Web Scraping

Part #4: Getting XML from RSS Feeds

What is an RSS feed?

RSS stands for "Really Simple Syndication." It's just a page of data conforming to the XML format that is updated frequently and can be processed in an automated way.

Exploring Some RSS Feeds

Many organizations have RSS feeds. Some links are provided below that will allow you to find some of these feeds. Spend a few minutes exploring some prominent examples using Google Chrome.

Note: If the RSS feed you are looking at is a collection of HTML-style tags, you are in the right place. If not, right-click and select "View page source."

Okay, time to explore:

- Local News:
  - **Chicago Tribune**: http://www.chicagotribune.com/cs-rssfeeds-htmlstory.html (http://www.chicagotribune.com/cs-rssfeeds-htmlstory.html)
  - **The Daily Herald**: http://www.dailyherald.com/rss/ (http://www.dailyherald.com/rss/)
  - **The Chicago Sun Times**: http://www.thesuntimes.com/section/feed (http://www.thesuntimes.com/section/feed)

- National/International News:
  - **Reuters**: https://www.reuters.com/tools/rss (https://www.reuters.com/tools/rss)
  - **USA Today**: https://www.usatoday.com/rss/ (https://www.usatoday.com/rss/)
  - **The New York Times**: http://www.nytimes.com/services/xml/rss/index.html (http://www.nytimes.com/services/xml/rss/index.html)
  - **BBC News**: http://www.bbc.com/news/10628494 (http://www.bbc.com/news/10628494)____

- Technology News:
  - **Wired.com**: https://www.wired.com/about/rss_feeds/ (https://www.wired.com/about/rss_feeds/)
  - **Ars Technica**: https://arstechnica.com/rss-feeds/ (https://arstechnica.com/rss-feeds/)
  - **CNET**: https://www.cnet.com/rss/ (https://www.cnet.com/rss/)

- Miscellaneous (Sports, Government, Science):
  - **ESPN**: http://www.espn.com/espn/news/story?page=rssinfo (http://www.espn.com/espn/news/story?page=rssinfo)
  - **Illinois Commerce Commission**: https://www.icc.illinois.gov/rss/ (https://www.icc.illinois.gov/rss/)
  - **US Congress**: https://www.congress.gov/rss (https://www.congress.gov/rss)

- **NASA**: https://www.nasa.gov/content/nasa-rss-feeds (https://www.nasa.gov/content/nasa-rss-feeds)

****Task \#1:**** After you explore a few of the feeds above, try to find an RSS feed for another website you are interested in. This may be a news website for a certain type of news you like to follow (video games, style/fashion, politics, etc). Then fill in the information below for the feed (or feeds) you found:

Organization(s):Buzzfeed

URL(s):https://www.buzzfeed.com/rss

Description(s): Had 3 sections. One was for the main sections in the wbesite, such as the homepage, quizes, and other mian sections of buzzfeed. The second section acts as more of a topic finder, and lists topics buzzfeed frequently talks about. The final one is for miscellenious items within buzzfeed thart have subjects with in buzzfeed, but arent as prominent

An RSS Feed from the Wall Street Journal

The beauty of an RSS feed is that its content is updated regularly, but the structure of its tags always stays the same. This allows you to extract up-to-date data in an automated fashion.

For example, here are the top stories from the "global news" section of *The Wall Street Journal* from their RSS feed: https://feeds.a.dj.com/rss/RSSWSJD.xml (https://feeds.a.dj.com/rss/RSSWSJD.xml)

After looking at this link in Chrome, explore it using Beautiful Soup:

