



Program Code: J620-002-4:2020

Program Name: FRONT-END SOFTWARE DEVELOPMENT

Title : Introduction to Web Scrapping

Name: Ooi Caaron

IC Number: 990701-07-5837

Date : 3/7/23

Introduction : Learning Web Scrapping, get the data from website

Conclusion : Still need to practice more

Here is the guidance in learning the following topics:

1. Overview of Web Applications & Web Scraping
2. Regular Expression
3. BeautifulSoup
4. Selenium

The classes shall be conducted with

1. Explanation
2. Online Demo
3. Q&A
4. Exercises

Tools - Python Libraries

1. Requests
2. Regex
3. BeautifulSoup
4. Selenium

Web Applications

What are Web Applications?

According to Wikipedia

'In computing, a web application or web app is a client–server computer program that the client (including the user interface and client-side logic) runs in a web browser. Common web applications include webmail, online retail sales, online banking, and online auctions.'

Reference:

https://en.wikipedia.org/wiki/Web_application (https://en.wikipedia.org/wiki/Web_application)

<https://developer.mozilla.org/en-US/docs/Learn> (<https://developer.mozilla.org/en-US/docs/Learn>)

Clients and servers

Computers connected to the web are called **clients** and **servers**. A simplified diagram of how they interact might look like this:



- Clients are the typical web user's internet-connected devices (for example, your computer connected to your Wi-Fi, or your phone connected to your mobile network) and web-accessing software available on those devices (usually a web browser like Firefox or Chrome).
- Servers are computers that store webpages, sites, or apps. When a client device wants to access a webpage, a copy of the webpage is downloaded from the server onto the client machine to be displayed in the user's web browser.



Anatomy of a Web Page

HTML

CSS

Javascript

Show some examples

In [2]:

```
webpage = '''
<!DOCTYPE html>
<html>
  <head>
    <title>This is a title</title>
  </head>
  <body>
    <p>Hello world!</p>
  </body>
</html>
'''
```

Web Scrapping

Web Scrapping is a technique to extract the data from the web pages in an **automated way**.

A web scrapping **script** can load and extract the **data** from multiple pages.

A web scrapping script contains Python codes and required libraries to perform the task.

The first library needed is **Requests**

Getting Started

Install the Request library

pip3 install requests

OR

conda install -c conda-forge requests

Ref:

<https://anaconda.org/conda-forge/request> (<https://anaconda.org/conda-forge/request>).

Requests

Requests (handles HTTP sessions and makes HTTP requests).

import requests

In [3]:

```
import requests

url='https://www.thestar.com.my/news/nation/2020/03/23/covid-19-current-situation-in-mal'

page = requests.get(url)

page.status_code
```

Out[3]:

200

Status code

200 OK

<https://developer.mozilla.org/en-US/docs/Web/HTTP/Status> (<https://developer.mozilla.org/en-US/docs/Web/HTTP/Status>)

In [4]:

```
# print the returned page as string

page.text
```

Out[4]:

```
'<!--default_base.blade.php-->\n<!DOCTYPE html>\n<html lang="en">\n<head>\n  \t<title>Covid-19: Cases up by 2,875 bringing total to 384,688 (updated daily) | The Star</title>\n  \t<link rel="icon" type="image/png" href="https://cdn.thestar.com.my/Themes/img/favicon.ico" />\n  \n  \n  <!--START: common.blade.php-->\n<meta http-equiv="Content-Type" content="text/html; charset=utf-8"/><script type="text/javascript">(window.NREUM||(NREUM={})).init={privacy:{cookies_enabled:false},ajax:{deny_list:["bam.nr-data.net"]}};(window.NREUM||(NREUM={})).loader_config={xpid:"VgIEVF9QChADU1hQAAUGUFc=",licenseKey:"ef8b08f3e1",applicationID:"379274163"};/*! For license information please see nr-loader-full-1.236.0.min.js.LICENSE.txt */\n(()=>{"use strict";var e,t,r={5763:(e,t,r)=>{r.d(t,{P_:(()=>f,Mt:(()=>g,C5:(()=>s,DL:(()=>m,OP:(()=>_,lF:(()=>T,Yu:(()=>y,Dg:(()=>h,CX:(()=>c,GE:(()=>b,sU:(()=>D)});var n=r(8632),i=r(9567);const o={beacon:n.ce.beacon,errorBeacon:n.ce.errorBeacon,licenseKey:void 0,applicationID:void 0,sa:void 0,queueTime:void 0,applicationTime:void 0,ttGuid:void 0,user:void 0,account:void 0,product:void 0,extra:void 0,jsAttributes:{},userAttributes:void 0,atts:void 0,transactionName:void 0,tNamePlain:void 0}.a={}:function s(e){if(!e)throw new Error("All info objec
```

