Experiment-04

Name: Atharva Paliwal

Roll No.: 40

Date: 12-09-2021

Aim: To perform web scrapping and then perform analysis.

(A) Extract paragraph data from Wikipedia along with image and display image and text extracted.

(B) Extract whether data from the given URL and perform various analytic tasks like find average temperature.

(C) Extract images based on category and save in separate folders that can be used as training images for machine learning algorithm.

Theory:

Web Scrapping Web scraping (also called web data extraction or data scraping) provides a solution for those who want to get access to structured web data in an automated fashion. Web scraping is useful if the public website you want to get data from doesn't have an API, or it does but provides only limited access to the data.

Steps to perform web scrapping

Step-1: Provide Url and download the web page using requests library Requests (HTTP) Library for Web Scraping – It is used for making various types of HTTP requests like GET, POST, etc. It is the most basic yet the most essential of all libraries.

- Requests allows you to send HTTP requests
- It enables downloading pages using the Python requests library. The requests library will make a GET request to a web server, which will download the HTML contents of a given web page for us.
- Download page using requests.get() method.
- After running our request, we get a Response object. This object has a status_code property, which indicates if the page was downloaded successfully.
- And a content property that gives the HTML content of the webpage as output.

In [ ]:

Step 2: Beautiful Soup for page parsing

Beautiful Soup Library for Web Scraping – Its work involves creating a parse tree for parsing content. A perfect starting library for beginners and very easy to work with.

- After the contents of the page are downloaded, BeautifulSoup a Python library is used for pulling data out of HTML and XML files.
- It needs an input (document or URL) to create a soup object as it cannot fetch a web page by itself.
- We have other modules such as regular expression, lxml for the same purpose.
- We then process the data in CSV or JSON or MySQL format.
- As all the tags are nested, we can move through the structure one level at a time. We can first select all the elements at the top level of the page using the children property of soup.
- Note that children returns a list generator, so we need to call the list function on it:
- If we want to extract a single tag, we can instead use the find_all() method, which will find all the instances of a tag on a page. Note that find_all() returns a list, so we'll have to loop through, or use list indexing, to extract text. If you instead only want to find the first instance of a tag, you can use the find method, which will return a single BeautifulSoup object.

In [ ]:

Code and Output

Part A:

In [1]:

```
!pip install selenium
!pip install requests
!pip install urllib3
!pip install bs4
```

Requirement already satisfied: selenium in d:\anaconda\lib\site-packages (3. 141.0)

Requirement already satisfied: urllib3 in d:\anaconda\lib\site-packages (fro m selenium) (1.25.11)

Requirement already satisfied: requests in d:\anaconda\lib\site-packages (2.

24.0)

Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in d:\anaconda\lib\site-packages (from requests) (1.25.11)

Requirement already satisfied: idna<3,>=2.5 in d:\anaconda\lib\site-packages (from requests) (2.10)

Requirement already satisfied: certifi>=2017.4.17 in d:\anaconda\lib\site-pa ckages (from requests) (2020.6.20)

Requirement already satisfied: chardet<4,>=3.0.2 in d:\anaconda\lib\site-pac kages (from requests) (3.0.4)

Requirement already satisfied: urllib3 in d:\anaconda\lib\site-packages (1.2 5.11)

Requirement already satisfied: bs4 in d:\anaconda\lib\site-packages (0.0.1) Requirement already satisfied: beautifulsoup4 in d:\anaconda\lib\site-packag es (from bs4) (4.9.3)

Requirement already satisfied: soupsieve>1.2; python_version >= "3.0" in d:\anaconda\lib\site-packages (from beautifulsoup4->bs4) (2.0.1)

In [ ]:

[2]:

```
# import required modules
import requests

# get URL
page = requests . get ( "https://en.wikipedia.org/wiki/Paragraph" )

# display status code
print ( page . status_code )

# display scrapped data
print ( page . content )
```

200

b'<!DOCTYPE html>\n<html class="client-nojs" lang="en" dir="ltr">\n<head> \n<meta charset="UTF-8"/>\n<title>Paragraph - Wikipedia</title>\n<script> document.documentElement.className="client-js";RLCONF={"wgBreakFrames":! 1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wg

DefaultDateFormat":"dmy","wgMonthNames":["","January","February","Marc h","April","May","June","July","August","September","October","Novembe r","December"],"wgRequestId":"88b2811d-7640-4718-97d3-5fdcf9cd644e","wgCS PNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageName":!1,"wgN amespaceNumber":0,"wgPageName":"Paragraph","wgTitle":"Paragraph","wgCurRe visionId":1039583963,"wgRevisionId":1039583963,"wgArticleId":230752,"wgIs Article":!0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":null,"wgUser Groups":["*"],"wgCategories":["Articles with limited geographic scope fro m June 2013","Articles containing Ancient Greek (to 1453)-language tex t","All articles with unsourced statements","Articles with unsourced stat ements from January 2020","Typography","Writing"],"wgPageContentLanguag e":"en","wgPageContentModel":"wikitext",\n"wgRelevantPageName":"Paragrap h","wgRelevantArticleId":230752,"wgIsProbablyEditable":!0,"wgRelevantPage IsProbablyEditable":!0,"wgRestrictionEdit":[],"wgRestrictionMove":[],"wgF

l dR P " {"t " {" t t " {"l l " 1}}} " M di Vi O Cli

[3]:

```
# import required modules
from bs4 import BeautifulSoup

# scrape webpage
soup = BeautifulSoup ( page . content , 'html.parser' )

# display scrapped data
print ( soup . prettify ())
```

