domain': 'foxnews.com/metadata/prism.channel'} fnc
8 category {'domain': 'foxnews.com/metadata/dc.source'} Fox News
9 category {'domain': 'foxnews.com/taxonomy'} fox-news/world/united-nations
10 category {'domain': 'foxnews.com/taxonomy'} fox-news/world/conflicts/north-korea
11 category {'domain': 'foxnews.com/taxonomy'} fox-news/world/world-regions/china
12 category {'domain': 'foxnews.com/taxonomy'} fox-news/world/world-regions/asia
13 category {'domain': 'foxnews.com/taxonomy'} fox-news/world/disasters/refugees
14 category {'domain': 'foxnews.com/section-path'} fox-news/world
15 category {'domain': 'foxnews.com/content-type'} article
16 pubDate {} Tue, 31 Oct 2023 11:35:29 -0400
```

```python
[{i.tag: i.text} for i in itemlist[0]]
#creates a list of dictionaries
```

```
[{'link': 'https://www.foxnews.com/world/un-tells-world-governments-not-repatriate-escaped-north-korean-refugees'},
 {'guid': 'https://www.foxnews.com/world/un-tells-world-governments-not-repatriate-escaped-north-korean-refugees'},
 {'title': 'UN tells world governments not to repatriate escaped North Korean refugees'},
 {'description': 'The United Nations is calling on member countries to refrain from repatriating escaped North Korean refugees.'},
 {'{http://purl.org/rss/1.0/modules/content/}encoded': '<p>The United Nations is calling on world governments to refrain from repatriating <a
href="https://www.foxnews.com/category/world/conflicts/north-korea" target="_blank">North Korean refugees.</a></p> <p><a
href="https://www.foxnews.com/category/world/united-nations" target="_blank">U.N. Deputy High Commissioner for Human Rights Nada al-Nashif</a> said Monday that member
countries must not send North Korean escapees back to the hermitic nation.</p> <p>She made the comment in a meeting with South Korean Unification Ministry official
Kang Jong-suk, according to Yonhap News Agency.</p> <p><a href="https://www.foxnews.com/world/china-forcibly-repatriates-hundreds-of-escaped-north-koreans-us-
ambassador-warns" target="_blank"><strong>CHINA FORCIBLY REPATRIATES HUNDREDS OF ESCAPED NORTH KOREANS, US AMBASSADOR WARNS</strong></a></p> <p><a
href="https://www.foxnews.com/category/world/world-regions/china" target="_blank">China has been accused</a> of sending back hundreds of escaped North Koreans over the
last few months.</p> <p>Elizabeth Salmón, the special rapporteur on the situation of human rights in the Democratic People's Republic of Korea, has warned that those
who are returned to North Korea after escape "would be subjected to torture, cruel, inhuman or degrading treatment and punishment and other serious human rights
violations(…) and even execution," according to NK News.</p> <p>Some refugees forcibly repatriated to the regime state have lived in the country for decades.</p> <p><a
href="https://www.foxnews.com/world/north-korean-foreign-minister-defends-russia-arms-sales-groundless-criticism" target="_blank"><strong>NORTH KOREAN FOREIGN MINISTER
DEFENDS RUSSIA ARMS SALES AGAINST \'GROUNDLESS CRITICISM\'</strong></a></p> <p>Ambassador Julie Turner, who was sworn in as U.S. special envoy for North Korean human
rights issues this month, first addressed the issue Friday.</p> <p>"I am gravely concerned by recent and credible reports that the PRC repatriated large numbers of
North Koreans, including as recently as last week," she said during an event at George Washington University\'s Institute for Korean Studies.</p> <p><a
href="https://www.foxnews.com/category/world/world-regions/south-korea" target="_blank">South Korean authorities</a> reported earlier this month that a small, wooden
boat of North Korean escapees was rescued near the country\'s sea border.</p> <p>The four refugees — reportedly a man and three women — claimed to be from the same
family.</p> <p>South Korea reports providing sanctuary for approximately 30,000 North Korean citizens who fled their oppressive conditions under <a
href="https://www.foxnews.com/category/person/kim-jong-un" target="_blank">supreme leader Kim Jong Un\'s regime.</a></p>'},
 {'{http://search.yahoo.com/mrss/}content': None},
 {'category': '57d1c085-cb9a-5921-993d-b0b1666025ce'},
 {'category': 'fnc'},
 {'category': 'Fox News'},
 {'category': 'fox-news/world/united-nations'},
 {'category': 'fox-news/world/conflicts/north-korea'},
 {'category': 'fox-news/world/world-regions/china'},
 {'category': 'fox-news/world/world-regions/asia'},
 {'category': 'fox-news/world/disasters/refugees'},
 {'category': 'fox-news/world'},
 {'category': 'article'},
 {'pubDate': 'Tue, 31 Oct 2023 11:35:29 -0400'}]
```