In [5]:

print the returned page as bytes

page.content

Out[5]:

```
b'<!--default_base.blade.php-->\n<!DOCTYPE html>\n<html lang="en">\n<head>\n    \t<title>Covid-19: Cases up by 2,875 bringing total to 384,688
(updated daily) | The Star</title>\n    \t<link rel="icon" type="image/pn
g" href="https://cdn.thestar.com.my/Themes/img/favicon.ico" />\n    \n
\n    <!--START: common.blade.php-->\n<meta http-equiv="Content-Type" c
ontent="text/html; charset=utf-8"/><script type="text/javascript">(windo
w.NREUM||(NREUM={})).init={privacy:{cookies_enabled:false},ajax:{deny_l
ist:["bam.nr-data.net"]}};(window.NREUM||(NREUM={})).loader_config={xpi
d:"VgIEVF9QChADU1hQAAUGUFc=",licenseKey:"ef8b08f3e1",applicationID:"379
274163"};;/*! For license information please see nr-loader-full-1.236.
0.min.js.LICENSE.txt */\n((()=>{"use strict";var e,t,r={5763:(e,t,r)=>
{r.d(t,{P_:(()=>f,Mt:(()=>g,C5:(()=>s,DL:(()=>m,OP:(()=>_,lF:(()=>T,Yu:(()=>y,
Dg:(()=>h,CX:(()=>c,GE:(()=>b,sU:(()=>D)});var n=r(8632),i=r(9567);const o=
{beacon:n.ce.beacon,errorBeacon:n.ce.errorBeacon,licenseKey:void 0,appl
icationID:void 0,sa:void 0,queueTime:void 0,applicationTime:void 0,ttGu
id:void 0,user:void 0,account:void 0,product:void 0,extra:void 0,jsAttr
ibutes:{},userAttributes:void 0,atts:void 0,transactionName:void 0,tNam
ePlain:void 0}.a={}:function s(e){if(!e)throw new Error("All info objec
```

In [6]:

print the returned page as bytes

page.encoding

Out[6]:

'UTF-8'

Compare

Open your Web Browser and compare the Source Code shown there and here

Understanding the Web Page

Where is the ?

HTML

CSS

Javascript

Show the Web Browser Developers Tools

Reduce impact

Do not query the webpage all the time

In [7]:

```
# How to save HTML Locally
import requests

def save_html(html, path):
    with open(path, 'wb') as f:
        f.write(html)

url = 'https://www.google.com'

r = requests.get(url)

save_html(r.content, 'google_com')

print(r.content[:100])
```

```
b'<!doctype html><html itemscope="" itemtype="http://schema.org/WebPage" l
ang="en-MY"><head><meta cont'
```

In [9]:

```
# How to open.read HTML from a local file

def open_html(path):
    with open(path, 'rb') as f:
        return f.read()

html = open_html('google_com')
```

In [9]:

html

Out[9]:

```
b'<!doctype html><html itemscope="" itemtype="http://schema.org/WebPag
e" lang="en-MY"><head><meta content="text/html; charset=UTF-8" http-equ
iv="Content-Type"><meta content="/images/branding/googleg/1x/googleg_st
andard_color_128dp.png" itemprop="image"><title>Google</title><script n
once="3JYHcJj+3Q7B9jM8+gCqHw==">(function(){window.google={kEI:'\tr2MYc
3QC0nJ1sQPgPGg6AE\'',kEXPI:'\0,1302536,56873,6058,207,4804,2316,383,246,
5,1354,5250,1122516,1197792,329475,51223,16115,17444,1953,9287,17572,13
26,3533,1361,283,9008,3027,17581,4020,978,13228,3847,4192,6430,7432,143
90,919,5081,885,708,1279,2214,528,149,1103,841,1982,213,4101,108,3406,6
06,2023,1777,520,1704,12966,3227,2845,7,12354,5096,15767,553,908,2,940,
6039,9718,3,576,1014,1,3371,2074,148,11323,2652,4,1252,276,2304,7039,20
57,2627,203,1811,16336,2039,2658,872,6483,32,5615,8013,1591,714,2132,16
786,5824,2533,992,3102,3138,6,908,3,3541,1,16524,283,912,5992,16728,171
5,2,14022,1931,784,255,2870,1680,743,5853,4974,2,3897,992,598,1160,419
2,1528,979,2381,2718,8527,1495,4148,4063,10,2,6,1,7771,4569,5709,278,27
1,1107,1853,1364,2308,91,3692,83,2,1514,2,2920,3147,547,267,3484,1038,6
625,2433,1300,1186,729,1877,702,68,2923,1104,651,2,1298,1156,285,2127,3
40.2.1.159.1078.189.3.2.20.3.2.121.15.736.1794.225.234.284.644.454.130.
```