<!DOCTYPE html>

<html class="client-nojs" dir="ltr" lang="en">

In [ ]:

<head>

<meta charset="utf-8"/>

<title>

Paragraph - Wikipedia

</title> <script>

document.documentElement.className="client-js";RLCONF={"wgBreakFrame
s":!1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":

["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","Februar
y","March","April","May","June","July","August","September","October","No
vember","December"],"wgRequestId":"88b2811d-7640-4718-97d3-5fdcf9cd644
e","wgCSPNonce":!1,"wgCanonicalNamespace":"","wgCanonicalSpecialPageNam
e":!1,"wgNamespaceNumber":0,"wgPageName":"Paragraph","wgTitle":"Paragrap
h","wgCurRevisionId":1039583963,"wgRevisionId":1039583963,"wgArticleId":2
30752,"wgIsArticle":!0,"wgIsRedirect":!1,"wgAction":"view","wgUserName":n
ull,"wgUserGroups":["*"],"wgCategories":["Articles with limited geographi c scope from June
2013","Articles containing Ancient Greek (to 1453)-lang t t" "All ti l ith d t t t " "A ti l ith

[4]:

```
# EXTRACTING PARAGRAPH

list (soup.children)

# find all occurance of p in HTML
# includes HTML tags
# Why p? Because we want paragraphs
print (soup.find_all ('p'))

print ('\n\n')
print ("================================================================================")
# return only text
# does not include HTML tags
print (soup.find_all ('p')[0].get_text())
print ("================================================================================")
```

[<p>A paragraph (from the Ancient Greek παράγραφος, parágraphos, "to write beside") is a self-contained unit of discourse in writing dealing with a particular point or idea. A paragraph consists of one or more sentences. <sup class="reference" id="cite_ref-UNC_1-0">

[1]</sup> Though not required by the syntax of any language, paragrap hs are usually an expected part of formal writing, used to organize longe r prose.<sup class="noprint Inlin e-Template Template-Fact" style="white-space:nowrap;">[<span title

In [ ]:

="This claim needs references to reliable sources. (January 2020)">citati on needed</i>]</sup>

, The oldest classical Greek and Latin writing had little or no sp ace between words and could be written in boustrophedon (alternating directions). Over tim t t di ti (l ft t i ht) b t d di d d < h f "/ ik

In [5]:

```
# EXTRACTING CUSTOMIZED TAG CONTENTS

object = soup.find(id="Numbering")

# find tags
items = object.find_all(class_="anchor")
result = items[0]

# display tags
print(result.prettify())
```

[6]:

```
# EXTRACTING IMAGE

image_tags = soup.findAll('img')
# print out image urls
for image_tag in image_tags:
    print(image_tag.get('src'))
```

//upload.wikimedia.org/wikipedia/commons/thumb/b/bd/Ambox_globe_content.svg/

48px-Ambox_globe_content.svg.png

//upload.wikimedia.org/wikipedia/commons/thumb/5/59/United_States_Constituti on.jpg/220px-United_States_Constitution.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/9/99/Wiktionary-logo-en-v2.sv g/16px-Wiktionary-logo-en-v2.svg.png

//upload.wikimedia.org/wikipedia/en/thumb/9/96/Symbol_category_class.svg/16p x-Symbol_category_class.svg.png

//en.wikipedia.org/wiki/Special:CentralAutoLogin/start?type=1x1

/static/images/footer/wikimedia-button.png

[10]: In [ ]:

```
# EXTRACTING TITLE
title = soup.find_all('title')
print(title[0].get_text())
```

Paragraph - Wikipedia

/static/images/footer/poweredby_mediawiki_88x31.png In [7]:

```
#getting jpg images
import re
image_tags = soup.findAll('img',{'src':re.compile('.jpg')})
# print out image urls
for image_tag in image_tags:
    print(image_tag['src']+'\n')
```

//upload.wikimedia.org/wikipedia/commons/thumb/5/59/United_States_Constituti
on.jpg/220px-United_States_Constitution.jpg

In [8]:

```
import urllib.request
urllib.request.urlretrieve("https://upload.wikimedia.org/wikipedia/commons/thumb/5/59/U
```

Out[8]: ('image1.jpg', <http.client.HTTPMessage at 0x18abef144f0>)

In [9]:

In [ ]:

```

```

```
from PIL import Image

#read the image
im = Image.open("image1.jpg")

#show image
im.show()
```

Image Extracted:

Part B:

In [1]:
```
import requests
from bs4 import BeautifulSoup
page = requests.get("https://forecast.weather.gov/MapClick.php?lat=37.7772&lon=-1
soup = BeautifulSoup(page.content, 'html.parser')

# display scrapped data
print(soup.prettify())
```

```
<!DOCTYPE html>
<html class="no-js">
 <head>
  <!-- Meta -->
  <meta content="width=device-width" name="viewport"/>
  <link href="http://purl.org/dc/elements/1.1/" rel="schema.DC"/>
  <title>
   National Weather Service
  </title>
  <meta content="National Weather Service" name="DC.title">
   <meta content="NOAA National Weather Service National Weather Service" na
me="DC.description"/>
   <meta content="US Department of Commerce, NOAA, National Weather Service"
name="DC.creator"/>
   <meta content="" name="DC.date.created" scheme="ISO8601"/>
   <meta content="EN-US" name="DC.language" scheme="DCTERMS.RFC1766"/>
   <meta content="weather, National Weather Service" name="DC.keywords"/>
   <meta content="NOAA's National Weather Service" name="DC.publisher"/>
   <meta content="National Weather Service" name="DC.contributor"/>
```