```
#4c counting child tags
child_counts = {}
#starting an empty dictionary


for e in root:
#for each element in the root
    child_count = len(e)
#the count is the length of all the elements (counting the elemtns)
    child_counts[e.tag] = child_count
#adds an entry with the key of the entry being the tag name of the current element and the value is the count of direct child tags


for tag, count in child_counts.items():
    print(f"The Tag: {tag}, has {count} Direct Child Tags")
#printing the results

    The Tag: channel, has 34 Direct Child Tags


#4c counting parent tags
parent_tag_count = 0
#setting parent tag count to 0

for e in root:
#for each element in the root
    parent_tag_count += 1
#add one to the the parent tag count


print("Number of top-level parent tags:", parent_tag_count)
#print the output

#i think there is only one parent tag becuase it contains all of the articles and datas and then there are many child and grandchildren tags

    Number of top-level parent tags: 1


element.tag
#this is the parent tag

    'channel'
```

(5a): Convert all of the items of your RSS XML into a pandas dataframe.

- An item in an RSS feed begins with an item tag < item > ... < /item >
- Each row represents 1 item
- Each column represents 1 child-tag to the parent <u>item</u> tag:
    - i.e.: Title, Link, PubDate, Category
        - Ignore children of these tags

(5b): Provide summary statistics about the dataframe.

- Consider using the .describe() method

[20 points]

▾ FOR EXAMPLE

https://api.foxsports.com/v1/rss?partnerKey=zBaFxRyGKCfxBagJG9b8pqLyndmvo7UU&tag=soccer image.png

```
# Enter your code here, printing relevant answers to console:
articles = []
#empty list of articles
for article in itemlist:
#for each article in the itemList
    articles.append({i.tag: i.text for i in article})
#append an article with the tag name as the keys and the text as the values
```

```
articles[:5]
#pulling first 5 articles
```

who are returned to North Korea after escape "would be subjected to torture, cruel, inhuman or degrading treatment and punishment and other serious human rights violations(…) and even execution," according to NK News.</p> <p>Some refugees forcibly repatriated to the regime state have lived in the country for decades.</p> NORTH KOREAN FOREIGN MINISTER DEFENDS RUSSIA ARMS SALES AGAINST \'GROUNDLESS CRITICISM\'</p> <p>Ambassador Julie Turner, who was sworn in as U.S. special envoy for North Korean human rights issues this month, first addressed the issue Friday.</p> <p>"I am gravely concerned by recent and credible reports that the PRC repatriated large numbers of North Koreans, including as recently as last week," she said during an event at George Washington University\'s Institute for Korean Studies.</p> <p>South Korean authorities reported earlier this month that a small, wooden boat of North Korean escapees was rescued near the country\'s sea border.</p> <p>The four refugees — reportedly a man and three women — claimed to be from the same family.</p> <p>South Korea reports providing sanctuary for approximately 30,000 North Korean citizens who fled their oppressive conditions under supreme leader Kim Jong Un\'s regime.</p>',
    '{http://search.yahoo.com/mrss/}content': None,
    'category': 'article',
    'pubDate': 'Tue, 31 Oct 2023 11:35:29 -0400'},
  {'link': 'https://www.foxnews.com/world/russias-defense-minister-accuses-us-fueling-tensions-warns-clash-between-nuclear-powers',
    'guid': 'https://www.foxnews.com/world/russias-defense-minister-accuses-us-fueling-tensions-warns-clash-between-nuclear-powers',
    'title': 'Russia's defense minister accuses US of fueling tensions, warns of clash between nuclear powers',
    'description': 'Russian Defense Minister Sergei Shoigu claimed the U.S. and its Western allies "deliberately undermined and destroyed" international security.',
    '{http://purl.org/rss/1.0/modules/content/}encoded': '<p>Russia's minister of defense on Monday blamed the United States of fueling tensions across the globe, claiming that the West's "steady escalation" threatens a military conflict between nuclear powers. </p> <p>Sergei Shoigu delivered remarks at a defense forum in Beijing, an annual event that China uses to promote military diplomacy and is attended by representatives from dozens of countries, including the U.S.</p> <p>"Washington for years has deliberately undermined and destroyed the foundations of international security and strategic stability, including the system of arms control agreements," Shoigu said at the Xiangshan Forum.</p> <p>Shoigu said that the U.S. and its Western allies are threatening Russia through NATO trying to expand its footprint in the Asia-Pacific under the pretense of seeking dialogue and collaboration with regional countries.</p> <p>KREMLIN SHORTENS QUARANTINE BEFORE MEETING PUTIN AS PRESIDENIAL ELECTIONS APPROACH</p> <p>"The Western line of steady escalation of the conflict with Russia carries the threat of a direct military clash between nuclear powers, which is fraught with catastrophic consequences," Shoigu said.</p> <p>Moscow, however, remained open to negotiations about the war in Ukraine if the conditions were right, Shoigu said, without elaborating.</p> <p>Shoigu's speech followed that of Zhang Youxia, China's second-ranking military official and vice