In [10]:

```
# for ipython notebook display
from IPython.core.display import display, HTML

display(HTML(str(html)))
```

C:\Users\User\AppData\Local\Temp\ipykernel_8188\2352304310.py:2: Deprecati
onWarning: Importing display from IPython.core.display is deprecated since
IPython 7.14, please import from IPython display
from IPython.core.display import display, HTML

b'Google

Search [Images \(https://www.google.com/imghp?hl=en&tab=wi\)](https://www.google.com/imghp?hl=en&tab=wi) [Maps \(https://maps.google.cc](https://maps.google.cc)
[Web History \(http://www.google.com.my/history/optout?hl=en\)](http://www.google.com.my/history/optout?hl=en) | [Settings \(/preferences?hl=en\)](https://preferences?hl=en)



Google

[Advanced search \(/advanced_search?hl=en-MY&authuser=0\)](#)

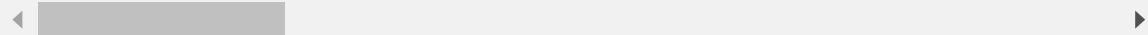
Google Search I'm Feeling Lucky

Google offered in: [Bahasa Melayu \(https://www.google.com/setprefs?sig=0_wv8KuU_6z10IHlpta1SVItEtNhc%3D&hl=ms&source=homepage&sa=X&ved=0ahUKEwixLTuxPH_AhV3r](https://www.google.com/setprefs?sig=0_wv8KuU_6z10IHlpta1SVItEtNhc%3D&hl=ms&source=homepage&sa=X&ved=0ahUKEwixLTuxPH_AhV3r)

[Advertising \(/intl/en/ads/\)](#) [Business Solutions \(http://www.google.com.my/intl/en/services/\)](#) [About Google \(/intl/en/about.html\)](#)

© 2023 - [Privacy \(/intl/en/policies/privacy/\)](#) - [Terms \(/intl/en/policies/terms/\)](#)

,



How to be a good scrapers/bots

Look for robots.txt at the root of the domain.