In [2]:
```
seven_day = soup.find(id="seven-day-forecast")
forecast_items = seven_day.find_all(class_="tombstone-container")
tonight = forecast_items[0]
print(tonight.prettify())
```

```
<div class="tombstone-container">
 <p class="period-name">
  Overnight
  <br/>
  <br/>
 </p>
 <p>
  <img alt="Overnight: Widespread haze before 2am. Mostly cloudy, with a low ar
ound 54. West wind around 11 mph. " class="forecast-icon" src="newimages/mediu
m/hz.png" title="Overnight: Widespread haze before 2am. Mostly cloudy, with a l
ow around 54. West wind around 11 mph. "/>
 </p>
 <p class="short-desc">
  Haze
 </p>
 <p class="temp temp-low">
  Low: 54 °F
 </p>
</div>
```

3

```
period = tonight.find(class_="period-name").get_text()
short_desc = tonight.find(class_="short-desc").get_text()
temp = tonight.find(class_="temp").get_text()
print(period)
print(short_d sc)
print(temp)
```

```
Overnight
Haze
Low: 54 °F
```

In [4]:

```
#extract the title attribute from the img tag.
img = tonight.find("img")
desc = img['title']
print(desc)
```

```
Overnight: Widespread haze before 2am. Mostly cloudy, with a low around 54. Wes
t wind around 11 mph.
```

In [5]:

```
#extract all information from the Page
period_tags = seven_day.select(".tombstone-container .period-name")
periods = [pt.get_text() for pt in period_tags]
periods
```

Out[5]:

```
['Overnight',
 'Friday',
 'FridayNight',
 'Saturday',
 'SaturdayNight',
 'Sunday',
 'SundayNight',
 'LaborDay',
 'MondayNight']
```

6

```
#get other three fields
short_descs = [sd.get_text() for sd in seven_day.select(".tombstone-container .sh
temps = [t.get_text() for t in seven_day.select(".tombstone-container .temp")]
descs = [d["title"] for d in seven_day.select(".tombstone-container img")]
print(short_d scs)
print(temps)
print(descs)
```

```
['Haze', 'BecomingSunny', 'Partly Cloudy', 'BecomingSunny', 'Mostly Clearand Br
eezythen PartlyCloudy', 'Mostly Sunny', 'Partly Cloudyand Breezythen MostlyClou
dy', 'Partly Sunny', 'Mostly Cloudy']
['Low: 54 °F', 'High: 70 °F', 'Low: 55 °F', 'High: 73 °F', 'Low: 56 °F', 'High:
77 °F', 'Low: 57 °F', 'High: 77 °F', 'Low: 57 °F']
['Overnight: Widespread haze before 2am. Mostly cloudy, with a low around 54. W
est wind around 11 mph. ', 'Friday: Mostly cloudy through mid morning, then gra
dual clearing, with a high near 70. Light west southwest wind becoming west 15
to 20 mph in the afternoon. Winds could gust as high as 25 mph. ', 'Friday Nigh
t: Partly cloudy, with a low around 55. West southwest wind 16 to 21 mph decrea
sing to 10 to 15 mph after midnight. Winds could gust as high as 28 mph. ', 'Sa
turday: Partly sunny, then gradually becoming sunny, with a high near 73. West
wind 8 to 18 mph, with gusts as high as 24 mph. ', 'Saturday Night: Mostly clea
r, with a low around 56. Breezy, with a west wind 17 to 22 mph decreasing to 11
to 16 mph in the evening. Winds could gust as high as 28 mph. ', 'Sunday: Mostl
y sunny, with a high near 77.', 'Sunday Night: Partly cloudy, with a low around
57. Breezy. ', 'Labor Day: Partly sunny, with a high near 77.', 'Monday Night:
Mostly cloudy, with a low around 57.']
```

8

```
import pandas as pd
weather = pd.DataFrame({
    "period": periods,
    "short_desc": short_descs,
    "temp": t mps,
    "desc":descs
})
weather
```

Out[8]:

	period	short_desc	temp	desc
0	Overnight	Haze	Low: 54 °F	Overnight: Widespread haze before 2am. Mostly ...
1	Friday	BecomingSunny	High: 70 °F	Friday: Mostly cloudy through mid morning, the...
2	FridayNight	Partly Cloudy	Low: 55 °F	Friday Night: Partly cloudy, with a low around...
3	Saturday	BecomingSunny	High: 73 °F	Saturday: Partly sunny, then gradually becomin...
4	SaturdayNight	Mostly Clearand Breezythen PartlyCloudy	Low: 56 °F	Saturday Night: Mostly clear, with a low aroun...
5	Sunday	Mostly Sunny	High: 77 °F	Sunday: Mostly sunny, with a high near 77.
6	SundayNight	Partly Cloudyand Breezythen MostlyCloudy	Low: 57 °F	Sunday Night: Partly cloudy, with a low around...
7	LaborDay	Partly Sunny	High: 77 °F	Labor Day: Partly sunny, with a high near 77.
8	MondayNight	Mostly Cloudy	Low: 57 °F	Monday Night: Mostly cloudy, with a low around...