```
    'title': 'Kremlin shortens quarantine before meeting Putin as presidential elections approach',
    'description': 'The Kremlin is reportedly reducing quarantine periods for individuals meeting in-person with Russian President Vladimir Putin ahead of presidential
elections.',
    '{http://purl.org/rss/1.0/modules/content/}encoded': '<p>The <a href="https://www.foxnews.com/category/world/world-regions/russia" target="_blank">Russian
government</a> is relaxing its quarantine requirements to meet with President Vladimir Putin as presidential elections approach.</p> <p>Individuals who wish to <a
href="https://www.foxnews.com/category/world/personalities/vladimir-putin" target="_blank">meet with Putin in person</a> are required to quarantine for a period of
time before their interaction.\xa0</p> <p>The Kremlin announced that the quarantine period is being lowered from two weeks to five days as presidential election season
draws closer, according to The Moscow Times.</p> <p><a href="https://www.foxnews.com/world/putin-oversees-russian-military-drill-simulating-massive-retaliatory-
nuclear-strike-reports" target="_blank"><strong>PUTIN OVERSEES RUSSIAN MILITARY DRILL SIMULATING \'MASSIVE RETALIATORY NUCLEAR STRIKE\': REPORTS</strong></a></p>
<p>The reduced restrictions are most likely related to the upcoming March 2024 presidential elections — Putin is expected to win his fifth term.</p> <p>Putin is one of
the few world leaders still screening his meetings for <a href="https://www.foxnews.com/category/health/infectious-disease/coronavirus" target="_blank">COVID-19
concerns.</a></p> <p>The Russian president regularly holds government meetings over video calls in order to skip quarantine protocol for high-ranking officials. The
entire Russian government has reportedly not met in person since March 2020.</p> <p>Putin has traveled extensively recently and <a
href="https://www.foxnews.com/world/xi-gushes-dear-friend-putin-duo-meet-china" target="_blank">met with various world leaders</a> domestically and abroad.</p> <p><a
href="https://www.foxnews.com/world/russia-says-bidens-comparison-putin-hamas-unacceptable-report" target="_blank"><strong>RUSSIA SAYS BIDEN\'S COMPARISON OF PUTIN AND
HAMAS IS \'UNACCEPTABLE\': REPORT</strong></a></p> <p><a href="https://www.foxnews.com/world/putin-no-rivals-presidency-election-spokesman-claims"
```

```python
df = pd.DataFrame(articles)
#creating a pandas dataframe out of the articles
df.head()
#displaying the head
```

	link	guid	title	description	{http://purl.org/r
0	https://www.foxnews.com/world/un-tells-world-g...	https://www.foxnews.com/world/un-tells-world-g...	UN tells world governments not to repatriate e...	The United Nations is calling on member countr...	<p>The
1	https://www.foxnews.com/world/russias-defense-...	https://www.foxnews.com/world/russias-defense-...	Russia's defense minister accuses US of fuelin...	Russian Defense Minister Sergei Shoigu claimed...	<p>Russ
2	https://www.foxnews.com/world/kremlin-shortens...	https://www.foxnews.com/world/kremlin-shortens...	Kremlin shortens quarantine before meeting Put...	The Kremlin is reportedly reducing quarantine ...	<p>The <a
3	https://www.foxnews.com/world/israeli-soldier-...	https://www.foxnews.com/world/israeli-soldier-...	Israeli soldier rescued in Gaza after being ca...	A 19-year-old Israeli soldier who was captured...	<p>An <a
4	https://www.foxnews.com/world/canada-backpedal...	https://www.foxnews.com/world/canada-backpedal...	Canada backpedals after US, dozen other nation...	Canada's foreign minister spoke in favor ...	<p>The <a

