

WEB SCRAPING FROM GOOGLE IMAGE

Menggunakan Anaconda Prompt
ft. Windows OS

by Arifa Nurul Fadlila





Arifa Nurul Fadlila (Ifa)

Seorang pemimpi

Mendambakan berjuta mimpi untuk masa depan
yang lebih baik.



arifanf



111201912160@mhs.dinus.ac.id



arifanf



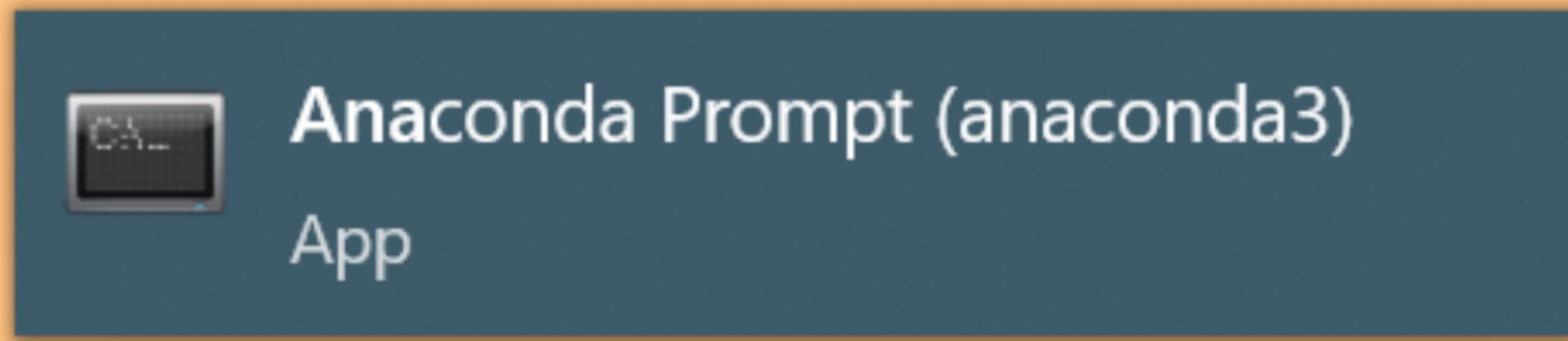
Pertama dan Harus dilakukan!

Install

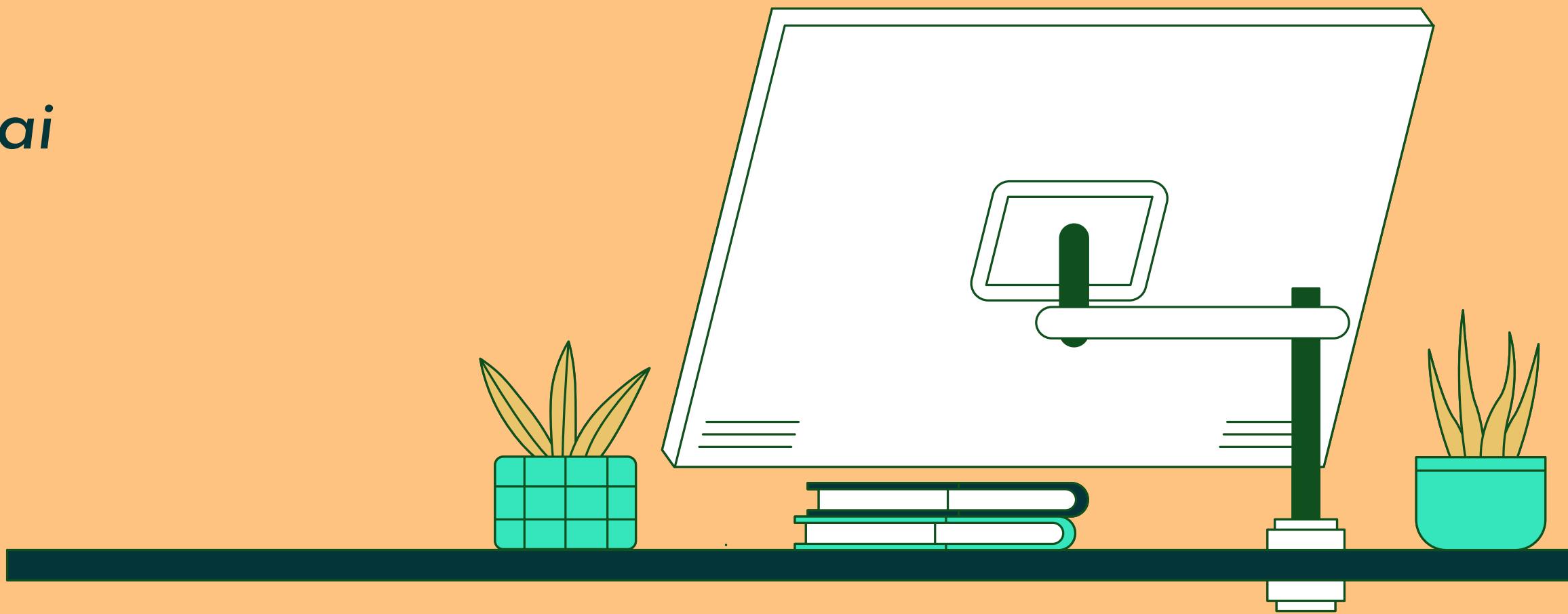


for Windows

Buka Anaconda Prompt



**Setelah installing selesai*



Perintah dasar



- **cd** : untuk masuk ke direktori
- **mkdir** : membuat direktori
- **dir/ls** : melihat isi direktori