In [9]:

```
import re
temp_nums = weather["temp"].str.extract("(?P<temp_num>\d+)", expand=False)
weather["temp_num"] = temp_nums.astype('int')
temp_nums
weather["temp_num"].mean()
```

Out[9]:

```
64.0
```

In [11]:

```
writer = pd.ExcelWriter('file_name.xlsx', engine='xlsxwriter')
df = pd.DataFrame(weather)
df.to_excel(writer)
writer.save()
```

Data Extracted:

	A	B	C	D	E	F	G
1		**period**	**short_desc**	**temp**	**desc**	**temp_num**	
2	**0**	Today	Sunny	High: 70 °F	Today: Sur	70	
3	**1**	Tonight	Partly Clou	Low: 55 °F	Tonight: P	55	
4	**2**	Monday	Sunny	High: 70 °F	Monday: S	70	
5	**3**	MondayNi	Mostly Cle	Low: 56 °F	Monday N	56	
6	**4**	Tuesday	Sunny	High: 74 °F	Tuesday: S	74	
7	**5**	TuesdayNi	Mostly Cle	Low: 57 °F	Tuesday N	57	
8	**6**	Wednesda	Mostly Sur	High: 69 °F	Wednesda	69	
9	**7**	Wednesda	Partly Clou	Low: 56 °F	Wednesda	56	
10	**8**	Thursday	Mostly Sur	High: 67 °F	Thursday:	67	
11							
12							

Part C:

In [1]:
```
import requests
from bs4 import BeautifulSoup

dogs_url = "https://bowwowinsurance.com.au/dogs/dogs-breeds/"
cats_url = "https://en.wikipedia.org/wiki/Cat"

dogs = requests.get(dogs_url)
cats = requests.get(cats_url)

dogSoup = BeautifulSoup(dogs.content, 'html.parser')
catSoup = BeautifulSoup(cats.content, 'html.parser')
#display scrapped data
print(" =  = = DOGS = = =")
print(dogSoup.prettify())

print(" =  = = CATS = = =")
print(catSoup.prettify())
```

```
 =  = = DOGS = = =
<!DOCTYPE html>
<html class="no-js no-svg" lang="en-AU">
 <head>
  <meta charset="utf-8"/>
  <meta content="no-cache" http-equiv="pragma"/>
  <meta content="no-cache" http-equiv="cache-control"/>
  <meta content="width=device-width, initial-scale=1" name="viewport"/>
  <meta content="3IeuFDx4Ya0ITzcRpx7RnclKjcd0QRvSLUuRr_XsXzA" name="google-s
ite-verification">
   <meta content="pl27ev169oymmm2uy37tl0iww4a68s" name="facebook-domain-veri
fication">
    <link href="http://gmpg.org/xfn/11" rel="profile"/>
    <link href="https://bowwowinsurance.com.au/wp-content/themes/bwm/assets/
images/favicon/favicon.ico?v=2" rel="icon" type="image/x-icon">
     <link disabled="" href="https://fonts.googleapis.com/css?family=Noto+Sa
ns:400,700" id="NotoSans" rel="stylesheet"/>
     <script type="text/javascript">
      window.onload = function()
     {
```

2

```
#Extracting images from dog url

image_tags = dogSoup.findAll('img')
# print out image urls
for image_tag in image_tags:
    print(image_tag.get('src'))
```

```
https://www.facebook.com/tr?id=431896870524265&ev=PageView&noscript=1 (http
s://www.facebook.com/tr?id=431896870524265&ev=PageView&noscript=1)
https://bowwowinsurance.com.au/wp-content/themes/bwm/assets/images/icons/ico
n-call-navy.svg (https://bowwowinsurance.com.au/wp-content/themes/bwm/asset
s/images/icons/icon-call-navy.svg)
https://bowwowinsurance.com.au/wp-content/themes/bwm/assets/images/icons/ico
n-login-navy.svg (https://bowwowinsurance.com.au/wp-content/themes/bwm/asset
s/images/icons/icon-login-navy.svg)
https://bowwowinsurance.com.au/wp-content/themes/bwm/assets/images/icons/ico
n-search-navy.svg (https://bowwowinsurance.com.au/wp-content/themes/bwm/asse
ts/images/icons/icon-search-navy.svg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/airedale-terrier-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aireda
le-terrier-700x700.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/akita-700x700.jpg
 (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/akita-700x700.jp
g)
https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_15250
126-Alaskan-Husky-in-front-of-white-background-thumbnail.jpg (https://bowwow
insurance.com.au/wp-content/uploads/2020/09/shutterstock_15250126-Alaskan-Hu
sky-in-front-of-white-background-thumbnail.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/alaskan-malamute-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/alaska
n-malamute-700x700.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/11/american-bulldog-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/11/americ
an-bulldog-700x700.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2021/04/shutterstock_14123
64536-THUMBNAIL-The-American-Eskimo-Dog-is-a-breed-of-companion-dog-originat
ing-in-Germany-The-American-Eskimo-is-a-member-of-the-Spitz-family.jpg (http
s://bowwowinsurance.com.au/wp-content/uploads/2021/04/shutterstock_141236453
6-THUMBNAIL-The-American-Eskimo-Dog-is-a-breed-of-companion-dog-originating-
in-Germany-The-American-Eskimo-is-a-member-of-the-Spitz-family.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2021/02/shutterstock_14448
33281-American-Foxhound-in-a-public-park-thumbnail-Bow-Wow-Meow-Pet-Insuranc
e.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2021/02/shutterstoc
k_1444833281-American-Foxhound-in-a-public-park-thumbnail-Bow-Wow-Meow-Pet-I
nsurance.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/american-staffords
hire-terrier-amstaff-american-staffy-700x700.jpg (https://bowwowinsurance.co
m.au/wp-content/uploads/2018/10/american-staffordshire-terrier-amstaff-ameri
can-staffy-700x700.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aussie-bulldog-thu
mb.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aussie-bul
ldog-thumb.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/blue-heeler-austra
lian-cattledog-700x700.jpg (https://bowwowinsurance.com.au/wp-content/upload
s/2018/10/blue-heeler-australian-cattledog-700x700.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_16352
```