In [1]:
```
from bs4 import BeautifulSoup       # Import BeautifulSoup
from urllib.request import urlopen  # Import urlopen

xml_page = urlopen("https://feeds.a.dj.com/rss/RSSWSJD.xml")   # Opens whatever p
bs_obj = BeautifulSoup(xml_page, 'xml')  # Extract xml data

print(bs_obj.prettify())  # Makes it more easily readible or 'pretty'
```

```
oor-billionaire-11588967043?mod=rss_Technology)
   </link>
   <description>
    The Tesla CEO is worth $39 billion on paper, but the electric-car maker,
 who recently announced he’s selling his houses and most of his worldly posse
ssions, needs a wad of money to exercise his latest payout.
   </description>
   <content:encoded/>
   <pubDate>
    Fri, 08 May 2020 18:00:00 -0400
   </pubDate>
   <guid isPermaLink="false">
    SB123186590461354347171045863706219917518 48
   </guid>
   <category domain="AccessClassName">
    PAID
   </category>
   <dj:articletype>
    Technology
   </dj:articletype>
```

Now you can get a list of all headlines:

In [2]:
```
headlines = bs_obj.find_all('title') # Extracts and creates a list of all the <t

print(headlines)
```

```
[<title>WSJ.com: WSJD</title>, <title>WSJ.com: WSJD</title>, <title>Amazon, Ber
kshire, JPMorgan Health-Care Venture Looking for New CEO</title>, <title>Uber R
edraws Road Map to Profit in Wake of Weakened Ridership</title>, <title>LiveXLi
ve to Join Podcast Fray With PodcastOne Acquisition</title>, <title>Google Pare
nt Alphabet Drops Controversial 'Smart City' Project</title>, <title>The Troubl
e With Coronavirus Contact-Tracing Apps</title>, <title>Elon Musk, Tech's Cash-
Poor Billionaire</title>, <title>Nintendo Scores on 'Animal Crossing' Sales Tha
nks to Coronavirus Lockdowns</title>, <title>Vista Follows Facebook Into Indi
a's Jio With $1.5 Billion Digital Bet</title>, <title>Spending Too Much Time On
line? Here Are Tips for Unplugging.</title>, <title>Facebook-Backed Libra Proje
ct Gets New CEO</title>, <title>China's WeChat Monitors Foreign Users to Refine
Censorship at Home</title>, <title>Grandma to the Rescue! Parents Get Virtual H
elp</title>, <title>Airbnb to Cut 25% of Workforce</title>, <title>Welcome Back
to the Office. Your Every Move Will Be Watched.</title>, <title>Hospitals Deplo
y Technology to Reduce ICU Staff Exposure to Covid-19</title>, <title>New York
City Apartment Renting Turns to Video Chats and Virtual Tours</title>]
```

If you want to strip out the `<title>` tags, use the *.getText()* method:

In [69]:
```
headlines = [story.getText() for story in headlines] # Creates a list of innerHTM

print(headlines)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-69-9eb3fd8388be> in <module>
----> 1 headlines = [story.getText() for story in headlines] # Creates a list o
f innerHTML for each <title> tag in headlines list
      2
      3 print(headlines)

<ipython-input-69-9eb3fd8388be> in <listcomp>(.0)
----> 1 headlines = [story.getText() for story in headlines] # Creates a list o
f innerHTML for each <title> tag in headlines list
      2
      3 print(headlines)

AttributeError: 'str' object has no attribute 'getText'
```

In [ ]:

In [ ]:

Hello Mr. Nichols! I hope you are well. While the above code segment says it doesnt work, it did do its job. In the cells below this one, you can see that the heading "Title" has been removed. My computer is acting wonky, and sorry for the trouble this may cause.

In [ ]:

In [ ]:

In this feed, the first two titles appear to be for the news website rather than for news stories themselves. This is an easy fix:

In [9]:
```
headlines = headlines[2:]

print(headlines)
```

```
['Amazon, Berkshire, JPMorgan Health-Care Venture Looking for New CEO', 'Uber R
edraws Road Map to Profit in Wake of Weakened Ridership', 'LiveXLive to Join Po
dcast Fray With PodcastOne Acquisition', "Google Parent Alphabet Drops Controve
rsial 'Smart City' Project", 'The Trouble With Coronavirus Contact-Tracing App
s', "Elon Musk, Tech's Cash-Poor Billionaire", "Nintendo Scores on 'Animal Cros
sing' Sales Thanks to Coronavirus Lockdowns", "Vista Follows Facebook Into Indi
a's Jio With $1.5 Billion Digital Bet", 'Spending Too Much Time Online? Here Ar
e Tips for Unplugging.', 'Facebook-Backed Libra Project Gets New CEO', "China's
WeChat Monitors Foreign Users to Refine Censorship at Home", 'Grandma to the Re
scue! Parents Get Virtual Help', 'Airbnb to Cut 25% of Workforce', 'Welcome Bac
k to the Office. Your Every Move Will Be Watched.', 'Hospitals Deploy Technolog
y to Reduce ICU Staff Exposure to Covid-19', 'New York City Apartment Renting T
urns to Video Chats and Virtual Tours']
```

Let's do the same thing with links to these stories by creating a list of all links

In [10]:
```
urls = bs_obj.find_all('link')            # Extracts and creates a list of all tl
urls = [link.getText() for link in urls]  # Creates a list of innerHTML for each
print(urls)
```

```
['http://online.wsj.com', '', 'http://online.wsj.com', 'https://www.wsj.com/art
icles/gawande-in-talks-about-leaving-helm-of-health-care-venture-haven-11588987
079?mod=rss_Technology', 'https://www.wsj.com/articles/ubers-first-quarter-loss
-balloons-on-coronavirus-impact-11588882349?mod=rss_Technology', 'https://www.w
sj.com/articles/livexlive-to-join-podcast-fray-with-podcastone-acquisition-1158
8939211?mod=rss_Technology', 'https://www.wsj.com/articles/alphabet-subsidiary-
sidewalk-labs-abandons-toronto-smart-city-project-11588867545?mod=rss_Technolog
y', 'https://www.wsj.com/articles/curbing-coronavirus-with-a-contact-tracing-ap
p-its-not-so-simple-11588996809?mod=rss_Technology', 'https://www.wsj.com/artic
les/elon-musk-techs-cash-poor-billionaire-11588967043?mod=rss_Technology', 'htt
ps://www.wsj.com/articles/nintendo-scores-on-animal-crossing-sales-thanks-to-co
ronavirus-lockdowns-11588842065?mod=rss_Technology', 'https://www.wsj.com/artic
les/vista-follows-facebook-into-indias-jio-with-1-5-billion-digital-bet-1158892
6647?mod=rss_Technology', 'https://www.wsj.com/articles/spending-too-much-time-
online-here-are-tips-for-unplugging-11588865659?mod=rss_Technology', 'https://w
ww.wsj.com/articles/facebooks-libra-adds-new-ceo-as-cryptocurrency-project-gets
-a-reboot-11588782600?mod=rss_Technology', 'https://www.wsj.com/articles/chinas
-wechat-monitors-foreign-users-to-refine-censorship-at-home-11588852802?mod=rss
_Technology', 'https://www.wsj.com/articles/its-grandparents-to-the-rescue-for-
stressed-working-from-home-parents-11588671001?mod=rss_Technology', 'https://ww
w.wsj.com/articles/airbnb-to-cut-25-of-workforce-as-coronavirus-stalls-global-t
ravel-11588707183?mod=rss_Technology', 'https://www.wsj.com/articles/lockdown-r
eopen-office-coronavirus-privacy-11588689725?mod=rss_Technology', 'https://www.
wsj.com/articles/hospitals-deploy-technology-to-reduce-icu-staff-exposure-to-co
vid-19-11588843801?mod=rss_Technology', 'https://www.wsj.com/articles/new-york-
city-apartment-renting-turns-to-video-chats-and-virtual-tours-11588888633?mod=r
ss_Technology']
```

This time, it looks like the third link is where we want to start (The second entry is an empty String!)