```
df = df[["link", "title", "guid", "description", "category", "pubDate"]]
#only pulling the columns we want
df.head()
#showing the head again to make sure it saved
```

	link	title	guid	description	categ
0	https://www.foxnews.com/world/un-tells-world-g...	UN tells world governments not to repatriate e...	https://www.foxnews.com/world/un-tells-world-g...	The United Nations is calling on member countr...	art
1	https://www.foxnews.com/world/russias-defense-...	Russia's defense minister accuses US of fuelin...	https://www.foxnews.com/world/russias-defense-...	Russian Defense Minister Sergei Shoigu claimed...	art
2	https://www.foxnews.com/world/kremlin-shortens...	Kremlin shortens quarantine before meeting Put...	https://www.foxnews.com/world/kremlin-shortens...	The Kremlin is reportedly reducing quarantine ...	art
3	https://www.foxnews.com/world/israeli-soldier-...	Israeli soldier rescued in Gaza after being ca...	https://www.foxnews.com/world/israeli-soldier-...	A 19-year-old Israeli soldier who was captured...	art
4	https://www.foxnews.com/world/canada-backpedal...	Canada backpedals after US, dozen other nation...	https://www.foxnews.com/world/canada-backpedal...	Canada's foreign minister spoke in favor ...	art

```
from datetime import datetime
from datetime import timedelta


#converting date time
def convert_date(date):
    return datetime.strptime(date, '%a, %d %b %Y %H:%M:%S %z')
#date is the value we are inputting into the function and then the percentages correspond to the conversion string


df['pubDate'] = df['pubDate'].apply(lambda x: convert_date(x))
#applying the convert date function to the pubDate column


df['pubDate']
#showing that this converted the date to make it more readable

0     2023-10-31 11:35:29-04:00
1     2023-10-31 10:09:15-04:00
2     2023-10-31 09:06:10-04:00
3     2023-10-31 09:02:20-04:00
4     2023-10-31 07:25:16-04:00
5     2023-10-31 07:23:32-04:00
6     2023-10-31 06:58:45-04:00
7     2023-10-31 06:52:29-04:00
8     2023-10-31 05:45:27-04:00
9     2023-10-31 04:53:28-04:00
10    2023-10-31 04:00:38-04:00
11    2023-10-30 20:45:46-04:00
12    2023-10-30 19:26:38-04:00
13    2023-10-30 19:21:57-04:00
14    2023-10-30 17:22:00-04:00
15    2023-10-30 15:33:26-04:00
16    2023-10-30 15:31:26-04:00
17    2023-10-30 15:29:51-04:00
18    2023-10-30 15:20:39-04:00
19    2023-10-30 15:04:57-04:00
20    2023-10-30 14:28:49-04:00
21    2023-10-30 14:13:56-04:00
```

```
22    2023-10-30 14:07:02-04:00
23    2023-10-30 09:25:17-04:00
24    2023-10-30 09:00:06-04:00
Name: pubDate, dtype: datetime64[ns, UTC-04:00]
```

```
df.describe()
#describing the data frame.
#showing we have 25 unique articles
```

```
<ipython-input-76-ea8415b8a3ee>:1: FutureWarning: Treating datetime data as categorical rather than numeric in `.describe` is deprecated and will be removed in a future
  df.describe()
```

	link	title	guid	description	category
count	25	25	25	25	25
unique	25	25	25	25	1
top	https://www.foxnews.com/world/un-tells-world-g...	UN tells world governments not to repatriate e...	https://www.foxnews.com/world/un-tells-world-g...	The United Nations is calling on member countr...	article
freq	1	1	1	1	25
first	NaN	NaN	NaN	NaN	NaN
last	NaN	NaN	NaN	NaN	NaN

```
df[df['title'].apply(lambda x: 'gaza' in x.lower())]
#looking for gaza within the articles that relate to the US
# i see that there are 4 articles that include gaza within the total 25 articles
```

	link	title	guid	description	catego
3	https://www.foxnews.com/world/israeli-soldier-...	Israeli soldier rescued in Gaza after being ca...	https://www.foxnews.com/world/israeli-soldier-...	A 19-year-old Israeli soldier who was captured...	art
5	https://www.foxnews.com/world/massachusetts-co...	Massachusetts couple, 1-year-old son stranded ...	https://www.foxnews.com/world/massachusetts-co...	A Massachusetts family remains stranded in Gaz...	art
6	https://www.foxnews.com/world/israel-hits-300-...	Israel hits 300 Hamas targets in Gaza, engages...	https://www.foxnews.com/world/israel-hits-300-...	Israel's military says hundreds of target...	art
24	https://www.foxnews.com/world/more-israeli-tro...	More Israeli troops pour into Gaza with IDF on...	https://www.foxnews.com/world/more-israeli-tro...	The Israeli military is escalating its militar...	art