72731-Cute-red-abricot-Australian-Cobberdog-Labradoodle-dog-pup-sitting-up-with-one-paw-high-in-air.-Mouth-closed.-Isolated-on-white-background.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_1635272731-Cute-red-abricot-Australian-Cobberdog-Labradoodle-dog-pup-sitting-up-with-one-paw-high-in-air.-Mouth-closed.-Isolated-on-white-background.jpg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/australian-kelpie-isolated-thumb-700x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/australian-kelpie-isolated-thumb-700x700.jpg)
None
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/icon-notsure-yellow.svg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/icon-notsure-yellow.svg)
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/dog-licking-face-of-guy-in-park.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/dog-licking-face-of-guy-in-park.jpg)
None
None
None
None
None
https://bowwowinsurance.com.au/wp-content/uploads/2018/06/icon-promo.svg (https://bowwowinsurance.com.au/wp-content/uploads/2018/06/icon-promo.svg)
None

3

```
#getting dog images
import re
image_tags = dogSoup.findAll('img', {'src':re.compile('.jpg')})
urls = []
# print out image urls
for image_tag in image_tags:
    print(image_tag['src']+'\n')
    urls.append(image_tag['src'])
print(urls)
```

```
https://bowwowinsurance.com.au/wp-content/uploads/2018/10/airedale-terrier-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aireda
le-terrier-700x700.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/akita-700x700.jpg
 (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/akita-700x700.jp
g)

https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_15250
126-Alaskan-Husky-in-front-of-white-background-thumbnail.jpg (https://bowwow
insurance.com.au/wp-content/uploads/2020/09/shutterstock_15250126-Alaskan-Hu
sky-in-front-of-white-background-thumbnail.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/alaskan-malamute-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/alaska
n-malamute-700x700.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/11/american-bulldog-7
00x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/11/americ
an-bulldog-700x700.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2021/04/shutterstock_14123
64536-THUMBNAIL-The-American-Eskimo-Dog-is-a-breed-of-companion-dog-originat
ing-in-Germany-The-American-Eskimo-is-a-member-of-the-Spitz-family.jpg (http
s://bowwowinsurance.com.au/wp-content/uploads/2021/04/shutterstock_141236453
6-THUMBNAIL-The-American-Eskimo-Dog-is-a-breed-of-companion-dog-originating-
in-Germany-The-American-Eskimo-is-a-member-of-the-Spitz-family.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2021/02/shutterstock_14448
33281-American-Foxhound-in-a-public-park-thumbnail-Bow-Wow-Meow-Pet-Insuranc
e.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2021/02/shutterstoc
k_1444833281-American-Foxhound-in-a-public-park-thumbnail-Bow-Wow-Meow-Pet-I
nsurance.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/american-staffords
hire-terrier-amstaff-american-staffy-700x700.jpg (https://bowwowinsurance.co
m.au/wp-content/uploads/2018/10/american-staffordshire-terrier-amstaff-ameri
can-staffy-700x700.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aussie-bulldog-thu
mb.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aussie-bul
ldog-thumb.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/blue-heeler-austra
lian-cattledog-700x700.jpg (https://bowwowinsurance.com.au/wp-content/upload
s/2018/10/blue-heeler-australian-cattledog-700x700.jpg)
```

https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_1635272731-Cute-red-abricot-Australian-Cobberdog-Labradoodle-dog-pup-sitting-up-with-one-paw-high-in-air.-Mouth-closed.-Isolated-on-white-background.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_1635272731-Cute-red-abricot-Australian-Cobberdog-Labradoodle-dog-pup-sitting-up-with-one-paw-high-in-air.-Mouth-closed.-Isolated-on-white-background.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/australian-kelpie-isolated-thumb-700x700.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/australian-kelpie-isolated-thumb-700x700.jpg)

https://bowwowinsurance.com.au/wp-content/uploads/2018/10/dog-licking-face-of-guy-in-park.jpg (https://bowwowinsurance.com.au/wp-content/uploads/2018/10/dog-licking-face-of-guy-in-park.jpg)