In [11]:
```
urls = urls[3:]

print(urls)
```

```
['https://www.wsj.com/articles/gawande-in-talks-about-leaving-helm-of-health-ca
re-venture-haven-11588987079?mod=rss_Technology', 'https://www.wsj.com/article
s/ubers-first-quarter-loss-balloons-on-coronavirus-impact-11588882349?mod=rss_T
echnology', 'https://www.wsj.com/articles/livexlive-to-join-podcast-fray-with-p
odcastone-acquisition-11588939211?mod=rss_Technology', 'https://www.wsj.com/art
icles/alphabet-subsidiary-sidewalk-labs-abandons-toronto-smart-city-project-115
88867545?mod=rss_Technology', 'https://www.wsj.com/articles/curbing-coronavirus
-with-a-contact-tracing-app-its-not-so-simple-11588996809?mod=rss_Technology',
'https://www.wsj.com/articles/elon-musk-techs-cash-poor-billionaire-1158896704
3?mod=rss_Technology', 'https://www.wsj.com/articles/nintendo-scores-on-animal-
crossing-sales-thanks-to-coronavirus-lockdowns-11588842065?mod=rss_Technology',
'https://www.wsj.com/articles/vista-follows-facebook-into-indias-jio-with-1-5-b
illion-digital-bet-11588926647?mod=rss_Technology', 'https://www.wsj.com/articl
es/spending-too-much-time-online-here-are-tips-for-unplugging-11588865659?mod=r
ss_Technology', 'https://www.wsj.com/articles/facebooks-libra-adds-new-ceo-as-c
ryptocurrency-project-gets-a-reboot-11588782600?mod=rss_Technology', 'https://w
ww.wsj.com/articles/chinas-wechat-monitors-foreign-users-to-refine-censorship-a
t-home-11588852802?mod=rss_Technology', 'https://www.wsj.com/articles/its-grand
parents-to-the-rescue-for-stressed-working-from-home-parents-11588671001?mod=rs
s_Technology', 'https://www.wsj.com/articles/airbnb-to-cut-25-of-workforce-as-c
oronavirus-stalls-global-travel-11588707183?mod=rss_Technology', 'https://www.w
sj.com/articles/lockdown-reopen-office-coronavirus-privacy-11588689725?mod=rss_
Technology', 'https://www.wsj.com/articles/hospitals-deploy-technology-to-reduc
e-icu-staff-exposure-to-covid-19-11588843801?mod=rss_Technology', 'https://www.
wsj.com/articles/new-york-city-apartment-renting-turns-to-video-chats-and-virtu
al-tours-11588888633?mod=rss_Technology']
```

Task #2: Write a function, *random_headline(headline_list, link_list)*, that accepts a list of headlines and a list of links as input and returns an output string in the format "HEADLINE, read more at LINK."

Note: Be sure to test out your function to make sure it works as expected. Show the results of your tests below. Also, not all headlines may have a link, and your two arrays may not be 'parallel'. Try re-running the cells for the most up-to-date listings!

HINT:

- Import random to create random #'s
- Create a function named random_headline(headline_list, link_list)
  - Generate a random number between 0 and one less than the length of headline_list
  - Create a variable to hold the value of headline_list at the index of the randomly generated number
  - Create a variable to hold the value of link_list at the index of the randomly generated number
  - Create an output string in the format "HEADLINE, read more at LINK" replacing HEADLINE and LINK with your variables
  - Return the output string

In [12]:
```
# Your code here
import random as rand

def random_headline(headline_list, link_list):
    head_length = len(headline_list)
    num = rand.randint(0, head_length-1)
    choosen_head = headline_list[num]
    choosen_link = link_list[num]
    output = choosen_head + ", read more at " + choosen_link
    return output
```

In [21]:
```
random_headline(headlines, urls)
```

Out[21]:
```
'The Trouble With Coronavirus Contact-Tracing Apps, read more at https://www.wsj.com/articles/curbing-coronavirus-with-a-contact-tracing-app-its-not-so-simple-11588996809?mod=rss_Technology' (https://www.wsj.com/articles/curbing-coronavirus-with-a-contact-tracing-app-its-not-so-simple-11588996809?mod=rss_Technology')
```