Website owner explicitly states what bots are allowed to do on their site

In [11]:

```
import requests


url = 'https://www.google.com/robots.txt'

user_agent = 'user-agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (

headers={'User-Agent':user_agent}

r = requests.get(url, headers=headers)

print(r.content)
```




```

b'User-agent: *\nDisallow: /search\nAllow: /search/about\nAllow: /search/s
tatic\nAllow: /search/howsearchworks\nDisallow: /sdch\nDisallow: /groups\n
Disallow: /index.html?\nDisallow: /?\nAllow: /?hl=\nDisallow: /?hl=*&\nAll
ow: /?hl=*&gws_rd=ssl$\nDisallow: /?hl=*&&gws_rd=ssl\nAllow: /?gws_rd=ssl
$\nAllow: /?pt1=true$\nDisallow: /imgres\nDisallow: /u/\nDisallow: /prefer
ences\nDisallow: /setprefs\nDisallow: /default\nDisallow: /m?\nDisallow: /
m/\nAllow: /m/finance\nDisallow: /wml?\nDisallow: /wml/?\nDisallow: /wm
l/search?\nDisallow: /xhtml?\nDisallow: /xhtml/?\nDisallow: /xhtml/search?
\nDisallow: /xml?\nDisallow: /imode?\nDisallow: /imode/?\nDisallow: /imod
e/search?\nDisallow: /jsky?\nDisallow: /jsky/?\nDisallow: /jsky/search?\nD
isallow: /pda?\nDisallow: /pda/?\nDisallow: /pda/search?\nDisallow: /sprin
t_xhtml\nDisallow: /sprint_wml\nDisallow: /pqa\nDisallow: /palm\nDisallow:
/gwt/\nDisallow: /purchases\nDisallow: /local?\nDisallow: /local_url\nDisa
llow: /shihui?\nDisallow: /shihui/\nDisallow: /products?\nDisallow: /produ
ct_\nDisallow: /products_\nDisallow: /products;\nDisallow: /print\nDisallo
w: /books/\nDisallow: /bkshp?*q=*\nDisallow: /books?*q=*\nDisallow: /book
s?*output=*\nDisallow: /books?*pg=*\nDisallow: /books?*jtp=*\nDisallow: /b
ooks?*jscmd=*\nDisallow: /books?*buy=*\nDisallow: /books?*zoom=*\nAllow: /
books?*q=related:*\nAllow: /books?*q=editions:*\nAllow: /books?*q=subject:
*\nAllow: /books/about\nAllow: /booksrightsholders\nAllow: /books?*zoom=1*
\nAllow: /books?*zoom=5*\nAllow: /books/content?*zoom=1*\nAllow: /books/co
ntent?*zoom=5*\nDisallow: /ebooks/\nDisallow: /ebooks?*q=*\nDisallow: /ebo
oks?*output=*\nDisallow: /ebooks?*pg=*\nDisallow: /ebooks?*jscmd=*\nDisall
ow: /ebooks?*buy=*\nDisallow: /ebooks?*zoom=*\nAllow: /ebooks?*q=related:*
\nAllow: /ebooks?*q=editions:*\nAllow: /ebooks?*q=subject:*\nAllow: /ebook
s?*zoom=1*\nAllow: /ebooks?*zoom=5*\nDisallow: /patents?\nDisallow: /paten
ts/download/\nDisallow: /patents/pdf/\nDisallow: /patents/related/\nDisall
ow: /scholar\nDisallow: /citations?\nAllow: /citations?user=\nDisallow: /c
itations?*cstart=\nAllow: /citations?view_op=new_profile\nAllow: /citation
s?view_op=top_venues\nAllow: /scholar_share\nDisallow: /s?\nDisallow: /map
s?\nAllow: /maps?*output=classic*\nAllow: /maps?*file=\nDisallow: /mapstt?
\nDisallow: /mapslt?\nDisallow: /mapabcpoi?\nDisallow: /maphp?\nDisallow:
/mapprint?\nDisallow: /maps/\nAllow: /maps/search/\nAllow: /maps/dir/\nAll
ow: /maps/d/\nAllow: /maps/reserve\nAllow: /maps/about\nAllow: /maps/match
\nDisallow: /maps/api/js/\nAllow: /maps/api/js\nDisallow: /mld?\nDisallow:
/staticmap?\nDisallow: /help/maps/streetview/partners/welcome/\nDisallow:
/help/maps/indoormaps/partners/\nDisallow: /lochp?\nDisallow: /center\nDis
allow: /ie?\nDisallow: /blogsearch/\nDisallow: /blogsearch_feeds\nDisallo
w: /advanced_blog_search\nDisallow: /uds/\nDisallow: /chart?\nDisallow: /t
ransit?\nAllow: /calendar$\nAllow: /calendar/about/\nDisallow: /cale
ndar/\nDisallow: /cl2/feeds/\nDisallow: /cl2/ical/\nDisallow: /coop/direct
ory\nDisallow: /coop/manage\nDisallow: /trends?\nDisallow: /trends/music?
\nDisallow: /trends/hottrends?\nDisallow: /trends/viz?\nDisallow: /trends/
embed.js?\nDisallow: /trends/fetchComponent?\nDisallow: /trends/beta\nDisa
llow: /trends/topics\nDisallow: /musica\nDisallow: /musicad\nDisallow: /mu
sicas\nDisallow: /musicl\nDisallow: /musics\nDisallow: /musicsearch\nDisal
low: /musicsp\nDisallow: /musiclp\nDisallow: /urchin_test/\nDisallow: /mov
ies?\nDisallow: /wapsearch?\nAllow: /safebrowsing/diagnostic\nAllow: /safe
browsing/report_badware/\nAllow: /safebrowsing/report_error/\nAllow: /safe
browsing/report_phish/\nDisallow: /reviews/search?\nDisallow: /orkut/album
s\nDisallow: /cbk\nDisallow: /recharge/dashboard/car\nDisallow: /recharge/
dashboard/static/\nDisallow: /profiles/me\nAllow: /profiles\nDisallow: /s
2/profiles/me\nAllow: /s2/profiles\nAllow: /s2/oz\nAllow: /s2/photos\nAllo
w: /s2/search/social\nAllow: /s2/static\nDisallow: /s2\nDisallow: /transco
nsole/portal/\nDisallow: /gcc/\nDisallow: /aclk\nDisallow: /cse?\nDisallo
w: /cse/home\nDisallow: /cse/panel\nDisallow: /cse/manage\nDisallow: /tbpr
oxy/\nDisallow: /imesync/\nDisallow: /shenghuo/search?\nDisallow: /suppor
t/forum/search?\nDisallow: /reviews/polls/\nDisallow: /hosted/images/\nDis
allow: /ppob/?\nDisallow: /ppob?\nDisallow: /accounts/ClientLogin\nDisallo
w: /accounts/ClientAuth\nDisallow: /accounts/o8\nAllow: /accounts/o8/id\nD
isallow: /topicsearch?q=\nDisallow: /xfx7/\nDisallow: /squared/api\nDisall