['https://bowwowinsurance.com.au/wp-content/uploads/2018/10/airedale-terrier-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/akita-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_15250126-Alaskan-Husky-in-front-of-white-background-thumbnail.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/alaskan-malamute-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/11/american-bulldog-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2021/04/shutterstock_1412364536-THUMBNAIL-The-American-Eskimo-Dog-is-a-breed-of-companion-dog-originating-in-Germany-The-American-Eskimo-is-a-member-of-the-Spitz-family.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2021/02/shutterstock_1444833281-American-Foxhound-in-a-public-park-thumbnail-Bow-Wow-Meow-Pet-Insurance.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/american-staffordshire-terrier-amstaff-american-staffy-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/aussie-bulldog-thumb.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/blue-heeler-australian-cattledog-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2020/09/shutterstock_1635272731-Cute-red-abricot-Australian-Cobberdog-Labradoodle-dog-pup-sitting-up-with-one-paw-high-in-air.-Mouth-closed.-Isolated-on-white-background.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/australian-kelpie-isolated-thumb-700x700.jpg', 'https://bowwowinsurance.com.au/wp-content/uploads/2018/10/dog-licking-face-of-guy-in-park.jpg']

In [5]:
```
import urllib.request

for i in range(len(urls)):
    urllib.request.urlretrieve(urls[i], "./dogs/dog"+str(i)+".jpg")
```

6

```
#Extracting images from catUrl

image_tags = catSoup.findAll('img')
# print out image urls
for image_tag in image_tags:
    print(image_tag.get('src'))
```

```
//upload.wikimedia.org/wikipedia/en/thumb/9/94/Symbol_support_vote.svg/19px-Symbol_support_vote.svg.png
//upload.wikimedia.org/wikipedia/en/thumb/1/1b/Semi-protection-shackle.svg/20px-Semi-protection-shackle.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/4/47/Sound-icon.svg/20px-Sound-icon.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/0/0b/Cat_poster_1.jpg/260px-Cat_poster_1.jpg
//upload.wikimedia.org/wikipedia/commons/7/74/Red_Pencil_Icon.png
//upload.wikimedia.org/wikipedia/commons/thumb/2/21/Wild-domestic-hybrid_cat_skulls.png/220px-Wild-domestic-hybrid_cat_skulls.png
//upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Tomb_of_Nakht_%287%29.jpg/220px-Tomb_of_Nakht_%287%29.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/5/5a/Scheme_cat_anatomy.svg/220px-Scheme_cat_anatomy.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/e/ef/Cat_skull.jpg/220px-Cat_skull.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/b/bc/Shed_domestic_cat_claw_sheaths.tiff/lossless-page1-110px-Shed_domestic_cat_claw_sheaths.tiff.png
//upload.wikimedia.org/wikipedia/commons/thumb/a/a1/BIOASTRONAUTICS_RESEARCH_Gov.archives.arc.68700.ogv/220px-seek%3D240-BIOASTRONAUTICS_RESEARCH_Gov.archives.arc.68700.ogv.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/7/76/TapetumLucidum.JPG/220px-TapetumLucidum.JPG
//upload.wikimedia.org/wikipedia/commons/thumb/b/bb/Kittyply_edit1.jpg/220px-Kittyply_edit1.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/4/4f/Felis_silvestris_catus_lying_on_rice_straw.jpg/220px-Felis_silvestris_catus_lying_on_rice_straw.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/5/5e/Domestic_Cat_Face_Shot.jpg/220px-Domestic_Cat_Face_Shot.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/d/da/Cat_tongue_macro.jpg/220px-Cat_tongue_macro.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/3/3b/Gato_enervado_pola_presencia_dun_can.jpg/220px-Gato_enervado_pola_presencia_dun_can.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Kot_z_mysz%C4%85.jpg/220px-Kot_z_mysz%C4%85.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Play_fight_between_cats.webmhd.webm/220px-seek%3D4-Play_fight_between_cats.webmhd.webm.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/e/e7/Cats_having_sex_in_Israel.jpg/220px-Cats_having_sex_in_Israel.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/8/84/Radiography_of_a_pregnant_cat.jpg/170px-Radiography_of_a_pregnant_cat.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/b/b5/1dayoldkitten.JPG/220px-1dayoldkitten.JPG
//upload.wikimedia.org/wikipedia/commons/thumb/b/b6/Felis_catus-cat_on_snow.jpg/220px-Felis_catus-cat_on_snow.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Feral_cat_Virginia_crop.jpg/170px-Feral_cat_Virginia_crop.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Mainecoon-lap.jpg/220px-Mai
```