Processing Other RSS Feeds

You already perused a few RSS feeds. This time, pick one of those feeds (or a new one) and explore it by writing code. As a reminder, here are some recommended feeds:

- Local News:
  - **Chicago Tribune**: http://www.chicagotribune.com/cs-rssfeeds-htmlstory.html (http://www.chicagotribune.com/cs-rssfeeds-htmlstory.html)
  - **The Daily Herald**: http://www.dailyherald.com/rss/ (http://www.dailyherald.com/rss/)
  - **The Chicago Sun Times**: http://www.thesuntimes.com/section/feed (http://www.thesuntimes.com/section/feed)

- National/International News:
  - **Reuters**: https://www.reuters.com/tools/rss (https://www.reuters.com/tools/rss)
  - **USA Today**: https://www.usatoday.com/rss/ (https://www.usatoday.com/rss/)
  - **The New York Times**: http://www.nytimes.com/services/xml/rss/index.html (http://www.nytimes.com/services/xml/rss/index.html)
  - **BBC News**: http://www.bbc.com/news/10628494 (http://www.bbc.com/news/10628494)____

- Technology News:
  - **Wired.com**: https://www.wired.com/about/rss_feeds/ (https://www.wired.com/about/rss_feeds/)
  - **Ars Technica**: https://arstechnica.com/rss-feeds/ (https://arstechnica.com/rss-feeds/)
  - **CNET**: https://www.cnet.com/rss/ (https://www.cnet.com/rss/)

- Miscellaneous (Sports, Government, Science):
  - **ESPN**: http://www.espn.com/espn/news/story?page=rssinfo (http://www.espn.com/espn/news/story?page=rssinfo)
  - **Illinois Commerce Commission**: https://www.icc.illinois.gov/rss/ (https://www.icc.illinois.gov/rss/)
  - **US Congress**: https://www.congress.gov/rss (https://www.congress.gov/rss)

- **NASA**: https://www.nasa.gov/content/nasa-rss-feeds (https://www.nasa.gov/content/nasa-rss-feeds)

Task #3: Pick any of the feeds above (or more than one). Experiment using Python code, and show the results of your experimentation below.

Note: This question is fairly open-ended, but at a minimum you must do the following:

- Create a "random headline"-style function like you did above. You should not expect code such as **headlines = headlines[2:]** or **urls = urls[3:]** to fit perfectly with your data since these were modifications that may be specific to the way **The Wall Street Journal**'s RSS feed is organized.
- You must show that you engaged with the feed(s) you picked using the Beautiful Soup module and at least one Python data structure (probably lists). You will need to analyze the XML for the feed you pick and write your code to fit with the data.

In [36]:
```
# Your code here
from bs4 import BeautifulSoup       # Import BeautifulSoup
from urllib.request import urlopen  # Import urlopen

DH_page = urlopen("http://www.dailyherald.com/rss/")   # Opens whatever page we 
DH_obj = BeautifulSoup(DH_page, 'xml')  # Extract xml data

print(DH_obj.prettify())  # Makes it more easily readible or 'pretty'
```

```
  </item>
  <item>
   <title>
    Testing in Illinois passes 20,000 a day, but can schools reopen this fal
l?
   </title>
   <link>
    http://www.dailyherald.com/news/20200508/testing-in-illinois-passes-20000
-a-day-but-can-schools-reopen-this-fall (http://www.dailyherald.com/news/2020
0508/testing-in-illinois-passes-20000-a-day-but-can-schools-reopen-this-fall)
   </link>
   <guid>
    http://www.dailyherald.com/article/20200508/news/200509300 (http://www.da
ilyherald.com/article/20200508/news/200509300)
   </guid>
   <pubDate>
    Fri, 8 May 2020 20:40:03 -0400
   </pubDate>
   <description>
    Testing for COVID-19 surpassed the 20,000-a-day milestone, officials sai
```

In [47]:
```
Description = DH_obj.find_all('description') # Extracts and creates a list of al

print(Description)
```

```
[<description/>, <description>Coronavirus outbreaks continue to ravage Illinois
nursing homes as new state data show at least 1,553 deaths associated with long
-term care facilities.</description>, <description>Testing for COVID-19 surpass
ed the 20,000-a-day milestone, officials said, but some measures of Illinois' s
uccess fighting the deadly virus lag, raising the question of whether schools c
ould reopen this fall -- a hope Gov. J.B. Pritzker expressed Friday.</descripti
on>, <description>Little Richard, one of the chief architects of rock 'n' roll
whose piercing wail, pounding piano and towering pompadour irrevocably altered
popular music while introducing black R&amp;amp;B to white America, died Saturday a
fter battling bone cancer. He was 87.</description>, <description>With social d
istancing precautions in place, the McHenry Outdoor Theater welcomes movie fans
this weekend for a prehistoric double bill under the stars.</description>, <des
cription>Sean Thomas, grandson of Wendy's founder Dave Thomas, plans to open hi
s own burger restaurant Monday in Kildeer. Fresh Stack Burger Co. will open for
curbside and third-party delivery service until restaurants are allowed to reop
en for dining in.</description>, <description>The Elgin Youth Symphony Orchestr
a will hold a virtual concert Sunday.</description>, <description>A Naperville
firefighter was injured at the scene of a fire in the 1700 block of Baybrook La
ne early Saturday. The residents of the two-story home were able to evacuate sa
fely, officials said.</description>, <description>Jeanne Hansen of Round Lake i
s back home for Mother's Day after three weeks in the hospital fighting COVID-1
9.</description>, <description>The state and county websites provide data on th
e number of cases on May 8 by county and, in many cases, by suburb. Plus, you c
an search by ZIP code.</description>, <description>Here's a song intended to li
ft you out of your coronavirus funk.</description>, <description>Two swans were
killed, most likely by a coyote or fox, after they were left in an outdoor pen
at an Itasca business park.</description>, <description>Cardinal Blase Cupich m
arks teacher appreciation week with a thank you video to educators.</descriptio
n>, <description>This year, many suburban restaurants are offering Mother's Day
brunches and dinners to-go so you can celebrate Mom at home.</description>, <de
scription>Little Richard, one of the chief architects of rock '□n' roll, has di
ed</description>, <description>Mother's Day weekend got off to an unseasonably
snowy start in areas of the Northeast thanks to the polar vortex</description>,
<description>South Korea's capital closed down more than 2,100 bars and other n
ightspots Saturday because of a new cluster of coronavirus infections, and Germ
any scrambled to contain fresh outbreaks at slaughterhouses, underscoring the d
angers authorities face as they try to reopen their economies.</description>, <
description>The rookies with the Pittsburgh Steelers are trying to find a way t
o get their careers off on the right foot amid the COVID-19 pandemic</descripti
on>, <description>Rank-and-file Chicago police officers have chosen a veteran c
op who's a longtime critic of the department's top brass as the new leader of t
heir police union</description>, <description>An event billed as the first tenn
is matches involving ranked players in the United States since the sport was pu
t on hold by the coronavirus gave the participants a chance to earn money and c
ompete - and gave fans at home a live event to watch on television</description
>, <description>Russian President Vladimir Putin has marked Victory Day, the an
niversary of the defeat of Nazi Germany in World War II, in a ceremony shorn of
its usual military parade and pomp by the coronavirus pandemic</description>]
```

In [60]:
```
Description = [story.getText() for story in Description] # Creates a list of inne

print(Description)
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-60-cb6373cb8980> in <module>
----> 1 Description = [story.getText() for story in Description] # Creates a li
st of innerHTML for each <title> tag in headlines list
      2
      3 print(Description)

<ipython-input-60-cb6373cb8980> in <listcomp>(.0)
----> 1 Description = [story.getText() for story in Description] # Creates a li
st of innerHTML for each <title> tag in headlines list
      2
      3 print(Description)

AttributeError: 'str' object has no attribute 'getText'
```

In [61]:
```
Description[3:]
```

Out[61]:
```
["Little Richard, one of the chief architects of rock 'n' roll whose piercing w
ail, pounding piano and towering pompadour irrevocably altered popular music wh
ile introducing black R&B to white America, died Saturday after battling bone c
ancer. He was 87.",
 'With social distancing precautions in place, the McHenry Outdoor Theater welc
omes movie fans this weekend for a prehistoric double bill under the stars.',
 "Sean Thomas, grandson of Wendy's founder Dave Thomas, plans to open his own b
urger restaurant Monday in Kildeer. Fresh Stack Burger Co. will open for curbsi
de and third-party delivery service until restaurants are allowed to reopen for
dining in.",
 'The Elgin Youth Symphony Orchestra will hold a virtual concert Sunday.',
 'A Naperville firefighter was injured at the scene of a fire in the 1700 block
of Baybrook Lane early Saturday. The residents of the two-story home were able
to evacuate safely, officials said.',
 "Jeanne Hansen of Round Lake is back home for Mother's Day after three weeks i
n the hospital fighting COVID-19.",
 'The state and county websites provide data on the number of cases on May 8 by
county and, in many cases, by suburb. Plus, you can search by ZIP code.',
 "Here's a song intended to lift you out of your coronavirus funk.",
 'Two swans were killed, most likely by a coyote or fox, after they were left i
n an outdoor pen at an Itasca business park.',
 'Cardinal Blase Cupich marks teacher appreciation week with a thank you video
to educators.',
 "This year, many suburban restaurants are offering Mother's Day brunches and d
inners to-go so you can celebrate Mom at home.",
 "Little Richard, one of the chief architects of rock '\x98n' roll, has died",
 "Mother's Day weekend got off to an unseasonably snowy start in areas of the N
ortheast thanks to the polar vortex",
 "South Korea's capital closed down more than 2,100 bars and other nightspots S
aturday because of a new cluster of coronavirus infections, and Germany scrambl
ed to contain fresh outbreaks at slaughterhouses, underscoring the dangers auth
orities face as they try to reopen their economies.",
 'The rookies with the Pittsburgh Steelers are trying to find a way to get thei
r careers off on the right foot amid the COVID-19 pandemic',
 "Rank-and-file Chicago police officers have chosen a veteran cop who's a longt
ime critic of the department's top brass as the new leader of their police unio
n",
 'An event billed as the first tennis matches involving ranked players in the U
nited States since the sport was put on hold by the coronavirus gave the partic
ipants a chance to earn money and compete - and gave fans at home a live event
to watch on television',
 'Russian President Vladimir Putin has marked Victory Day, the anniversary of t
he defeat of Nazi Germany in World War II, in a ceremony shorn of its usual mil
itary parade and pomp by the coronavirus pandemic']
```