```

```

ow: /squared/search\nDisallow: /squared/table\nDisallow: /qnasearch?\nDisa
llow: /app/updates\nDisallow: /sidewiki/entry/\nDisallow: /quality_form?\n
Disallow: /labs/popgadget/search\nDisallow: /buzz/post\nDisallow: /compres
siontest/\nDisallow: /analytics/feeds/\nDisallow: /analytics/partners/comm
ents/\nDisallow: /analytics/portal/\nDisallow: /analytics/uploads/\nAllow:
/alerts/manage\nAllow: /alerts/remove\nDisallow: /alerts/\nAllow: /alerts/
$\nDisallow: /ads/search?\nDisallow: /ads/plan/action_plan?\nDisallow: /ad
s/plan/api/\nDisallow: /ads/hotels/partners\nDisallow: /phone/compare/?\nD
isallow: /travel/clk\nDisallow: /travel/flights/s/\nDisallow: /hotelfinde
r/rpc\nDisallow: /hotels/rpc\nDisallow: /commercesearch/services/\nDisallo
w: /evaluation/\nDisallow: /chrome/browser/mobile/tour\nDisallow: /compar
e/*/apply*\nDisallow: /forms/perks/\nDisallow: /shopping/suppliers/search
\nDisallow: /ct/\nDisallow: /edu/cs4hs/\nDisallow: /trustedstores/s/\nDisa
llow: /trustedstores/tm2\nDisallow: /trustedstores/verify\nDisallow: /adwo
rds/proposal\nDisallow: /shopping?*\nDisallow: /shopping/product/\nDisallo
w: /shopping/seller\nDisallow: /shopping/ratings/account/metrics\nDisallo
w: /shopping/ratings/merchant/immersivedetails\nDisallow: /shopping/review
er\nDisallow: /about/careers/applications/\nDisallow: /about/careers/appli
cations-a/\nDisallow: /landing/signout.html\nDisallow: /webmasters/sitemap
s/ping?\nDisallow: /ping?\nDisallow: /gallery/\nDisallow: /landing/now/ont
ap/\nAllow: /searchhistory/\nAllow: /maps/reserve\nAllow: /maps/reserve/pa
rtners\nDisallow: /maps/reserve/api/\nDisallow: /maps/reserve/search\nDisa
llow: /maps/reserve/bookings\nDisallow: /maps/reserve/settings\nDisallow:
/maps/reserve/manage\nDisallow: /maps/reserve/payment\nDisallow: /maps/res
erve/receipt\nDisallow: /maps/reserve/sellersignup\nDisallow: /maps/reserv
e/payments\nDisallow: /maps/reserve/feedback\nDisallow: /maps/reserve/term
s\nDisallow: /maps/reserve/m/\nDisallow: /maps/reserve/b/\nDisallow: /map
s/reserve/partner-dashboard\nDisallow: /about/views/\nDisallow: /intl/*/ab
out/views/\nDisallow: /local/cars\nDisallow: /local/cars/\nDisallow: /loca
l/dealership/\nDisallow: /local/dining/\nDisallow: /local/place/products/
\nDisallow: /local/place/reviews/\nDisallow: /local/place/rap/\nDisallow:
/local/tab/\nDisallow: /localservices/*\nAllow: /finance\nAllow: /js/\nDis
allow: /nonprofits/account/\nDisallow: /fbx\nDisallow: /viewer\nDisallow:
/landing/cmsnext-root/\n\n# AdsBot\nUser-agent: AdsBot-Google\nDisallow: /
maps/api/js/\nAllow: /maps/api/js\nDisallow: /maps/api/place/js/\nDisallo
w: /maps/api/staticmap\nDisallow: /maps/api/streetview\n\n# Crawlers of ce
rtain social media sites are allowed to access page markup when google.co
m/imgres* links are shared. To learn more, please contact images-robots-al
lowlist@google.com.\nUser-agent: Twitterbot\nAllow: /imgres\nAllow: /searc
h\nDisallow: /groups\nDisallow: /hosted/images/\nDisallow: /m/\n\nUser-age
nt: facebookexternalhit\nAllow: /imgres\nAllow: /search\nDisallow: /groups
\nDisallow: /hosted/images/\nDisallow: /m/\n\nSitemap: https://www.google.com/sitemap.xml\n' (https://www.google.com/sitemap.xml\n')