necoon-lap.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/8/87/Louvre_egyptologie_21.jpg/200px-Louvre_egyptologie_21.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/Cat_birds_MAN_Napoli_Inv9993.jpg/200px-Cat_birds_MAN_Napoli_Inv9993.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/PSM_V37_D105_English_tabby_cat.jpg/200px-PSM_V37_D105_English_tabby_cat.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/0/0c/Black_Cat_%287983739954%29.jpg/220px-Black_Cat_%287983739954%29.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Cat03.jpg/28px-Cat03.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/1/18/Okapi2.jpg/32px-Okapi2.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/4/44/Wapiti_from_Wagon_Trails.jpg/32px-Wapiti_from_Wagon_Trails.jpg
//upload.wikimedia.org/wikipedia/commons/thumb/4/47/Sound-icon.svg/45px-Sound-icon.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/9/99/Wiktionary-logo-en-v2.svg/16px-Wiktionary-logo-en-v2.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/d/df/Wikispecies-logo.svg/14px-Wikispecies-logo.svg.png
//upload.wikimedia.org/wikipedia/en/thumb/4/4a/Commons-logo.svg/12px-Commons-logo.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/d/df/Wikibooks-logo-en-noslogan.svg/16px-Wikibooks-logo-en-noslogan.svg.png
//upload.wikimedia.org/wikipedia/commons/thumb/f/fa/Wikiquote-logo.svg/13px-Wikiquote-logo.svg.png
//upload.wikimedia.org/wikipedia/en/thumb/9/96/Symbol_category_class.svg/16px-Symbol_category_class.svg.png
//upload.wikimedia.org/wikipedia/en/thumb/e/e2/Symbol_portal_class.svg/16px-Symbol_portal_class.svg.png
//upload.wikimedia.org/wikipedia/en/thumb/8/8a/OOjs_UI_icon_edit-ltr-progressive.svg/10px-OOjs_UI_icon_edit-ltr-progressive.svg.png
//en.wikipedia.org/wiki/Special:CentralAutoLogin/start?type=1x1
/static/images/footer/wikimedia-button.png
/static/images/footer/poweredby_mediawiki_88x31.png

7

```
#getting cat images
import re
image_tags = catSoup.findAll('img', {'src':re.compile('.jpg')})
catUrls = []
# print out image urls
for image_tag in image_tags:
    print(image_tag['src']+'\n')
    catUrls.append(image_tag['src'])
print(catUrls)
```

```
//upload.wikimedia.org/wikipedia/commons/thumb/0/0b/Cat_poster_1.jpg/260px-Cat_
poster_1.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Tomb_of_Nakht_%287%29.jpg/2
20px-Tomb_of_Nakht_%287%29.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/e/ef/Cat_skull.jpg/220px-Cat_sku
ll.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/a/a1/BIOASTRONAUTICS_RESEARCH_Go
v.archives.arc.68700.ogv/220px-seek%3D240-BIOASTRONAUTICS_RESEARCH_Gov.archive
s.arc.68700.ogv.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/b/bb/Kittyply_edit1.jpg/220px-Ki
ttyply_edit1.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/4/4f/Felis_silvestris_catus_lyin
g_on_rice_straw.jpg/220px-Felis_silvestris_catus_lying_on_rice_straw.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/5/5e/Domestic_Cat_Face_Shot.jpg/
220px-Domestic_Cat_Face_Shot.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/d/da/Cat_tongue_macro.jpg/220px-
Cat_tongue_macro.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/3/3b/Gato_enervado_pola_presenci
a_dun_can.jpg/220px-Gato_enervado_pola_presencia_dun_can.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Kot_z_mysz%C4%85.jpg/220px-
Kot_z_mysz%C4%85.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Play_fight_between_cats.web
mhd.webm/220px-seek%3D4-Play_fight_between_cats.webmhd.webm.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/e/e7/Cats_having_sex_in_Israel.j
pg/220px-Cats_having_sex_in_Israel.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/8/84/Radiography_of_a_pregnant_c
at.jpg/170px-Radiography_of_a_pregnant_cat.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/b/b6/Felis_catus-cat_on_snow.jp
g/220px-Felis_catus-cat_on_snow.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Feral_cat_Virginia_crop.jp
g/170px-Feral_cat_Virginia_crop.jpg
```

//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Mainecoon-lap.jpg/220px-Mainecoon-lap.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/8/87/Louvre_egyptologie_21.jpg/200px-Louvre_egyptologie_21.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/Cat_birds_MAN_Napoli_Inv9993.jpg/200px-Cat_birds_MAN_Napoli_Inv9993.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/PSM_V37_D105_English_tabby_cat.jpg/200px-PSM_V37_D105_English_tabby_cat.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/0/0c/Black_Cat_%287983739954%29.jpg/220px-Black_Cat_%287983739954%29.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Cat03.jpg/28px-Cat03.jpg

//upload.wikimedia.org/wikipedia/commons/thumb/1/18/Okapi2.jpg/32px-Okapi2.jpg

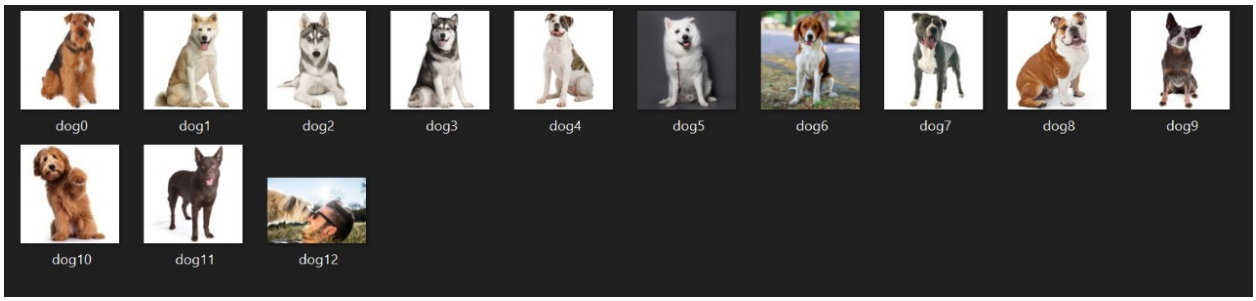
//upload.wikimedia.org/wikipedia/commons/thumb/4/44/Wapiti_from_Wagon_Trails.jpg/32px-Wapiti_from_Wagon_Trails.jpg