Membuat Virtual Environment Baru

```
(base) C:\Users\ASUS>conda create -n sharing python=3.8
Collecting package metadata (current_repodata.json): done
Solving environment: done

## Package Plan ##

environment location: C:\Users\ASUS\anaconda3\envs\sharing

added / updated specs:
- python=3.8
```

conda create -n sharing python=3.8

*p.s. nama environment 'sharing' dapat disesuaikan

-n : -name

Mengaktifkan Virtual Environment yang dibuat

```
(base) C:\Users\ASUS>conda activate sharing
```

Jika virtual environment telah aktif, maka namanya akan muncul menggantikan base.

```
(sharing) C:\Users\ASUS>
```

Membuat Direktori Baru

```
(sharing) D:\Free-Sharing>mkdir Sharing
```

*Membuat direktori baru dapat dilakukan di local C: maupun D: atau lainnya
tergantung dengan keinginan dan space komputer kita :)

Jika ingin di local D maka arahkan ke local D:

```
(sharing) C:\Users\ASUS>D:
```

Install git untuk Git Clone

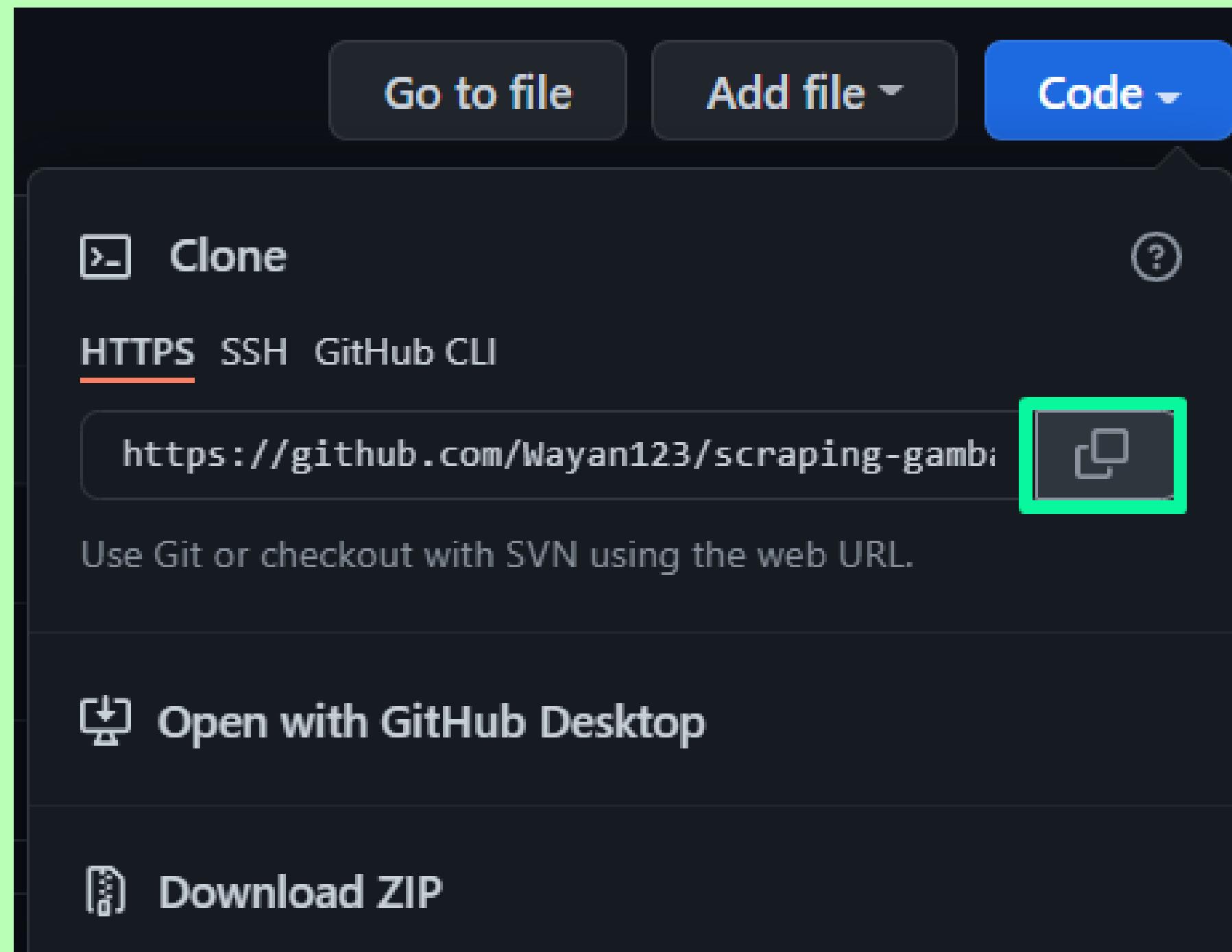
```
(sharing) D:\Free-Sharing\Sharing>conda install git
```

Tunggu hingga done, dan ketika ada pilihan [y]/n? selalu pilih y

Kemudian, lakukan git clone

```
git clone git address
```

Cara Mendapatkan Git Address



Klik *Code* dan *Copy link git address*



Git Clone

```
(sharing) D:\Free-Sharing\Sharing>git clone https://github.com/Wayan123/scraping-gambar-google.git
```

git clone https://github.com/Wayan123/scraping-gambar-google.git

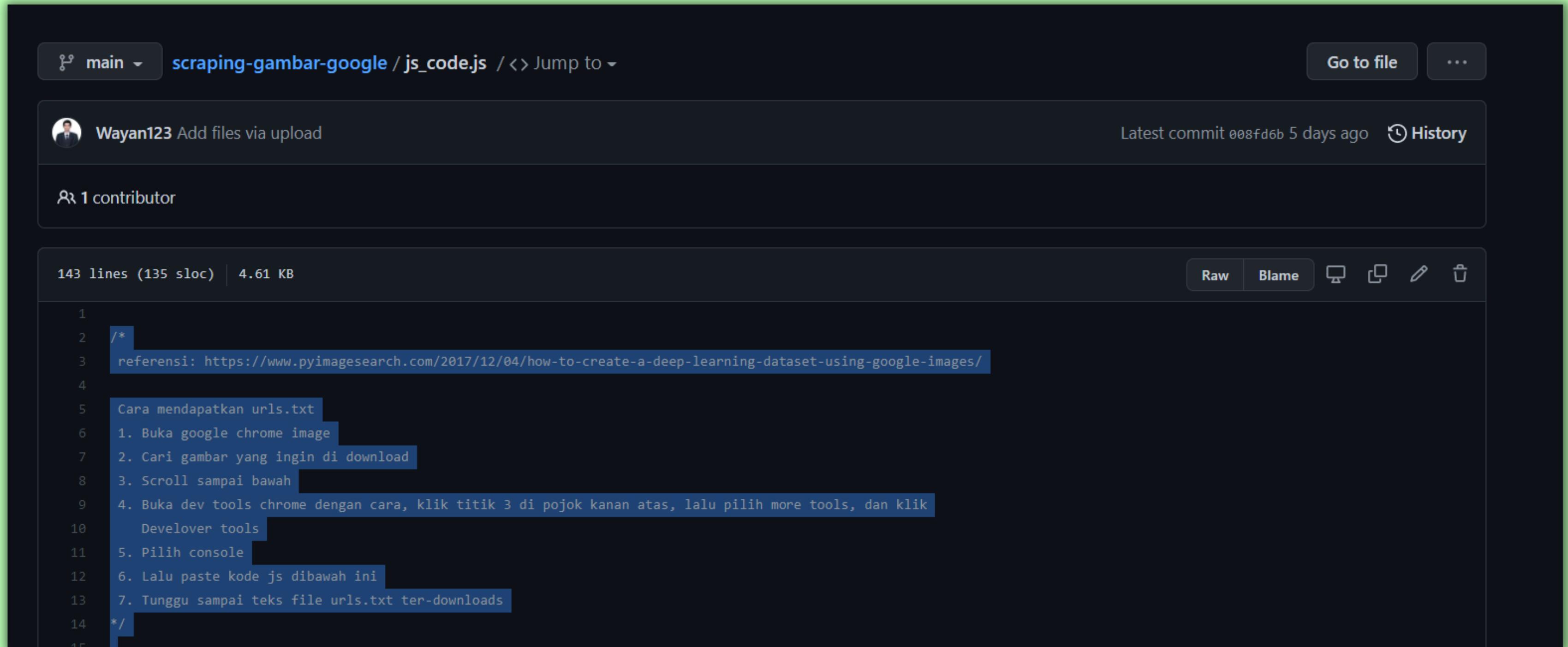
Tunggu hingga done, dan masuk ke direktori scraping-gambar-google

```
(sharing) D:\Free-Sharing\Sharing>cd scraping-gambar-google
```

Scraping Time



Copy Code js_code.js



The screenshot shows a GitHub repository page for 'scraping-gambar-google / js_code.js'. The page includes navigation links for 'main' and 'Jump to', and buttons for 'Go to file' and '...'. It displays information about the file being 143 lines (135 sloc) and 4.61 KB. The code itself is a JavaScript script for scraping Google Images, with comments explaining the steps to obtain URLs from Google Images. Lines 1 through 143 are shown, with line 1 highlighted in blue.

```
1
2 /*
3 referensi: https://www.pyimagesearch.com/2017/12/04/how-to-create-a-deep-learning-dataset-using-google-images/
4
5 Cara mendapatkan urls.txt
6 1. Buka google chrome image
7 2. Cari gambar yang ingin di download
8 3. Scroll sampai bawah
9 4. Buka dev tools chrome dengan cara, klik titik 3 di pojok kanan atas, lalu pilih more tools, dan klik
   Develop tools
10 5. Pilih console
11 6. Lalu paste kode js dibawah ini
12 7. Tunggu sampai teks file urls.txt ter-downloads
13
14 */
15
```

Dari **line 1 - 143** atau **klik Raw** dan **copy semua code**

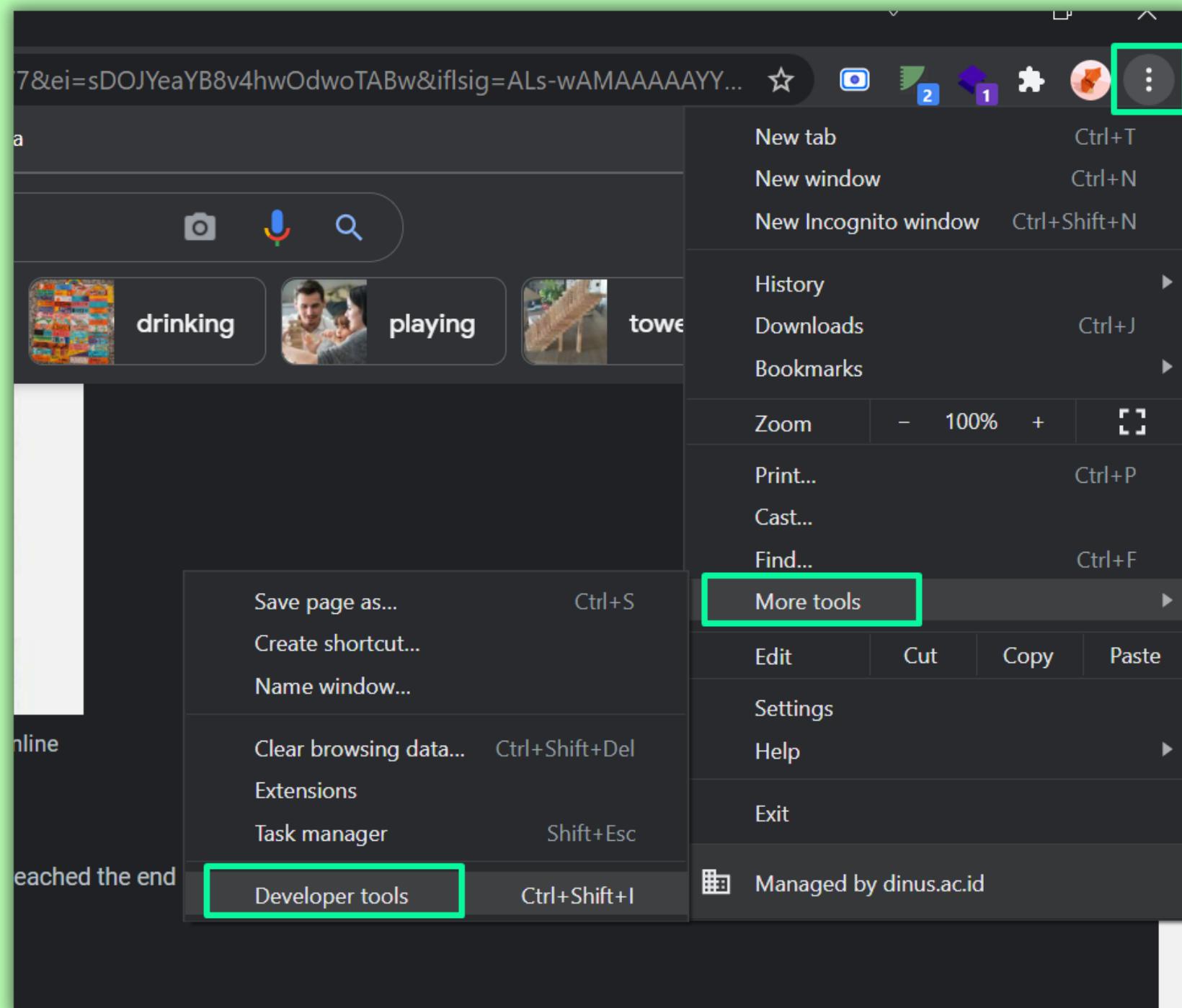
Cari Gambar di Google Image

A screenshot of a Google Images search results page for the query "jenga". The search bar at the top contains the word "jenga". Below the search bar, there are tabs for All, Images (which is selected), Shopping, Videos, Maps, More, Tools, Collections, and SafeSearch. A row of related search terms is shown below the tabs, including "game", "wedding", "giant", "falling", "blocks", "drinking", "playing", "tower", and "party". The main area displays a grid of image thumbnails. The first row contains six images: a Jenga box, a tall stack of blocks falling, a tall stack of blocks standing upright, a tall stack of numbered blocks, a tall stack of blocks, and a tall stack of blocks next to a box. The second row contains four images: a hand playing with numbered blocks, three boxes for "GO! STACKO!", a tall stack of blocks, and two boxes for "Classic Jenga".

Scroll hingga muncul kata 'Looks like you've reached the end'



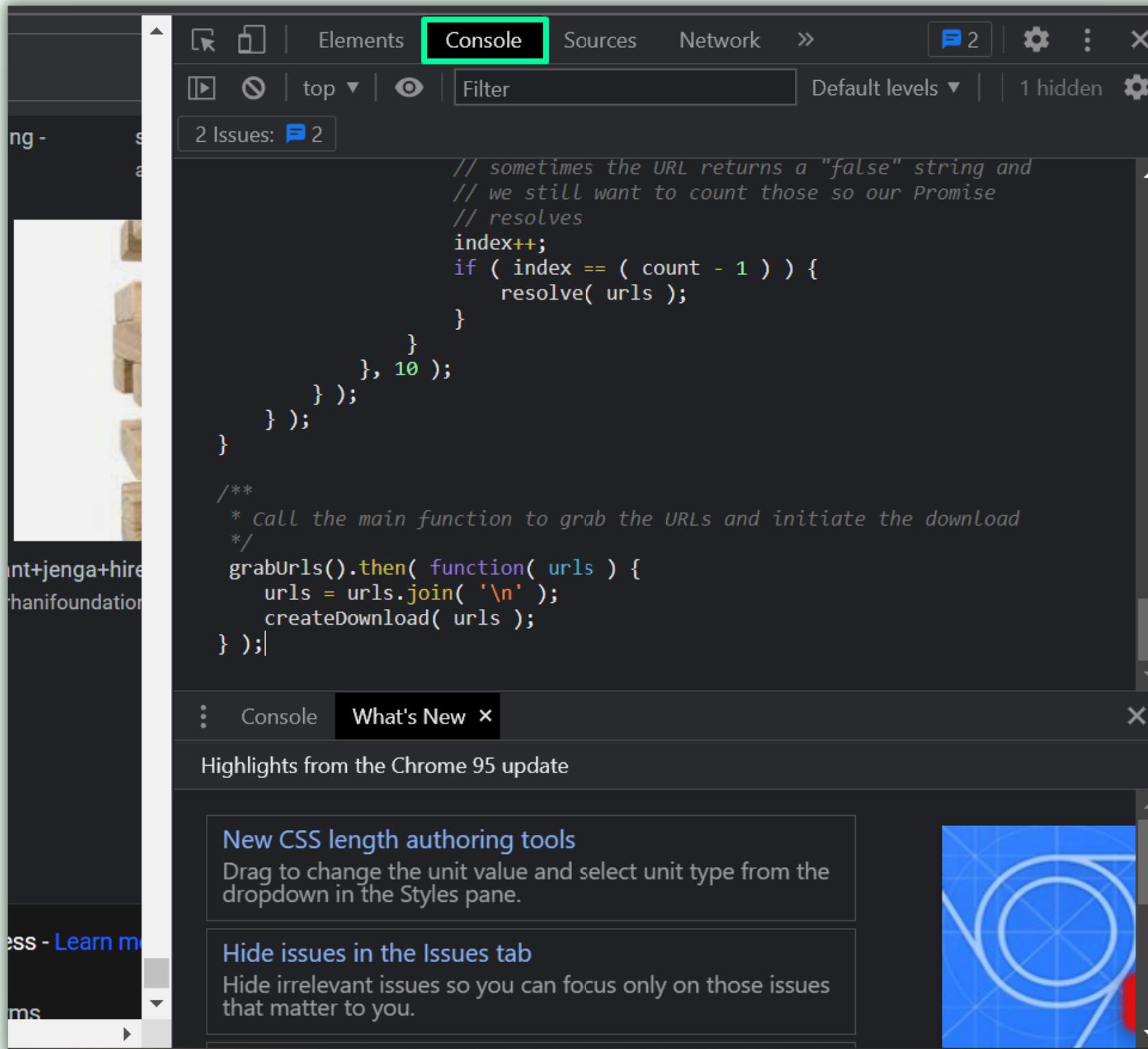
Buka Developer Tools



*atau tekan
Ctrl+Shift+I

Three dot menu - more tools - Developer tools

Paste code js_code.js pada Console



The screenshot shows the Chrome DevTools interface with the 'Console' tab selected. The console window displays the following JavaScript code:

```
// sometimes the URL returns a "false" string and
// we still want to count those so our Promise
// resolves
index++;
if ( index == ( count - 1 ) ) {
    resolve( urls );
}
}, 10 );
} );
}

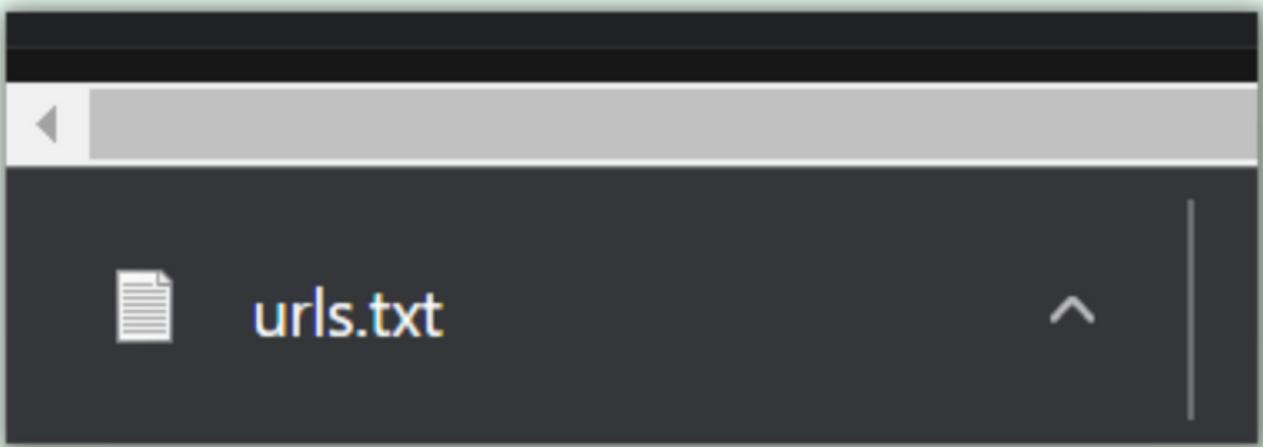
/**
 * Call the main function to grab the URLs and initiate the download
 */
grabUrls().then( function( urls ) {
    urls = urls.join( '\n' );
    createDownload( urls );
} );
```

Below the code, the console output shows:

```
: Console What's New x
Highlights from the Chrome 95 update

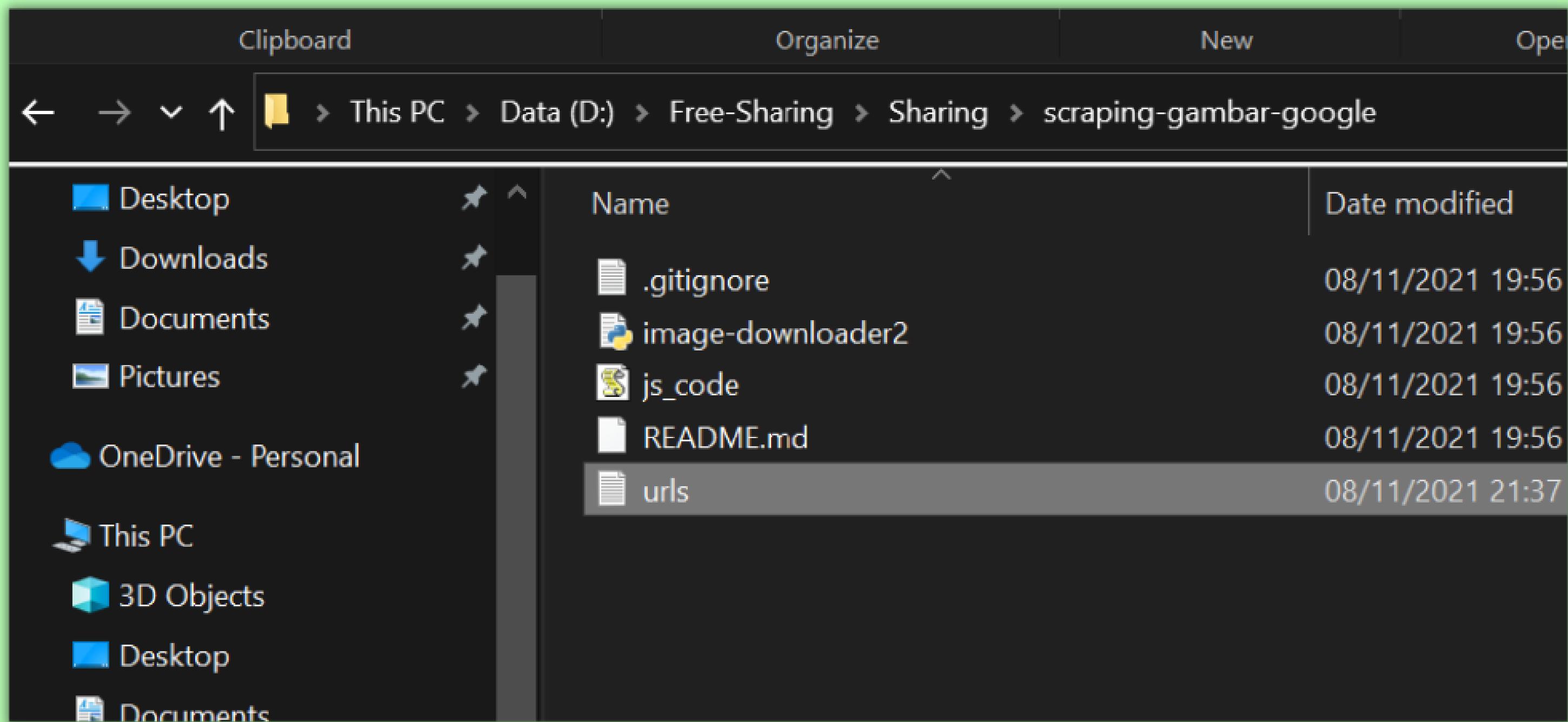
New CSS length authoring tools
Drag to change the unit value and select unit type from the dropdown in the Styles pane.

Hide issues in the Issues tab
Hide irrelevant issues so you can focus only on those issues that matter to you.
```



*Lalu tekan Enter, hingga
urls.txt terdownload*

Pindahkan urls.txt ke Direktori Scraping



*Pada folder/direktori **scraping-gambar-google**

Buat Direktori Baru untuk Gambar

```
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>mkdir gambar-jenga
```

Install opencv-python, requests, dan imutils

```
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>pip install opencv-python requests imutils
```

pip install opencv-python requests imutils

**Installing dapat ditulis bersamaan atau install masing-masing*

Cek Isi Folder

```
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>dir
Volume in drive D is Data
Volume Serial Number is 368A-473B

Directory of D:\Free-Sharing\Sharing\scraping-gambar-google

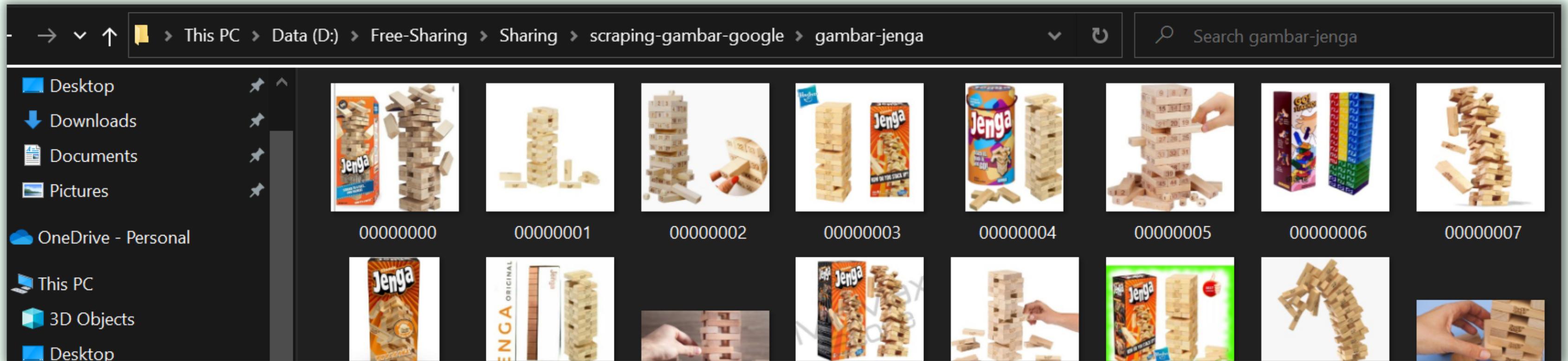
08/11/2021  21:45    <DIR>          .
08/11/2021  21:45    <DIR>          ..
08/11/2021  19:56            1.928 .gitignore
08/11/2021  21:45    <DIR>          gambar-jenga
08/11/2021  19:56            3.395 image-downloader2.py
08/11/2021  19:56            4.718 js_code.js
08/11/2021  19:56            102 README.md
08/11/2021  21:37            55.193 urls.txt
                           5 File(s)        65.336 bytes
                           3 Dir(s)  130.168.799.232 bytes free
```

*Menggunakan **dir** pada windows atau **ls** jika bisa

Download Gambar

```
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>python image-downloader2.py -u urls.txt -o gambar-jenga
```

python image-downloader2.py -u urls.txt -o gambar-jenga



***Tunggu hingga semua gambar terdownload**

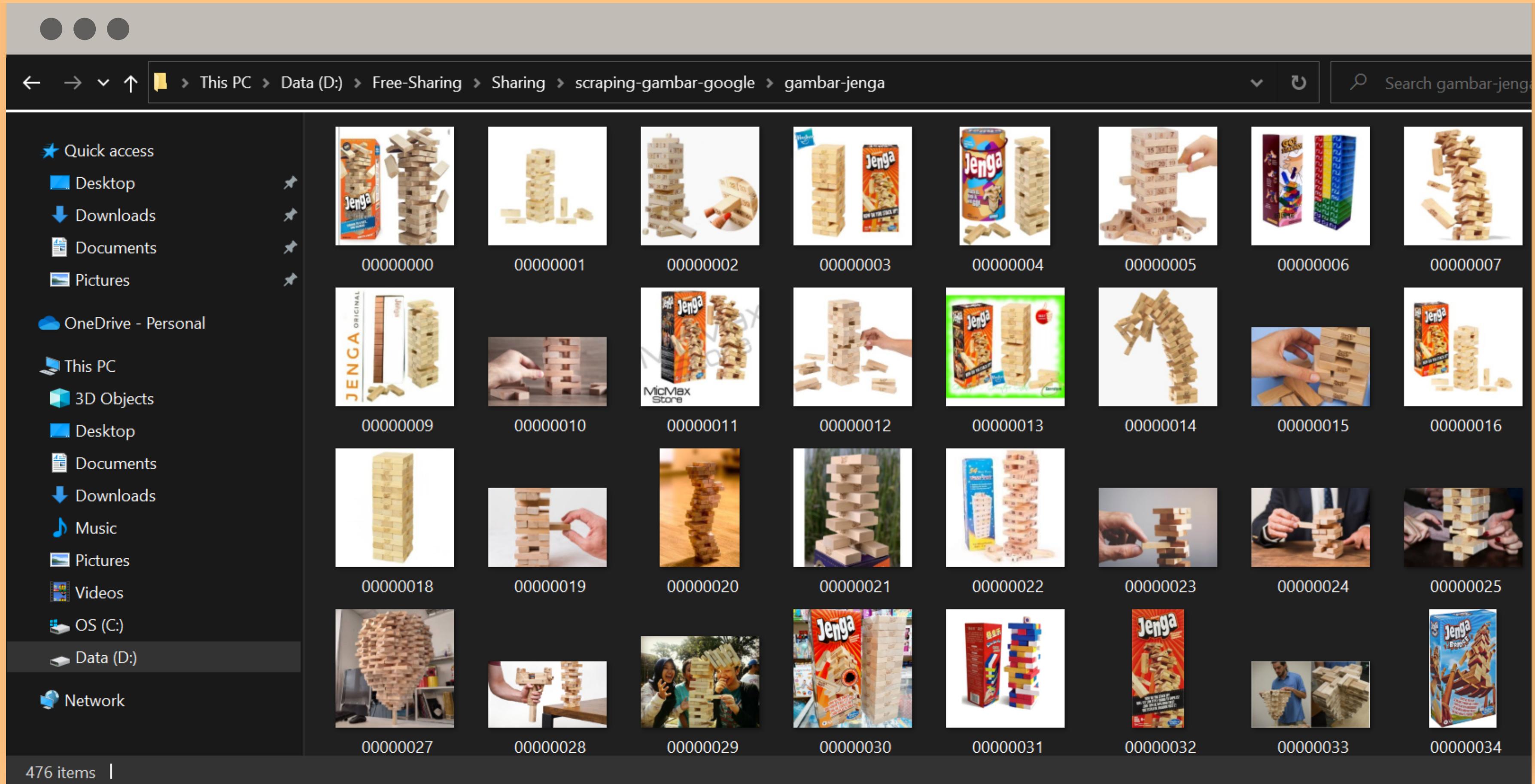


After Downloading . . .

```
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>python image-downloader2.py -u urls.txt -o gambar-jenga
image-downloader2.py:59: DeprecationWarning: The binary mode of fromstring is deprecated, as it behaves surprisingly on
unicode inputs. Use frombuffer instead
    nparr = np.fromstring(data, np.uint8)
Error downloading https://www.target.com.au/medias/static_content/product/images/full/71/41/A1327141.jpg?impolicy=produ
ct_portrait_hero : HTTPSConnectionPool(host='www.target.com.au', port=443): Read timed out. (read timeout=60)
Error downloading https://cdn-gamesworldau.pressidium.com/wp-content/uploads/2018/02/jenga-2.jpg : HTTPSConnectionPool(
host='cdn-gamesworldau.pressidium.com', port=443): Read timed out.
Error downloading https://www.multiplication.com/sites/default/files/images/jenga-box.jpg : HTTPSConnectionPool(host='w
ww.multiplication.com', port=443): Max retries exceeded with url: /sites/default/files/images/jenga-box.jpg (Caused by
NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x0000019EA9EA3D60>: Failed to establish a new connec
tion: [WinError 10060] A connection attempt failed because the connected party did not properly respond after a period
of time, or established connection failed because connected host has failed to respond'))
Error downloading https://www.multiplication.com/sites/default/files/images/jenga-write.jpg : HTTPSConnectionPool(host=
'www.multiplication.com', port=443): Max retries exceeded with url: /sites/default/files/images/jenga-write.jpg (Caused
by NewConnectionError('<urllib3.connection.HTTPSConnection object at 0x0000019EAA6DC820>: Failed to establish a new co
nection: [WinError 10060] A connection attempt failed because the connected party did not properly respond after a per
iod of time, or established connection failed because connected host has failed to respond'))
libpng warning: iCCP: known incorrect sRGB profile

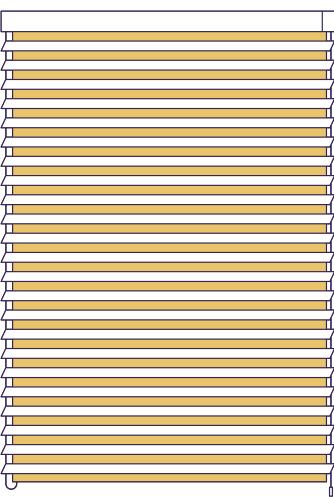
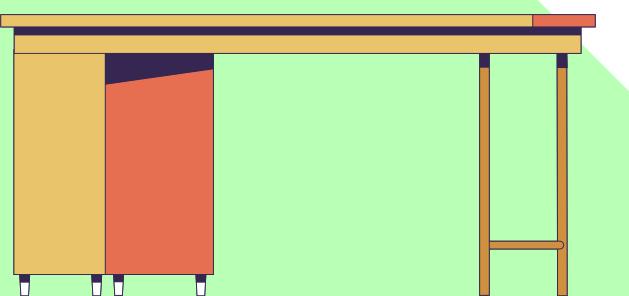
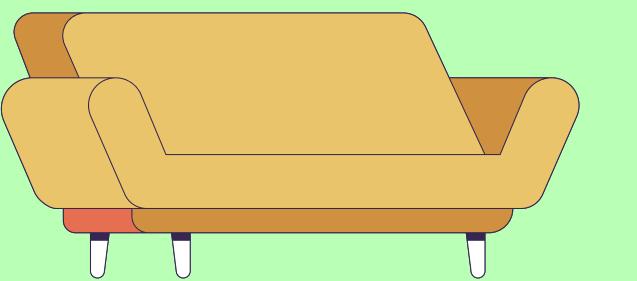
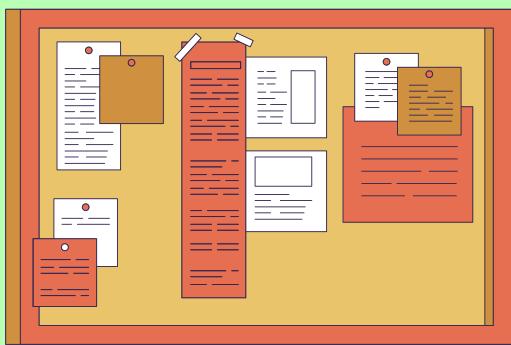
(sharing) D:\Free-Sharing\Sharing\scraping-gambar-google>
```

Processing code end



Gambar terdownload dalam folder

Thank you.



Made in
P.