In [62]:
```
url = DH_obj.find_all('link')            # Extracts and creates a list of all the
url = [link.getText() for link in url]  # Creates a list of innerHTML for each <
print(url)
```

```
['http://www.dailyherald.com/', 'http://www.dailyherald.com/news/20200508/covid
-19-nursing-home-deaths-climb-to-1553-xadx2014-48-of-state-total', 'http://www.
dailyherald.com/news/20200508/testing-in-illinois-passes-20000-a-day-but-can-sc
hools-reopen-this-fall', 'http://www.dailyherald.com/entlife/20200509/little-ri
chard-flamboyant-rock-n-roll-pioneer-dead-at-87', 'http://www.dailyherald.com/n
ews/20200508/night-out-at-the-movies-mchenry-drive-in-delivers-double-bill-with
-pandemic-precautions', 'http://www.dailyherald.com/news/20200509/fresh-stack-b
urger-co-set-to-open-monday-in-kildeer', 'http://www.dailyherald.com/news/20200
509/eyso-creates-virtual-way-to-conclude-44th-season', 'http://www.dailyherald.
com/news/20200509/firefighter-injured-at-scene-of-naperville-fire', 'http://ww
w.dailyherald.com/news/20200509/mothers-day-means-even-more-this-year-for-hanse
n', 'http://www.dailyherald.com/news/20200508/may-8-covid-19-cases-per-county-s
earch-by-zip-code', 'http://www.dailyherald.com/entlife/20200509/a-song-to-make
-you-smile-i-saved-the-world-today', 'http://www.dailyherald.com/news/20200508/
2-swans-found-dead-in-pen-at-itasca-business-park', 'http://www.dailyherald.co
m/news/20200509/watch-cardinal-cupich-thanks-educators', 'http://www.dailyheral
d.com/entlife/20200507/mothers-day-2020-suburban-restaurants-offer-special-to-g
o-meals-this-year', 'http://www.dailyherald.com/article/20200509/news/30509997
6/', 'http://www.dailyherald.com/article/20200509/news/305099961/', 'http://ww
w.dailyherald.com/business/20200509/reopenings-bring-new-cases-in-south-korea-v
irus-fears-in-italy', 'http://www.dailyherald.com/article/20200509/sports/30509
9963/', 'http://www.dailyherald.com/article/20200509/news/305099962/', 'http://
www.dailyherald.com/article/20200509/sports/305099964/', 'http://www.dailyheral
d.com/article/20200509/news/305099988/']
```

In [63]:
```
url = url[3:]
print(url)
```

```
['http://www.dailyherald.com/entlife/20200509/little-richard-flamboyant-rock-n-
roll-pioneer-dead-at-87', 'http://www.dailyherald.com/news/20200508/night-out-a
t-the-movies-mchenry-drive-in-delivers-double-bill-with-pandemic-precautions',
'http://www.dailyherald.com/news/20200509/fresh-stack-burger-co-set-to-open-mon
day-in-kildeer', 'http://www.dailyherald.com/news/20200509/eyso-creates-virtual
-way-to-conclude-44th-season', 'http://www.dailyherald.com/news/20200509/firefi
ghter-injured-at-scene-of-naperville-fire', 'http://www.dailyherald.com/news/20
200509/mothers-day-means-even-more-this-year-for-hansen', 'http://www.dailyhera
ld.com/news/20200508/may-8-covid-19-cases-per-county-search-by-zip-code', 'htt
p://www.dailyherald.com/entlife/20200509/a-song-to-make-you-smile-i-saved-the-w
orld-today', 'http://www.dailyherald.com/news/20200508/2-swans-found-dead-in-pe
n-at-itasca-business-park', 'http://www.dailyherald.com/news/20200509/watch-car
dinal-cupich-thanks-educators', 'http://www.dailyherald.com/entlife/20200507/mo
thers-day-2020-suburban-restaurants-offer-special-to-go-meals-this-year', 'htt
p://www.dailyherald.com/article/20200509/news/305099976/', 'http://www.dailyher
ald.com/article/20200509/news/305099961/', 'http://www.dailyherald.com/busines
s/20200509/reopenings-bring-new-cases-in-south-korea-virus-fears-in-italy', 'ht
tp://www.dailyherald.com/article/20200509/sports/305099963/', 'http://www.daily
herald.com/article/20200509/news/305099962/', 'http://www.dailyherald.com/artic
le/20200509/sports/305099964/', 'http://www.dailyherald.com/article/20200509/ne
ws/305099988/']
```

In [64]:
```
import random as rand

def random_ad(Description_list, link_list):
    head_length = len(Description_list)
    num = rand.randint(0, head_length-1)
    choosen_head = Description_list[num]
    choosen_link = link_list[num]
    output = str(choosen_head) + ". Please read more at " + str(choosen_link)
    return output
```

In [65]:
```
random_ad(Description, url)
```

Out[65]:
```
'The rookies with the Pittsburgh Steelers are trying to find a way to get their
careers off on the right foot amid the COVID-19 pandemic. Please read more at h
ttp://www.dailyherald.com/article/20200509/news/305099988/' (http://www.dailyhe
rald.com/article/20200509/news/305099988/')
```