['//upload.wikimedia.org/wikipedia/commons/thumb/0/0b/Cat_poster_1.jpg/260px-Cat_poster_1.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/9/9e/Tomb_of_Nakht_%287%29.jpg/220px-Tomb_of_Nakht_%287%29.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/e/ef/Cat_skull.jpg/220px-Cat_skull.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/a/a1/BIOASTRONAUTICS_RESEARCH_Gov.archives.arc.68700.ogv/220px-seek%3D240-BIOASTRONAUTICS_RESEARCH_Gov.archives.arc.68700.ogv.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/b/bb/Kittyply_edit1.jpg/220px-Kittyply_edit1.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/4/4f/Felis_silvestris_catus_lying_on_rice_straw.jpg/220px-Felis_silvestris_catus_lying_on_rice_straw.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/5/5e/Domestic_Cat_Face_Shot.jpg/220px-Domestic_Cat_Face_Shot.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/d/da/Cat_tongue_macro.jpg/220px-Cat_tongue_macro.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/3/3b/Gato_enervado_pola_presencia_dun_can.jpg/220px-Gato_enervado_pola_presencia_dun_can.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Kot_z_mysz%C4%85.jpg/220px-Kot_z_mysz%C4%85.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Play_fight_between_cats.webmhd.webm/220px-seek%3D4-Play_fight_between_cats.webmhd.webm.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/e/e7/Cats_having_sex_in_Israel.jpg/220px-Cats_having_sex_in_Israel.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/8/84/Radiography_of_a_pregnant_cat.jpg/170px-Radiography_of_a_pregnant_cat.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/b/b6/Felis_catus-cat_on_snow.jpg/220px-Felis_catus-cat_on_snow.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/9/97/Feral_cat_Virginia_crop.jpg/170px-Feral_cat_Virginia_crop.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/4/45/Mainecoon-lap.jpg/220px-Mainecoon-lap.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/8/87/Louvre_egyptologie_21.jpg/200px-Louvre_egyptologie_21.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/Cat_birds_MAN_Napoli_Inv9993.jpg/200px-Cat_birds_MAN_Napoli_Inv9993.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/b/b0/PSM_V37_D105_English_tabby_cat.jpg/200px-PSM_V37_D105_English_tabby_cat.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/0/0c/Black_Cat_%287983739954%29.jpg/220px-Black_Cat_%287983739954%29.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Cat03.jpg/28px-Cat03.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/1/18/Okapi2.jpg/32px-Okapi2.jpg', '//upload.wikimedia.org/wikipedia/commons/thumb/4/44/Wapiti_from_Wagon_Trails.jpg/32px-Wapiti_from_Wagon_Trails.jpg']

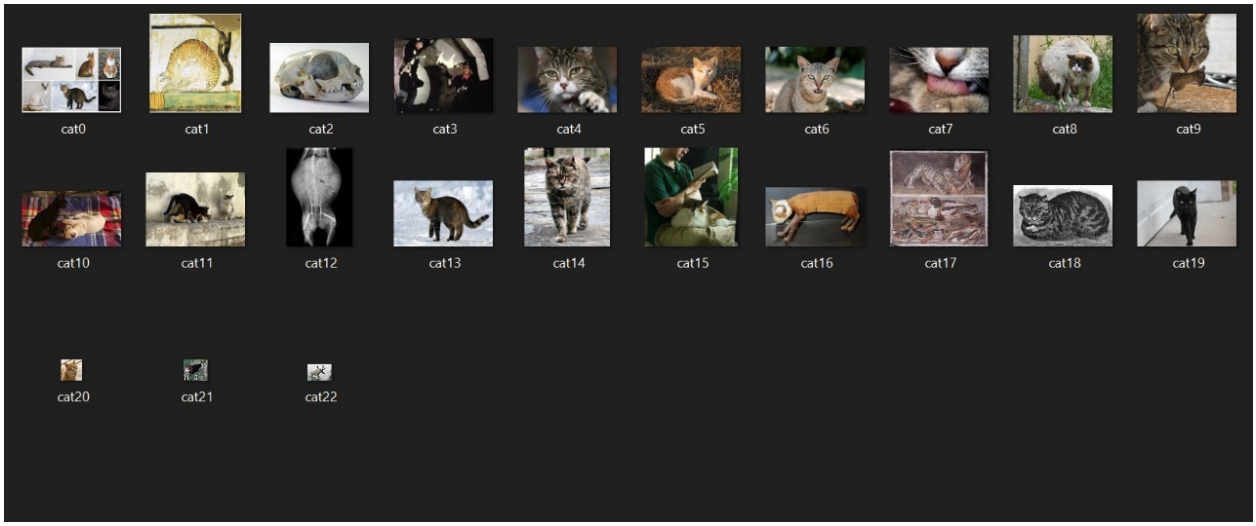
In [9]:
```
import urllib.request

for i in range(len(catUrls)):
    urllib.request.urlretrieve("https:" + catUrls[i], "./cats/cat"+str(i)+".jpg")
```

Dogs folder Created:



Cats folder Created:



***** END *****