```

In [12]:

```
# Change to Byte to String  
str(r.content, 'utf-8')
```

Out[12]:

```
'User-agent: *\nDisallow: /search\nAllow: /search/about\nAllow: /search/st
atic\nAllow: /search/howsearchworks\nDisallow: /sdch\nDisallow: /groups\nD
isallow: /index.html?\nDisallow: /?\nAllow: /?hl=\nDisallow: /?hl=*&\nAllo
w: /?hl=*&gws_rd=ssl$\nDisallow: /?hl=*&&gws_rd=ssl\nAllow: /?gws_rd=ssl
$\nAllow: /?pt1=true$\nDisallow: /imgres\nDisallow: /u/\nDisallow: /prefer
ences\nDisallow: /setprefs\nDisallow: /default\nDisallow: /m?\nDisallow: /
m/\nAllow: /m/finance\nDisallow: /wml?\nDisallow: /wml/?\nDisallow: /wm
l/search?\nDisallow: /xhtml?\nDisallow: /xhtml/?\nDisallow: /xhtml/search?
\nDisallow: /xml?\nDisallow: /imode?\nDisallow: /imode/?\nDisallow: /imod
e/search?\nDisallow: /jsky?\nDisallow: /jsky/?\nDisallow: /jsky/search?\nD
isallow: /pda?\nDisallow: /pda/?\nDisallow: /pda/search?\nDisallow: /sprin
t_xhtml\nDisallow: /sprint_wml\nDisallow: /pqa\nDisallow: /palm\nDisallow:
/gwt/\nDisallow: /purchases\nDisallow: /local?\nDisallow: /local_url\nDisa
llow: /shihui?\nDisallow: /shihui/\nDisallow: /products?\nDisallow: /produ
ct_\nDisallow: /products_\nDisallow: /products;\nDisallow: /print\nDisallo
w: /books/\nDisallow: /bkshp?*q=*\nDisallow: /books?*q=*\nDisallow: /book
s?*output=*\nDisallow: /books?*pg=*\nDisallow: /books?*jtp=*\nDisallow: /b
ooks?*jscmd=*\nDisallow: /books?*buy=*\nDisallow: /books?*zoom=*\nAllow: /
books?*q=related:*\nAllow: /books?*q=editions:*\nAllow: /books?*q=subject:
*\nAllow: /books/about\nAllow: /booksrightsholders\nAllow: /books?*zoom=1*
\nAllow: /books?*zoom=5*\nAllow: /books/content?*zoom=1*\nAllow: /books/co
ntent?*zoom=5*\nDisallow: /ebooks/\nDisallow: /ebooks?*q=*\nDisallow: /ebo
oks?*output=*\nDisallow: /ebooks?*pg=*\nDisallow: /ebooks?*jscmd=*\nDisall
ow: /ebooks?*buy=*\nDisallow: /ebooks?*zoom=*\nAllow: /ebooks?*q=related:*
\nAllow: /ebooks?*q=editions:*\nAllow: /ebooks?*q=subject:*\nAllow: /eboo
k?*zoom=1*\nAllow: /ebooks?*zoom=5*\nDisallow: /patents?\nDisallow: /paten
ts/download/\nDisallow: /patents/pdf/\nDisallow: /patents/related/\nDisall
ow: /scholar\nDisallow: /citations?\nAllow: /citations?user=\nDisallow: /c
itations?*cstart=\nAllow: /citations?view_op=new_profile\nAllow: /citation
s?view_op=top_venues\nAllow: /scholar_share\nDisallow: /s?\nDisallow: /map
s?\nAllow: /maps?*output=classic*\nAllow: /maps?*file=\nDisallow: /mapstt?
\nDisallow: /mapslt?\nDisallow: /mapabcpoi?\nDisallow: /maphp?\nDisallow:
/mapprint?\nDisallow: /maps/\nAllow: /maps/search/\nAllow: /maps/dir/\nAll
ow: /maps/d/\nAllow: /maps/reserve\nAllow: /maps/about\nAllow: /maps/match
\nDisallow: /maps/api/js/\nAllow: /maps/api/js\nDisallow: /mld?\nDisallow:
/staticmap?\nDisallow: /help/maps/streetview/partners/welcome/\nDisallow:
/help/maps/indoormaps/partners/\nDisallow: /lochp?\nDisallow: /center\nDis
allow: /ie?\nDisallow: /blogsearch/\nDisallow: /blogsearch_feeds\nDisallo
w: /advanced_blog_search\nDisallow: /uds/\nDisallow: /chart?\nDisallow: /t
ransit?\nAllow: /calendar$\nAllow: /calendar/about/\nDisallow: /cale
ndar/\nDisallow: /cl2/feeds/\nDisallow: /cl2/ical/\nDisallow: /coop/direct
ory\nDisallow: /coop/manage\nDisallow: /trends?\nDisallow: /trends/music?
\nDisallow: /trends/hottrends?\nDisallow: /trends/viz?\nDisallow: /trends/
embed.js?\nDisallow: /trends/fetchComponent?\nDisallow: /trends/beta\nDisa
llow: /trends/topics\nDisallow: /musica\nDisallow: /musicad\nDisallow: /mu
sicas\nDisallow: /musicl\nDisallow: /musics\nDisallow: /musicsearch\nDisal
low: /musicsp\nDisallow: /musiclp\nDisallow: /urchin_test/\nDisallow: /mov
ies?\nDisallow: /wapsearch?\nAllow: /safebrowsing/diagnostic\nAllow: /safe
browsing/report_badware/\nAllow: /safebrowsing/report_error/\nAllow: /safe
browsing/report_phish/\nDisallow: /reviews/search?\nDisallow: /orkut/album
s\nDisallow: /cbk\nDisallow: /recharge/dashboard/car\nDisallow: /recharge/
dashboard/static/\nDisallow: /profiles/me\nAllow: /profiles\nDisallow: /s
2/profiles/me\nAllow: /s2/profiles\nAllow: /s2/oz\nAllow: /s2/photos\nAllo
w: /s2/search/social\nAllow: /s2/static\nDisallow: /s2\nDisallow: /transco
nsole/portal/\nDisallow: /gcc/\nDisallow: /aclk\nDisallow: /cse?\nDisallo
w: /cse/home\nDisallow: /cse/panel\nDisallow: /cse/manage\nDisallow: /tbpr
oxy/\nDisallow: /imesync/\nDisallow: /shenghuo/search?\nDisallow: /suppor
t/forum/search?\nDisallow: /reviews/polls/\nDisallow: /hosted/images/\nDis
allow: /ppob/?\nDisallow: /ppob?\nDisallow: /accounts/ClientLogin\nDisallo
w: /accounts/ClientAuth\nDisallow: /accounts/o8\nAllow: /accounts/o8/id\nD
isallow: /topicsearch?q=\nDisallow: /xfx7/\nDisallow: /squared/api\nDisall
```

```

ow: /squared/search\nDisallow: /squared/table\nDisallow: /qnasearch?\nDisa
in[13]:
flow: /app/updates\nDisallow: /sidewiki/entry/\nDisallow: /quality_form?\n
# Print the new line
Disallow: /labs/pogadget/search\nDisallow: /buzz/post\nDisallow: /compres
siontest/\nDisallow: /analytics/feeds/\nDisallow: /analytics/partners/comm
ents/\nDisallow: /analytics/portal/\nDisallow: /analytics/uploads/\nAllow:
print(str(r.content, utf-8))
/alerts/manage\nAllow: /alerts/remove\nDisallow: /alerts/\nAllow: /alerts/
$/\nDisallow: /ads/search?\nDisallow: /ads/plan/action_plan?\nDisallow: /ad
s/plan/api/\nDisallow: /ads/hotels/partners\nDisallow: /phone/compare/?\nD
User-agent:
Disallow: /travel/clk\nDisallow: /travel/flights/s/\nDisallow: /hotelfinde
Disallow: /search
Disallow: /hotels/rpc\nDisallow: /commercesearch/services/\nDisallo
Allow: /search/about
Allow: /evaluation/\nDisallow: /chrome/browser/mobile/tour\nDisallow: /compar
e/*\nDisallow: /search/static
Allow: /apply*\nDisallow: /forms/perks/\nDisallow: /shopping/suppliers/search
Disallow: /search/howsearchworks
Disallow: /ct/\nDisallow: /edu/cs4hs/\nDisallow: /trustedstores/s/\nDisa
Disallow: /sdch
Disallow: /trustedstores/tm2\nDisallow: /trustedstores/verify\nDisallow: /adwo
Disallow: /groups
Disallow: /proposals\nDisallow: /shopping?*\nDisallow: /shopping/product/\nDisallo
Disallow: /index.html?
w: /shopping/seller\nDisallow: /shopping/ratings/account/metrics\nDisallo
Disallow: /
w: /shopping/ratings/merchant/immersivedetails\nDisallow: /shopping/review
er\nDisallow: /about/careers/applications/\nDisallow: /about/careers/appli
cations-a/\nDisallow: /landing/signout.html\nDisallow: /webmasters/sitemap
Allow: /?hl=*&gws_rd=ssl$
Disallow: /ping?\nDisallow: /ping?\nDisallow: /gallery/\nDisallow: /landing/now/ont
Disallow: /?hl=*&gws_rd=ssl$
ap/\nAllow: /searchhistory/\nAllow: /maps/reserve\nAllow: /maps/reserve/pa
rtners\nDisallow: /maps/reserve/api/\nDisallow: /maps/reserve/search\nDisa
Allow: /?ptl=true$
Allow: /maps/reserve/bookings\nDisallow: /maps/reserve/settings\nDisallow:
Disallow: /imgres
/maps/reserve/manage\nDisallow: /maps/reserve/payment\nDisallow: /maps/res
Disallow: /u/
erve/receipt\nDisallow: /maps/reserve/sellerssignup\nDisallow: /maps/reserv
e/preferences
e/payments\nDisallow: /maps/reserve/feedback\nDisallow: /maps/reserve/term
Disallow: /setprefs
s\nDisallow: /maps/reserve/m/\nDisallow: /maps/reserve/b/\nDisallow: /map
s/reserve/partner-dashboard\nDisallow: /about/views/\nDisallow: /intl/*\nab
out/views/\nDisallow: /local/cars\nDisallow: /local/cars/\nDisallow: /loca
l/dealership/\nDisallow: /local/dining/\nDisallow: /local/place/products/
\nDisallow: /local/place/reviews/\nDisallow: /local/place/rap/\nDisallow:
/local/tab/\nDisallow: /localservices/*\nAllow: /finance\nAllow: /js/\nDis
allow: /nonprofits/account/\nDisallow: /fbx\nDisallow: /uviewer\nDisallow:
/landing/cmsnext-root/\n\n# AdsBot\nUser-agent: AdsBot-Google\nDisallow: /
maps/api/js/\nAllow: /maps/api/js\nDisallow: /maps/api/place/js/\nDisallo
w: /maps/api/staticmap\nDisallow: /maps/api/streetview\n\n# Crawlers of ce
rtain social media sites are allowed to access page markup when google.co
m/imgres* links are shared. To learn more, please contact images-robots-al
lowlist@google.com.\nUser-agent: Twitterbot\nAllow: /imgres\nAllow: /searc
h\nDisallow: /groups\nDisallow: /hosted/images/\nDisallow: /m/\n\nUser-age
nt: facebookexternalhit\nAllow: /imgres\nAllow: /search\nDisallow: /groups
\nDisallow: /hosted/images/\nDisallow: /m/\n\nSitemap: https://www.google.
com/sitemap.xml\n' (https://www.google.com/sitemap.xml\n')

```

In [14]:

```
import requests

url = 'https://www.google.com/robots.txt'

user_agent = 'user-agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (
headers={'User-Agent':user_agent}

r = requests.get(url, headers=headers)

print(r.text)
```

```
User-agent: *
Disallow: /search
Allow: /search/about
Allow: /search/static
Allow: /search/howsearchworks
Disallow: /sdch
Disallow: /groups
Disallow: /index.html?
Disallow: /?
Allow: /?hl=
Disallow: /?hl=*%
Allow: /?hl=*%gws_rd=ssl$
Disallow: /?hl=*%*%gws_rd=ssl
Allow: /?gws_rd=ssl$
Allow: /?pt1=true$
Disallow: /imgres
Disallow: /u/
Disallow: /preferences
Disallow: /setprefs
```

Using r.text

Requests makes educated guesses about the encoding of the response based on the HTTP headers.

The text encoding guessed by Requests is used when you access r.text.

You can find out what encoding Requests is using, and change it, using the r.encoding property: