Instruction how to make HTML copy of a news article webpage.
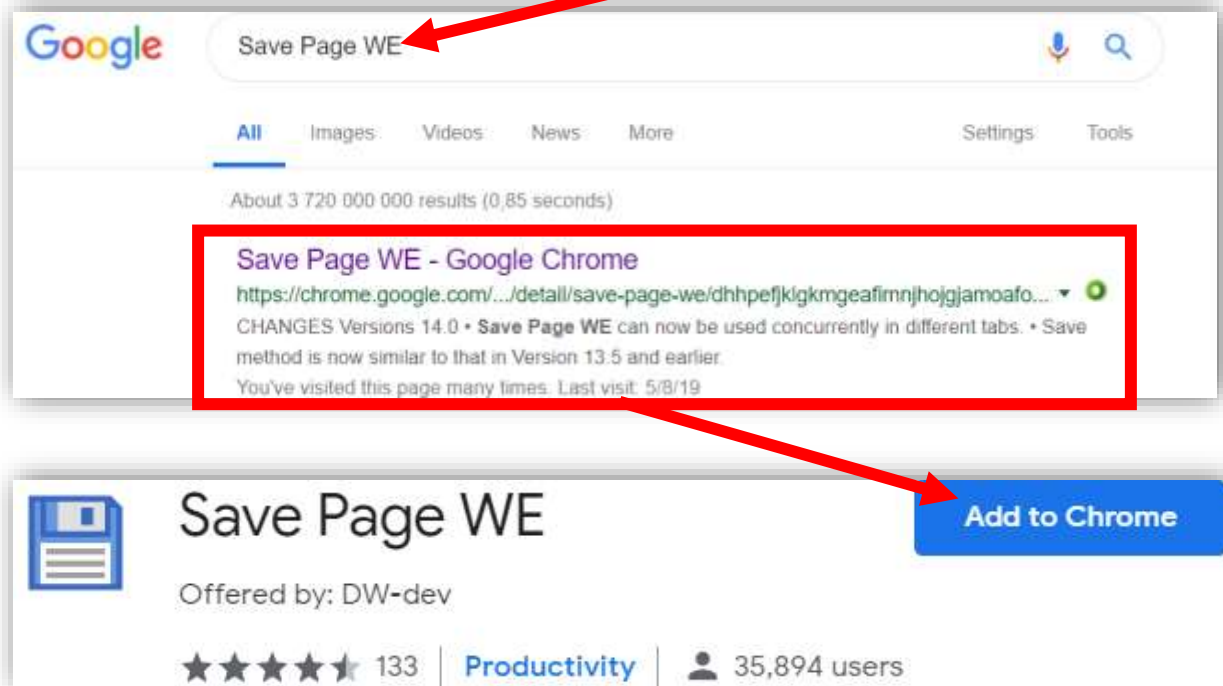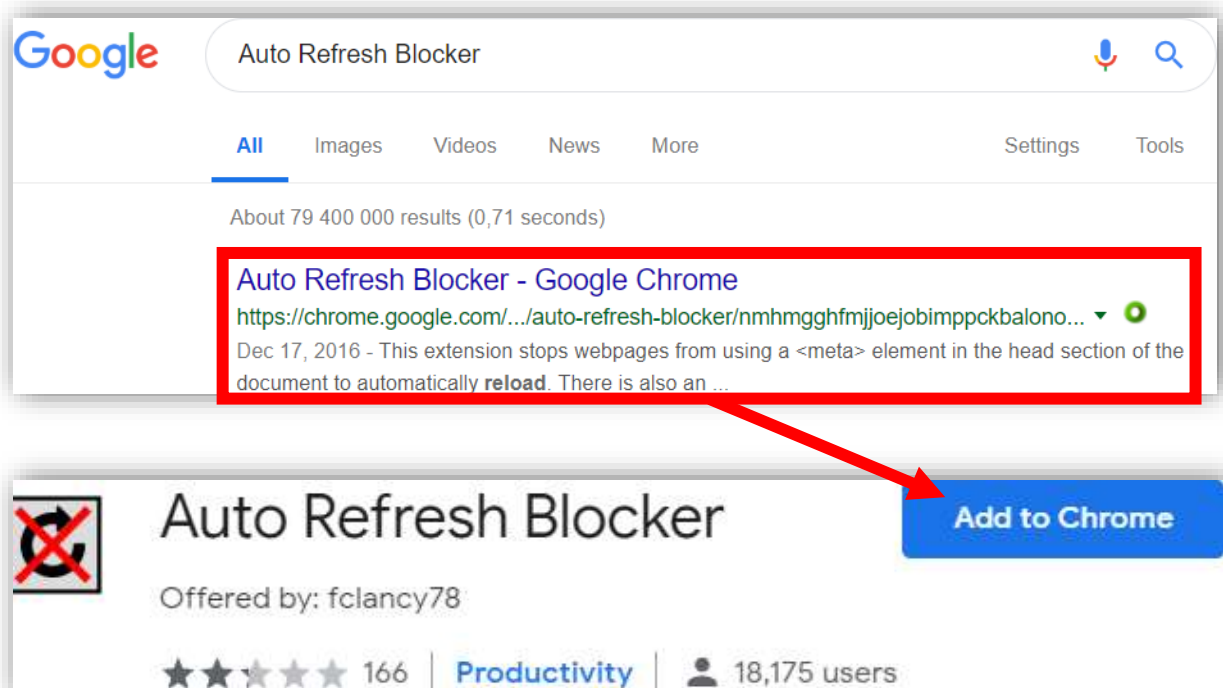
**Step 1:** Install and open Google Chrome browser.

Step 2: Download "Save Page WE" Chrome extension.

Step 3: Download "Auto Refresh Blocker" Chrome extension. Its very common that news websites are using auto refresh meta tag that will automatically refresh a webpage after 10-30 minutes. You should add this extension to disable webpage auto refreshing.

Step 4: Download "Notepad++" text editor.
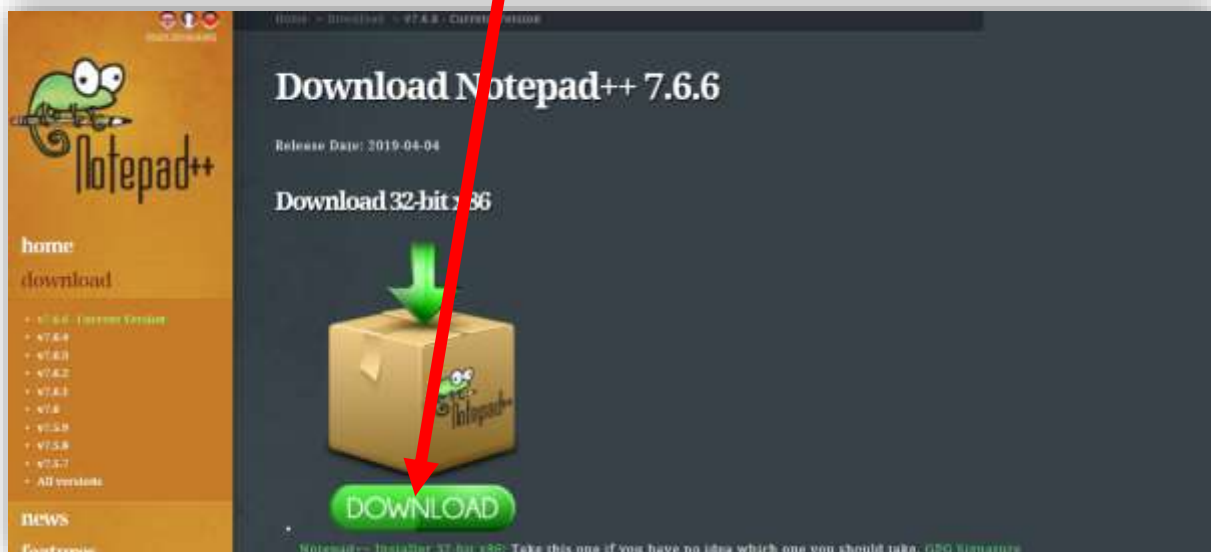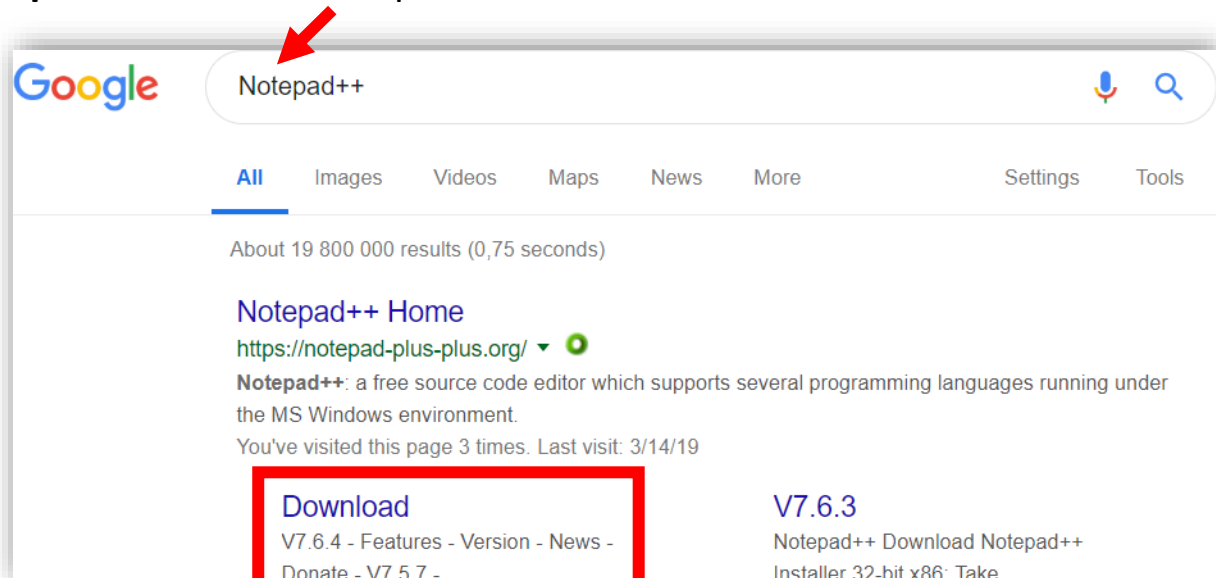
Step 5: Open a news article URL in Google Chrome browser from which you have to create a HTML copy. Just as an example, I am going to use an article from a Polish news website. Here is a link to this article:

https://podroze.onet.pl/plaze/plaza-maya-bay-pozostanie-zamknieta-przez-kolejne-dwa-lata/2bw6hh9
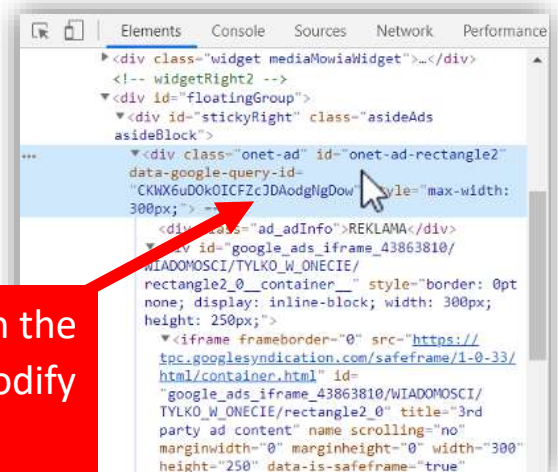
On steps 6-13 we are going to remove specific elements from a news page before we save it to our computer. Here is how to remove or modify elements using Google Chrome browser:

1 Right-click on element you have to delete or change.
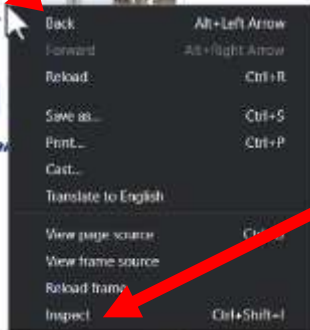
2 Click **Inspect** to open elements code.

3 Select element in the code to delete or modify

It will appear in blue colour on the page.

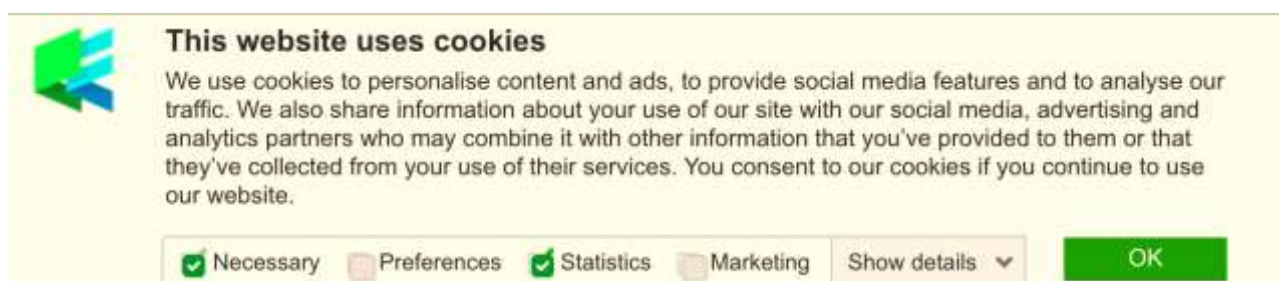**4** Click Delete button to delete selected element.

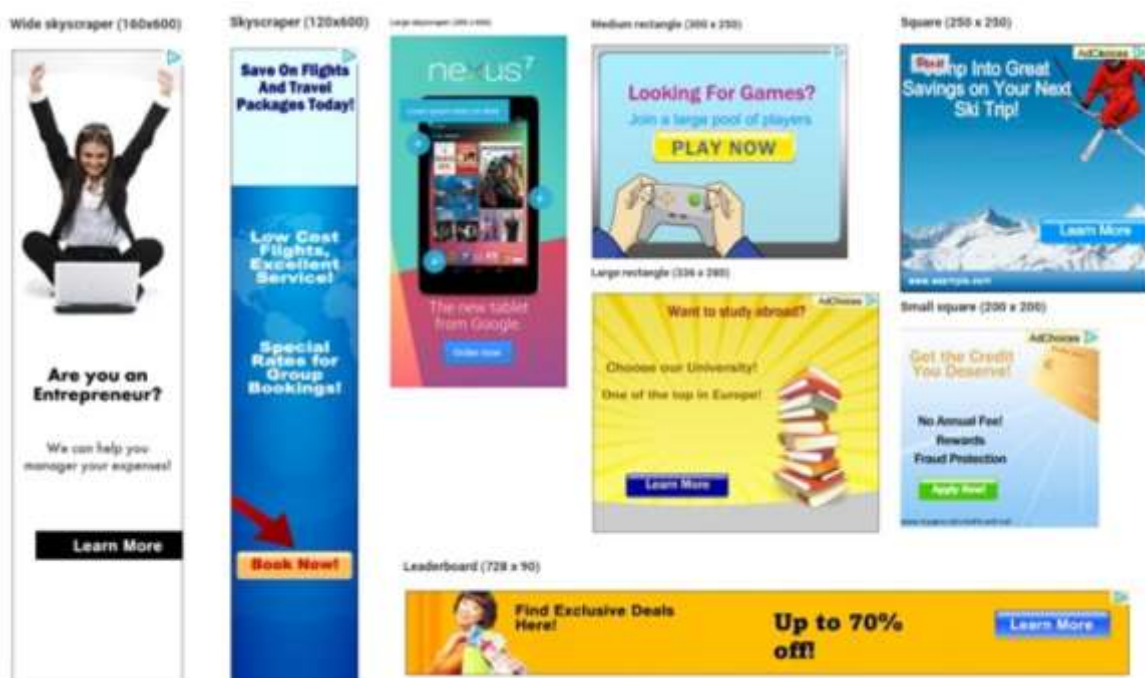**5** Use Ctrl + Z to Undo and Ctrl + Y to Redo action.

Step 6: Delete **popup windows** or any **cookie notification elements** in the page. Here is an example of **popup window**:

Here is an example of **Cookie notification:**

Step 7: Remove advertising banners from entire article. Here are some examples how advertising banners looks on the website:

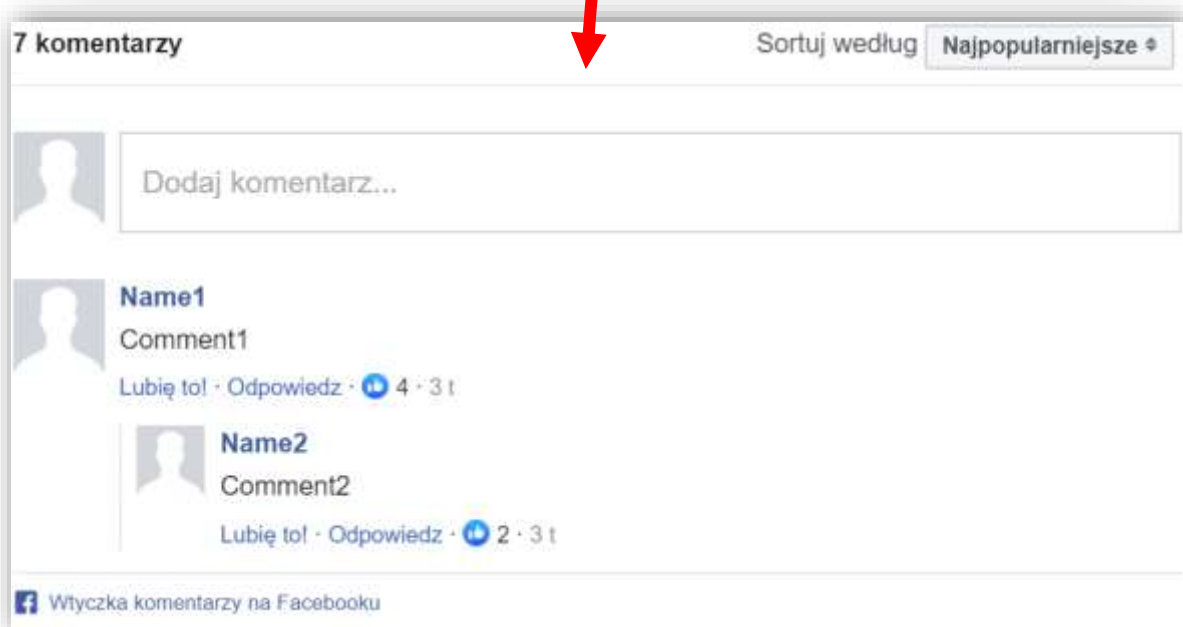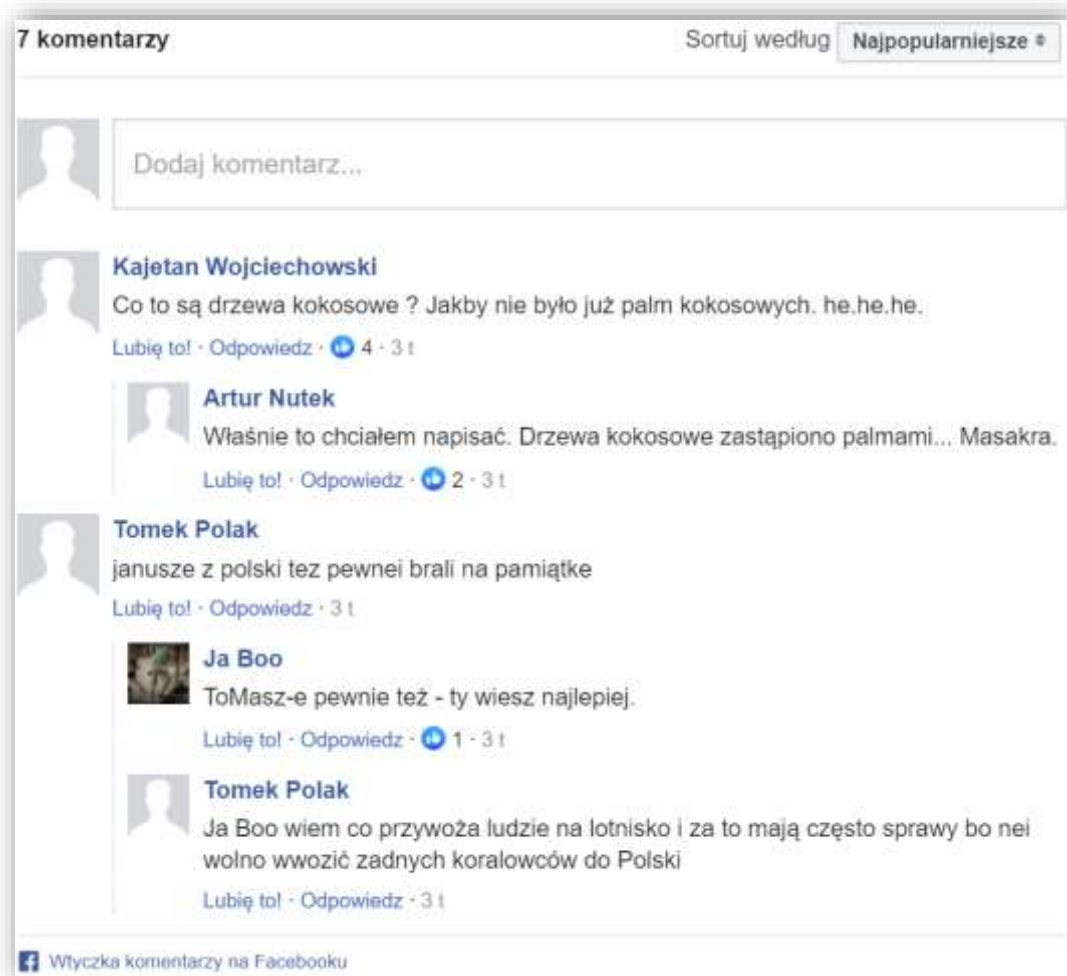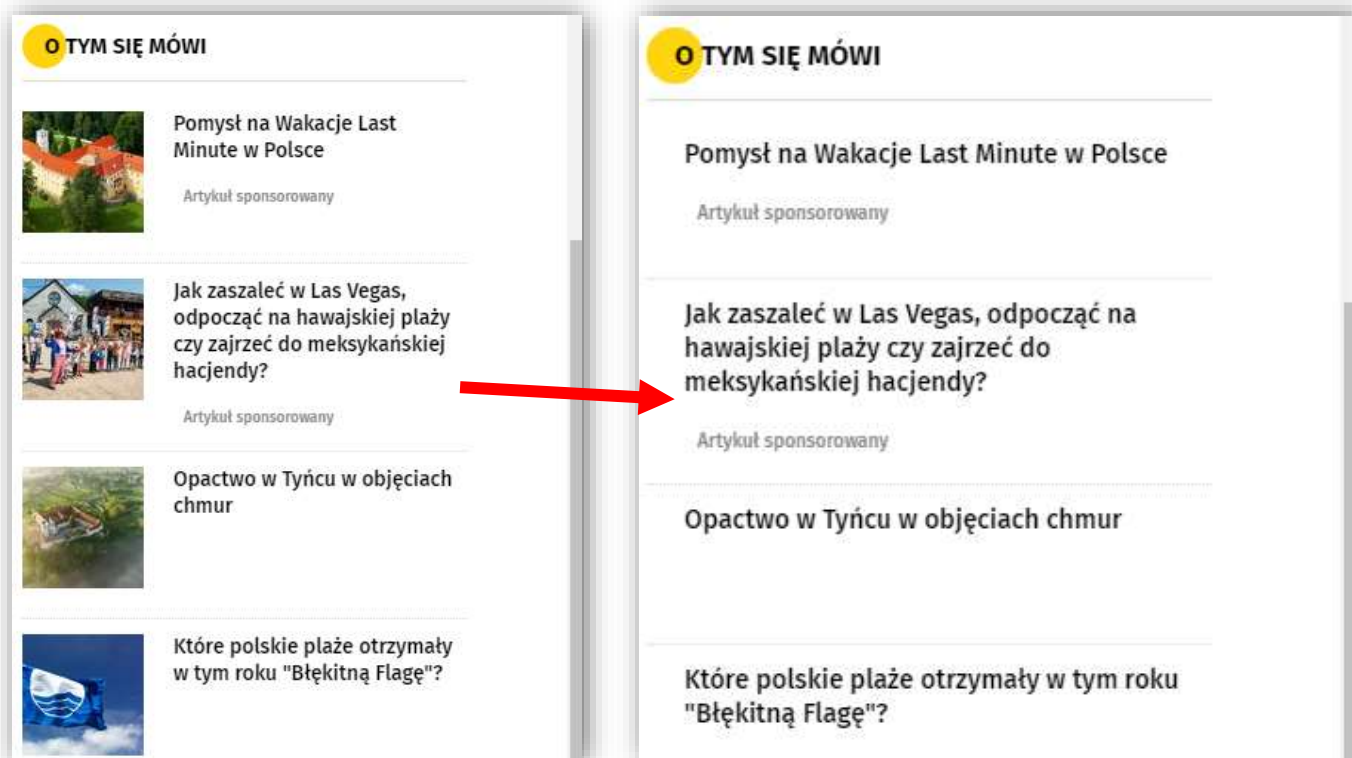**Step 8:** Remove article title and replace it with word: "Title".

Step 9: Remove all names of main image:

Step 10: Find a comment section below the article and remove almost all comments leaving just one comment with one response to it, just as shown in next before and after image examples:

Step 11: Article sidebar must contain only text elements (no images). Remove all images from the sidebar as shown in this example:

O TYM SIĘ MÓWI

Pomysł na Wakacje Last Minute w Polsce

Artykuł sponsorowany

Jak zaszaleć w Las Vegas, odpocząć na hawajskiej plaży czy zajrzeć do meksykańskiej hacjendy?

Artykuł sponsorowany

Opactwo w Tyńcu w objęciach chmur

Które polskie plaże otrzymały w tym roku "Błękitną Flagę"?

Step 12: Remove all links to other articles that you can see on the article leaving about 5 rows of the links to other articles in the bottom of the page, just as shown in the next image:

Then remove all pictures from those 5 rows of articles links as shown in next picture:

NAJPOPULARNIEJSZE DZIŚ W ONECIE

Magiera: dla niektórych będzie to ostatni występ z orłem na piersi

Pomysł na Wakacje Last Minute w Polsce

Artykuł sponsorowany

Jak zamaskować szerokie biodra? Te elementy garderoby w tym pomogą!

Elektrownia słoneczna na blokowisku wyprodukowała więcej prądu niż zakładano

Francja: Policja oskarżona o brutalność. Tysiące rannych, 24 osoby straciły oko

Sensacyjne zwycięstwo Ruiza z Joshuą. Największa zaskoczenie XXI wieku

"Big Brother": Program opuściła Izabela Mączka. Różnica głosów była minimalna

Wedel miał firmowy samolot przed II wojną światową. Zrzucał z nieba czekoladę

Europosłowie PO nie podali ręki politykom PiS. Skrytykowali ich partyjni koledzy

Jak zostać milionerem? To pierwszy nawyk bogaczy

Ceny wszystkiego rosną powoli. Prawie tego nie widać. Chyba, że chodzi o ziemniaki
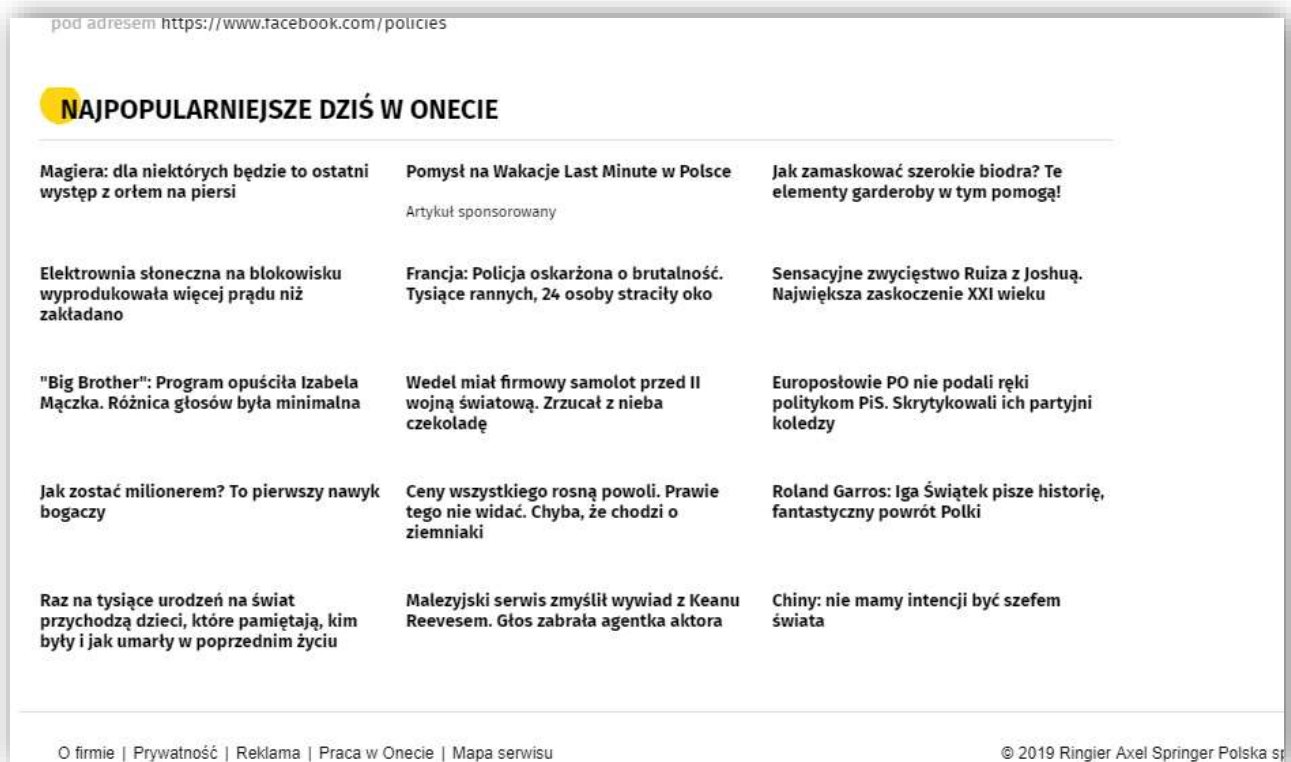
Roland Garros: Iga Świątek pisze historię, fantastyczny powrót Polki

Raz na tysiące urodzeń na świat przychodzą dzieci, które pamiętają, kim były i jak umarły w poprzednim życiu

Malezyjski serwis zmyślił wywiad z Keanu Reevesem. Głos zabrała agentka aktora

Chiny: nie mamy intencji być szefem świata

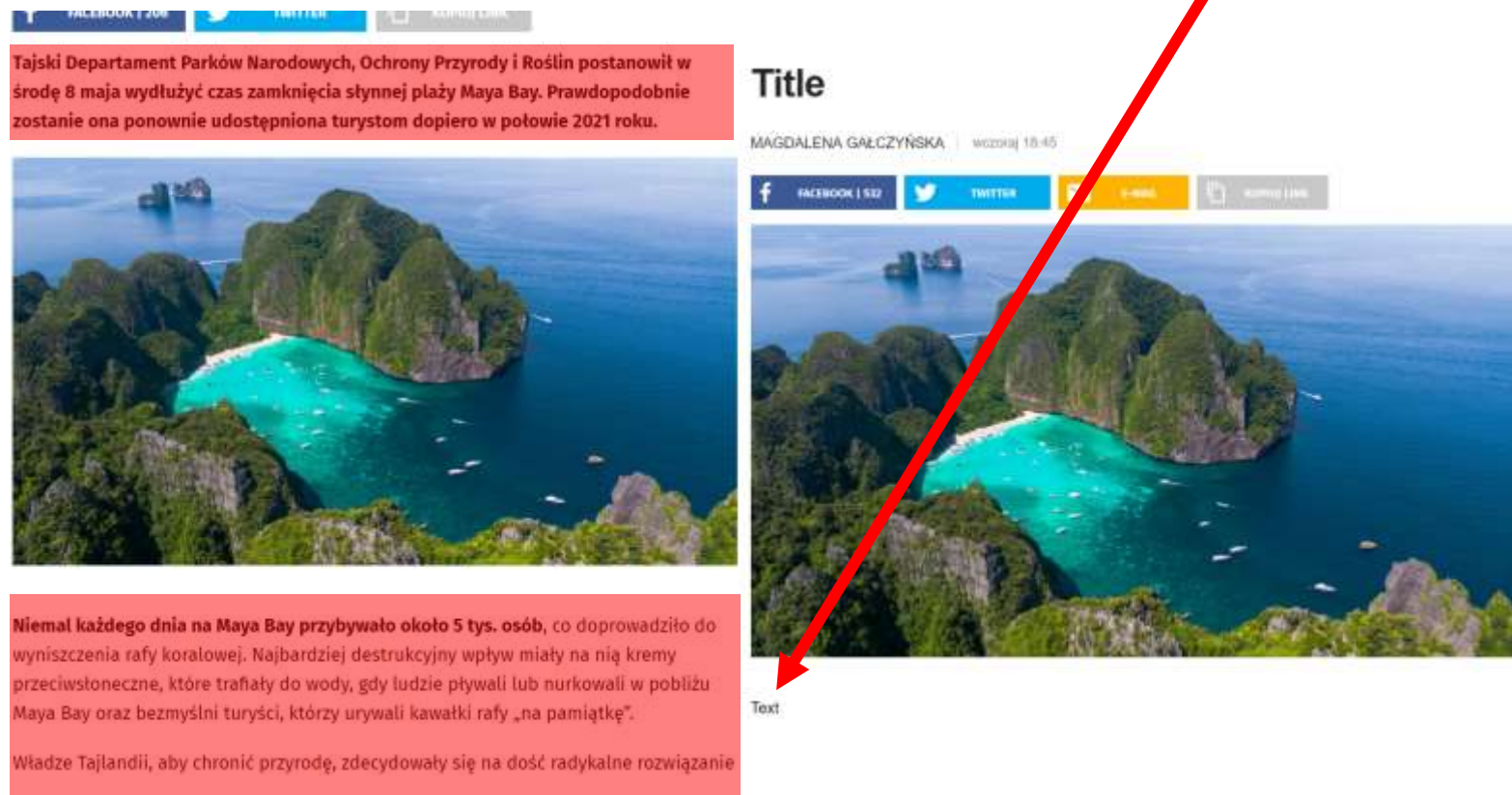O firmie | Prywatność | Reklama | Praca w Onecie | Mapa serwisu

© 2019 Ringier Axel Springer Polska sp

Step 13: Remove article content and replace it with a word "Text", it should be located right after article main image.

Tajski Departament Parków Narodowych, Ochrony Przyrody i Roślin postanowił w środę 8 maja wydłużyć czas zamknięcia słynnej plaży Maya Bay. Prawdopodobnie zostanie ona ponownie udostępniona turystom dopiero w połowie 2021 roku.

Niemal każdego dnia na Maya Bay przybywało około 5 tys. osób, co doprowadziło do wyniszczenia rafy koralowej. Najbardziej destrukcyjny wpływ miały na nią kremy przeciwsłoneczne, które trafiały do wody, gdy ludzie pływali lub nurkowali w pobliżu Maya Bay oraz bezmyślni turyści, którzy urywali kawałki rafy „na pamiątkę".

Władze Tajlandii, aby chronić przyrodę, zdecydowały się na dość radykalne rozwiązanie

Title

MAGDALENA GAŁCZYŃSKA   wczoraj 18:45

FACEBOOK | 532   TWITTER   E-MAIL   KOPIUJ LINK

Text

When removing any parts of the page, be careful and make sure you don't delete next elements:

Don't delete social media buttons

FACEBOOK | 205   TWITTER   KOPIUJ LINK   7   SKOMENTUJ

Don't delete Details who wrote the article

Źródło: Bangkok Post

Don't delete Article publication dates, they are found in the beginning and the end of the article.

9 maja, 16:01

Data utworzenia: 9 maja 2019 16:01

f

7

‹ wróć

This is how page will look before saving it with "Save Page WE":

Step 14: Now we are ready to save a webpage as HTML file. Before saving scroll webpage above the fold in Chrome browser in order to save a page with the whole navigation menu.
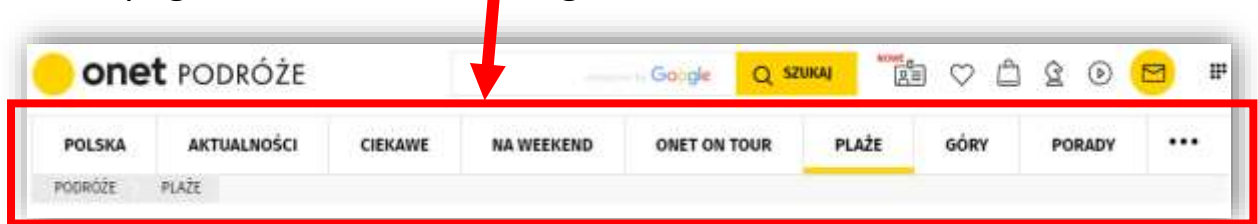
Then right click "Save Page WE" icon and select Save Basic option.
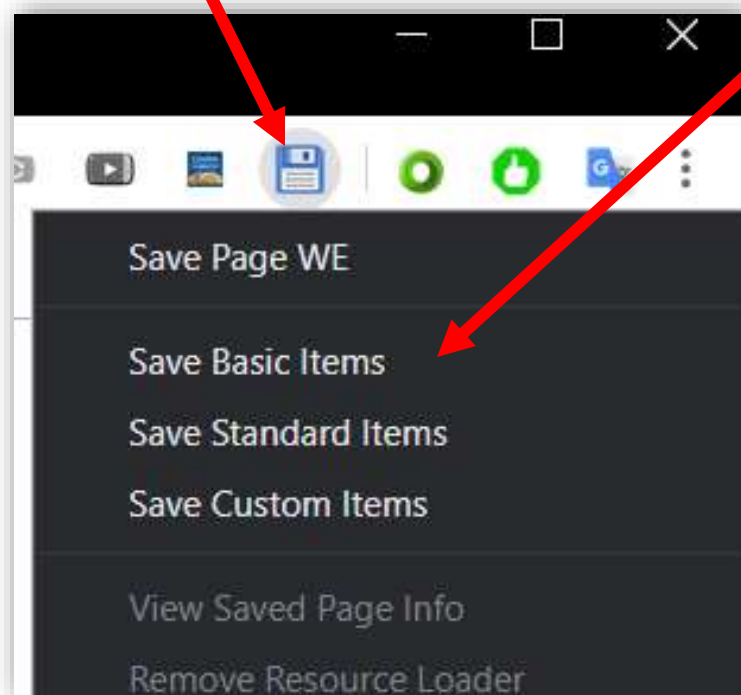
Step 15: Rename downloaded HTML file to index.
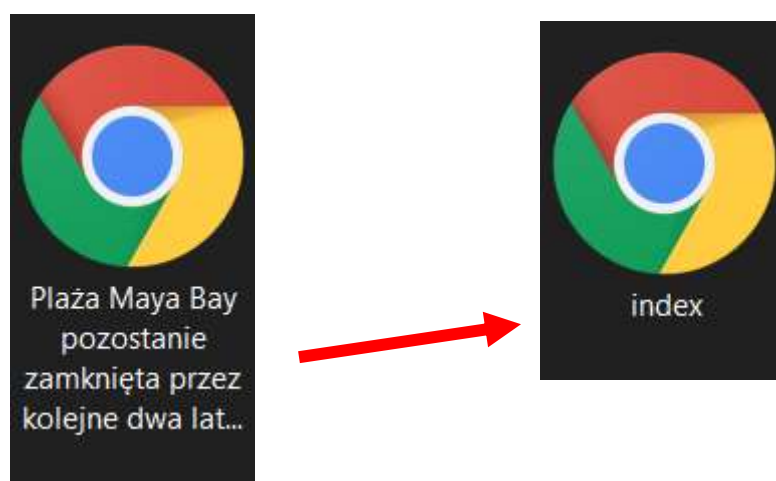
Step 16: Open HTML file in Chrome browser and check if index.html looks visually identical to original webpage you made a copy from.

Sometimes elements may be missing, like icons, logo or comment section. **If something is missing, you have to restore those elements.**

For example, if comment section is not visible, highlight it from original page and copy-

paste it to HTML editor like Quackit: https://www.quackit.com/html/online-html-editor
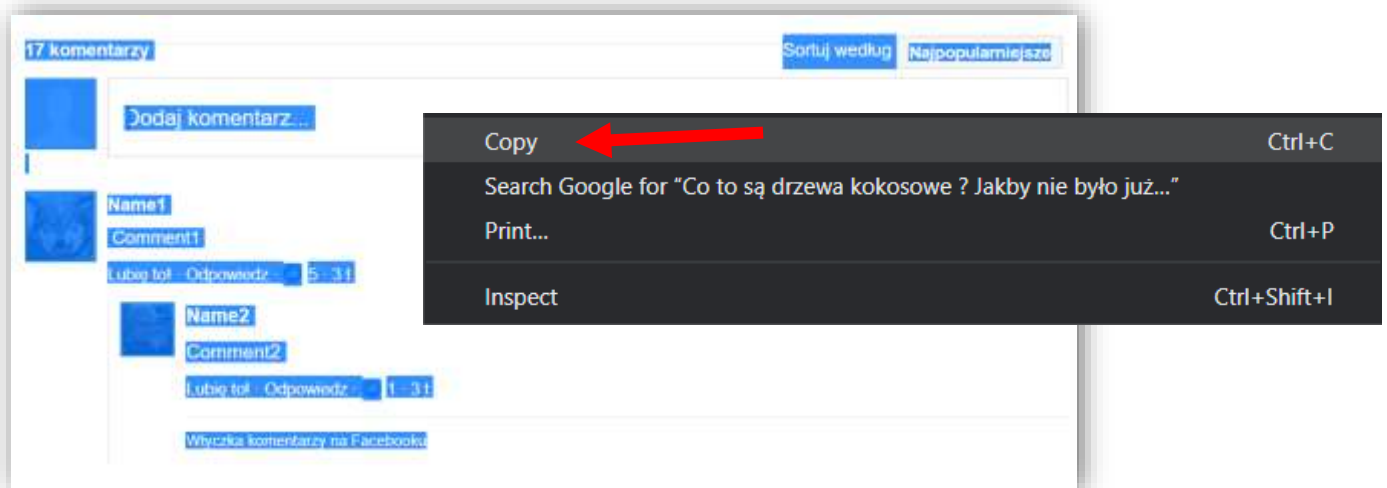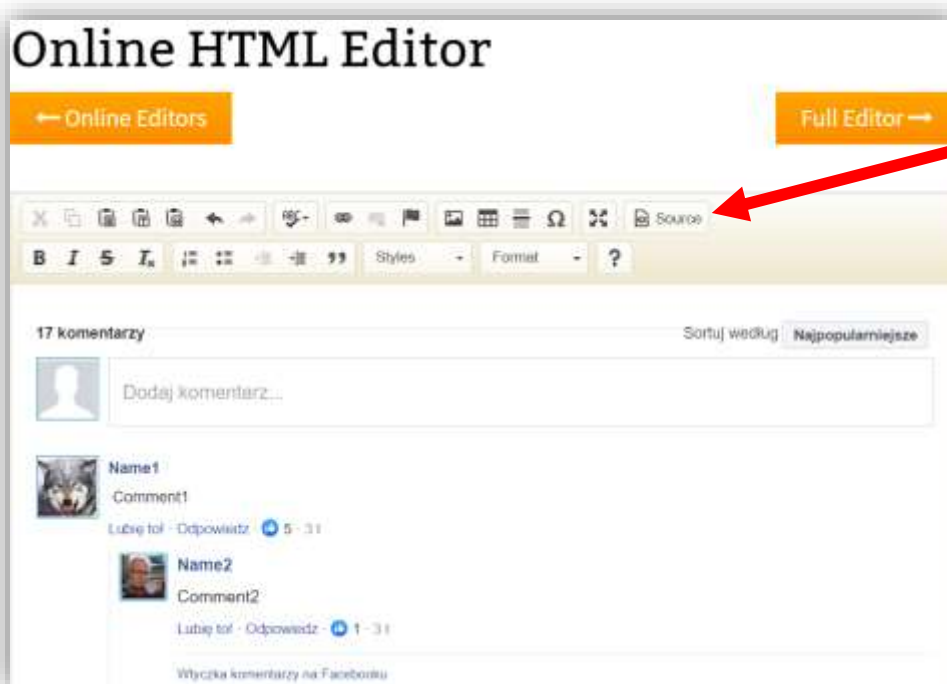
Then go to Source tab, copy HTML code and replace it with not working HTML element inside "Notepad++" editor.

Step 17: Open Notepad++ text editor.
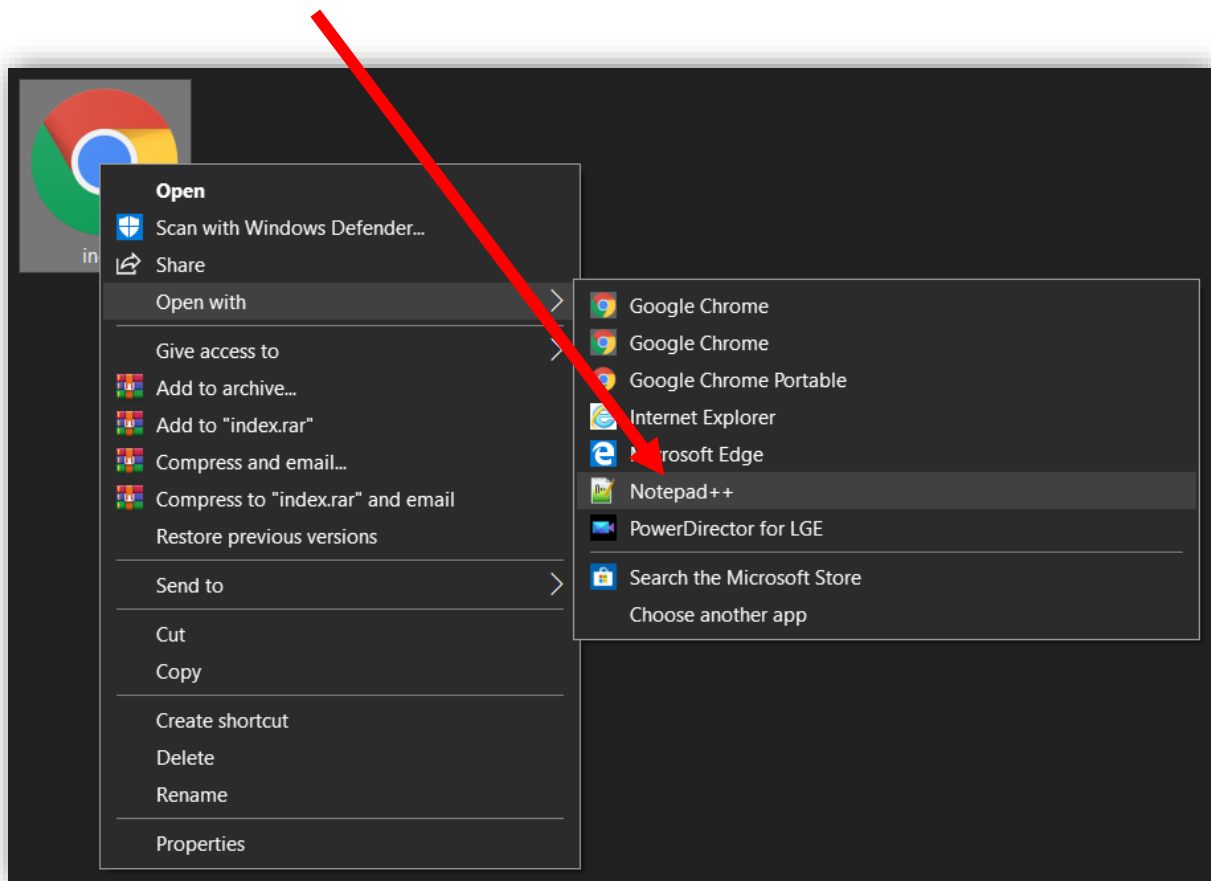
You will be executing codes that will remove or modify index.html page elements. Follow these 5 steps:

3 Fill out "Find what"

2 Click "Replace"

1 Highlight all text: Ctrl+A

4 Fill out "Replace with"

5 Click "Replace all"

You can check here how much lines were replaced.
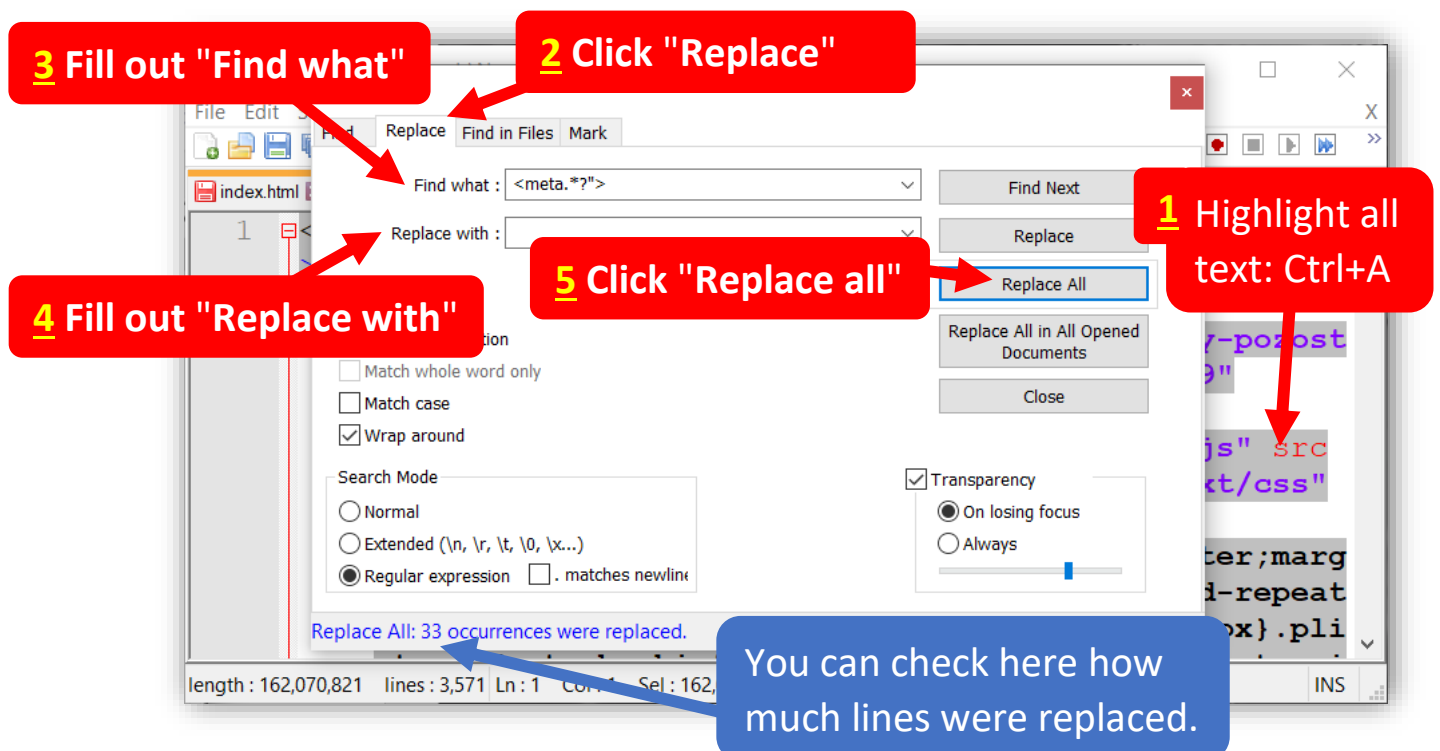
Fill out "Find what" and "Replace with" fields with these codes and then click "Replace All" button. After each execution, save and reload HTML page in Chrome browser to recheck if page has not changed visually from the original.

After each execution check original and changed pages carefully side by side, because sometimes a replacement can remove some parts of the code that are responsible for some webpage elements. **If some part of the page has changed visually in the browser, you have to find what part of the code causes this and fix it.**

Codes to execute:

1. Find what: **<meta.*?">**
 Replace with:

2. Find what: **<base.*?">**
 Replace with:

3. Find what: **<link.*?">**
 Replace with:

4. Find what: **<iframe.*?</iframe>**
 Replace with:

5. Find what: **<!--.*?-->**
 Replace with:

6. Find what: **<script.*?</script>**
 Replace with:

7. Find what: **(href=".*?")**
 Replace with: **href="javascript:void(0);"**

Step 18: Cut all CSS from HTML and paste it to separate **style.css** file. Selecting styles one by one from HTML is very long process, just use next three codes instead:

1. Remove everything in HTML file besides CSS style:
Find what: **(?s)^.*?(<style.*?</style>)|(?s).*\z**
Replace with: **(?1\1\r\n)**

2. After you paste style to style.css file, delete all style tags using:
Find what: **<style.*?>**
Replace with:

Find what: **</style.*?>**
Replace with:

3. Then remove all CSS from HTML file:
Find what: **<style.*?</style>**
Replace with:

Step 19: Add next line in the beginning of <head> tag:
<link href="style.css" rel="stylesheet">

Step 20: Remove all empty lines from HTML:
Find what: **^[\t]*$\r?\n**
Replace with:

Step 21: Remove unnecessary HTML to reduce size of **index.html** file as much as possible.

Open **index.html** in Chrome browser and Notepad ++ side by side and remove all HTML code elements that does not appear visually in a browser and does not change page visual look. Remember to save and recheck page in Chrome each time you remove HTML code.

Good example of elements to remove are **dropdown menu options**, because they are visible only after clicking or mouseover**:**

Remove all elements that does not change visual look of a HTML page:
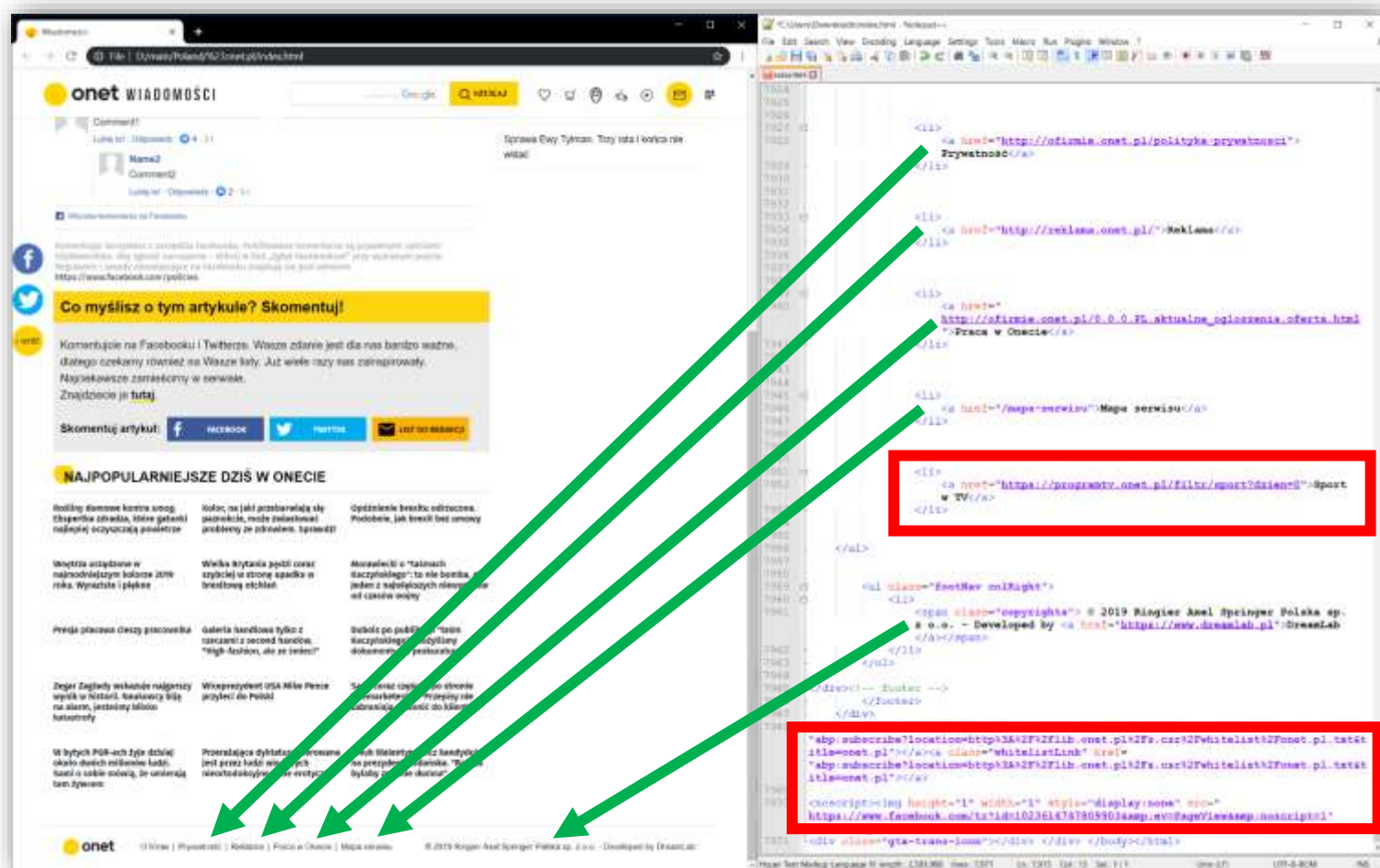
Green color arrows show that we keep elements which are found on the article and are responsible for page visual look.

Remove elements that are highlighted in red, because they does not change page visual appearance after they are removed.

Do not skip this step. In the end HTML and CSS files should weigh together without images no more than 300 kilobytes.

Step 22: Replace **article publishing date** to show always that article was published _yesterday_ and change a year near copyright symbol.

- Next JavaScript will display current year:

```
<script type="text/javascript"> var mydate=new
Date;mydate.setDate(mydate.getDate()),document.write(""+mydate.getFullYear()
); </script>
```
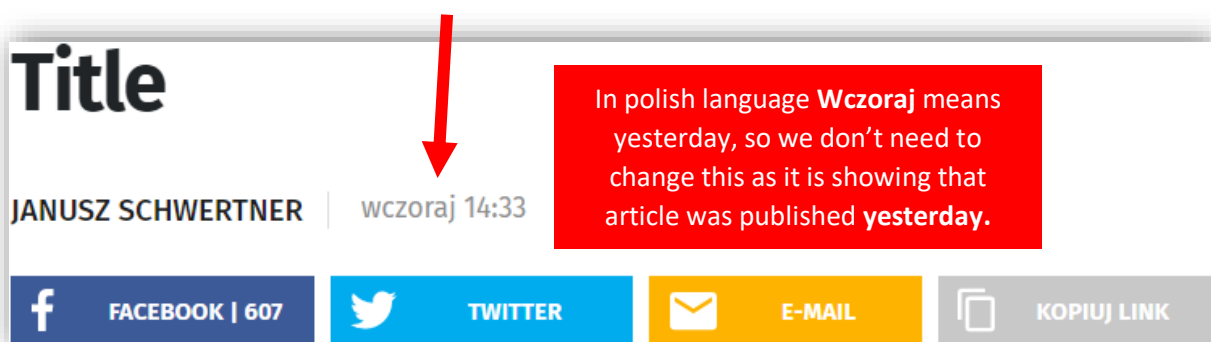
Usually a year date is visible on the bottom of a webpage, where **Copyright** symbol **©** is located:

© 2019 Ringier Axel Springer Polska sp. z o.o. - Developed by DreamLab

Then change article publishing dates. They are found before and after the article. Keep original format of the dates.

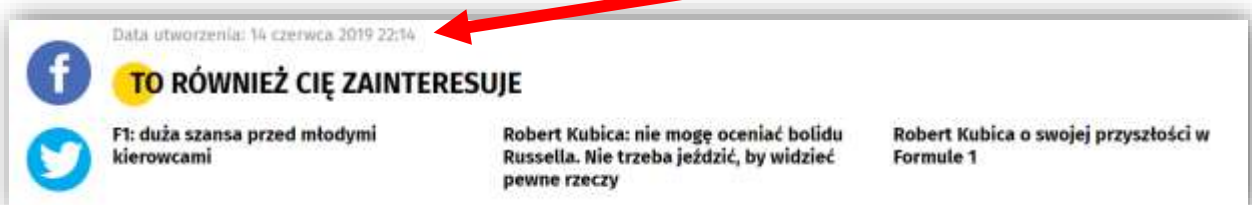- Date when article was written before the article:

In polish language **Wczoraj** means yesterday, so we don't need to change this as it is showing that article was published **yesterday.**

- Date when article was written after the article:

When we have to modify date that has day, month and year format, replace it with JavaScript below so it would always show that article was published yesterday:

<script type="text/javascript">

```
        var month=new
Array;month[0]="1",month[1]="2",month[2]="3",month[3]="4",month[4]="5",
month[5]="6",month[6]="7",month[7]="8",month[8]="9",month[9]="10",mont
h[10]="11",month[11]="12";var mydate=new
Date;mydate.setDate(mydate.getDate()),document.write(""+mydate.getDate()-
1+"."+month[mydate.getMonth()]+"."+mydate.getFullYear());
        </script>
```

In order to preserve original format, we change yellow numbers to Polish months and dots to whitespaces:

```
<script type="text/javascript">
        var month=new
Array;month[0]="Stycznia",month[1]="Lutego",month[2]="Marca",month[3]="
Kwietnia",month[4]="Maja",month[5]="Czerwca",month[6]="Lipca",month[7]=
"Sierpnia",month[8]="Września",month[9]="Października",month[10]="Listopa
da",month[11]="Grudnia";var mydate=new
Date;mydate.setDate(mydate.getDate()),document.write(""+mydate.getDate()-
1+" "+month[mydate.getMonth()]+" "+mydate.getFullYear());
        </script>
```
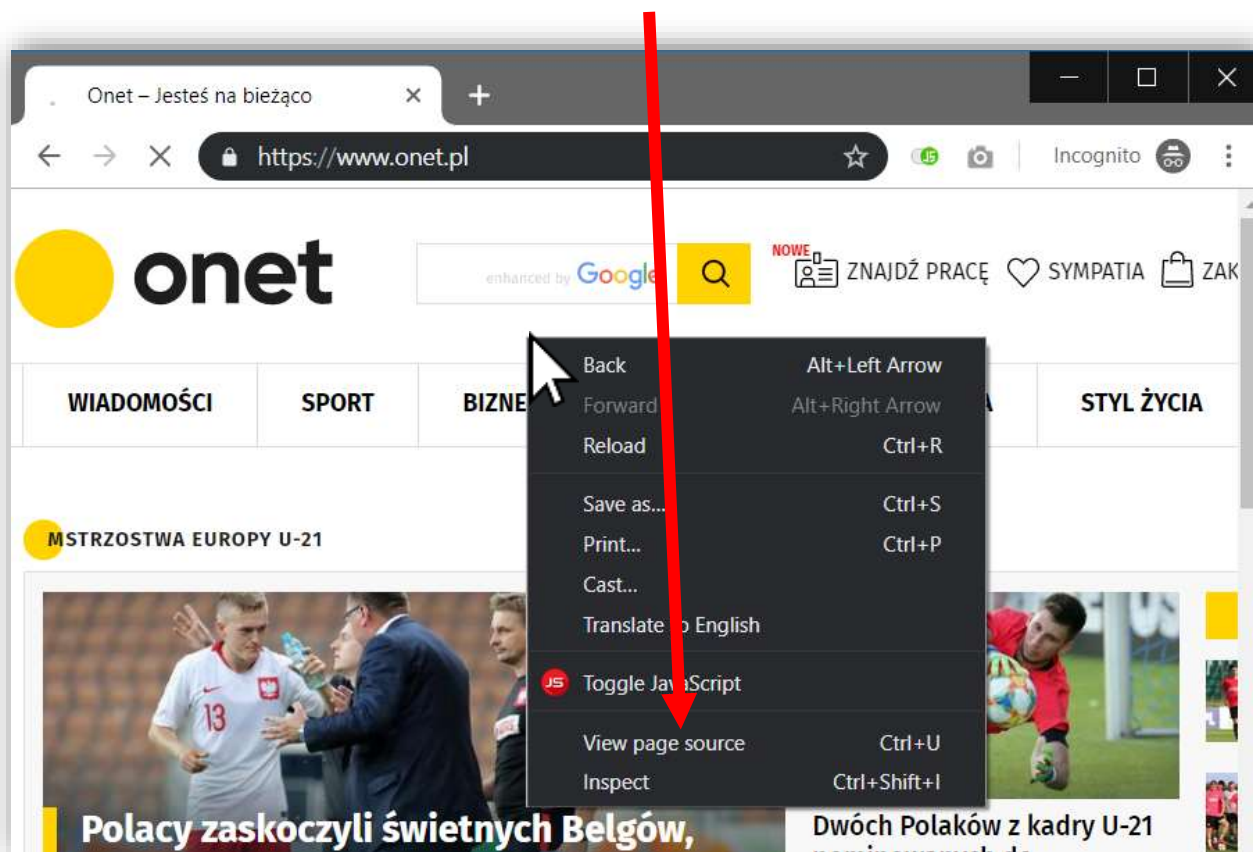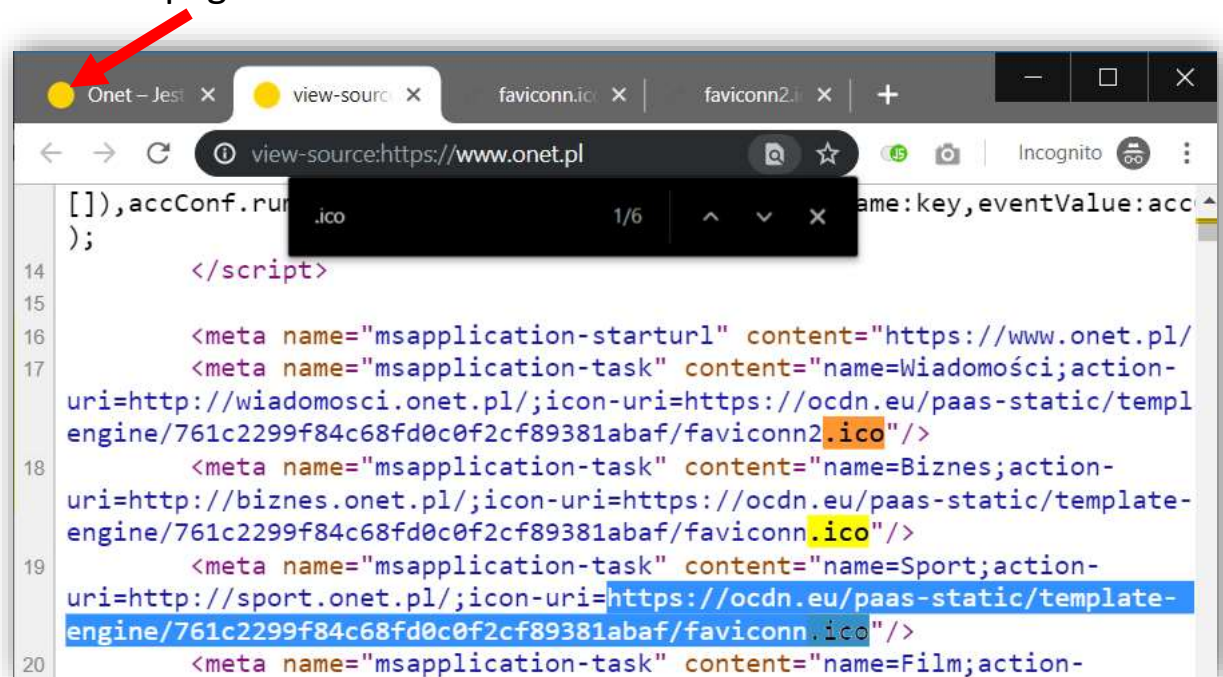
**Step 23:** Replace article main image to a.jpg image:
**<img width="100%" height="100%" src="a.jpg">**
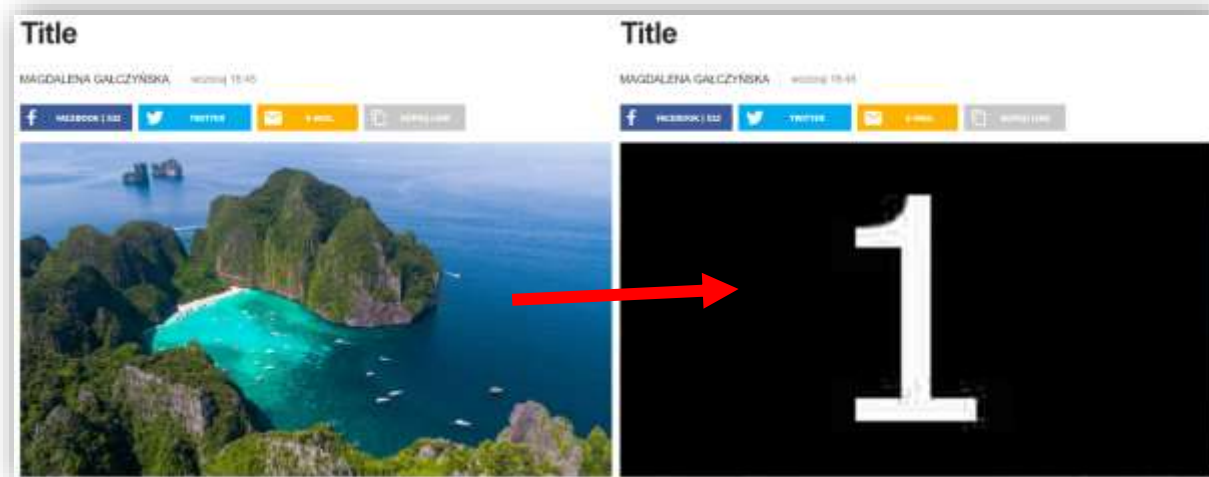


**Step 24:** Add these two lines inside the <head> tag:
**<link rel="icon" href="favicon.ico" type="image/x-icon"/>**
**<link rel="shortcut icon" href="favicon.ico" type="image/x-icon"/>**

Go to Home page **(domain.com)** of a news website you are copying and here right click and open Page Source:
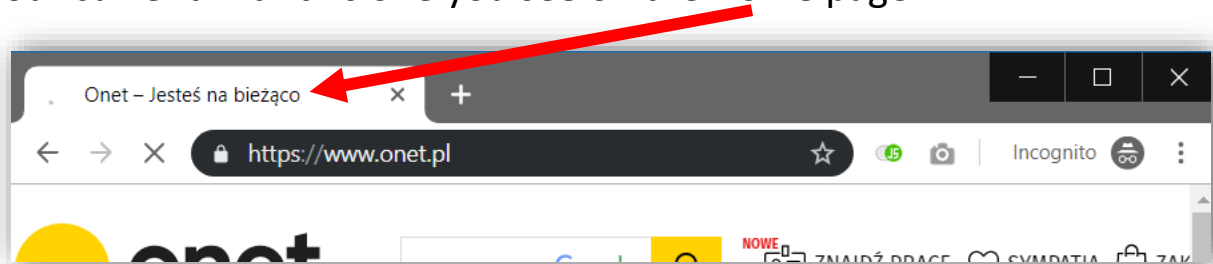
Inside Page Source search **.ico** to find favicon picture you see near page title of the page:
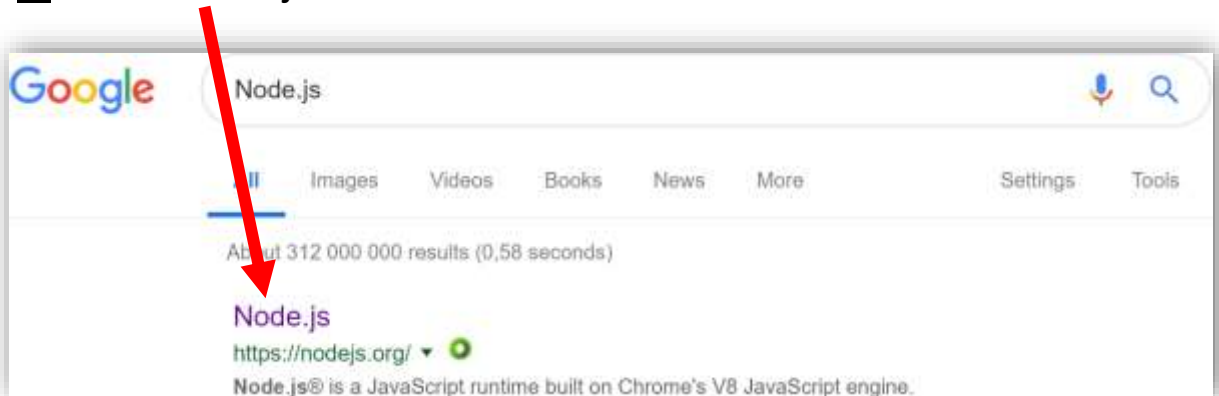


Download and rename it to **favicon.ico** and put together with other files.

**Step 25:** Search also on a Page Source name of the Title and replace your current with this one you see on the Home page.
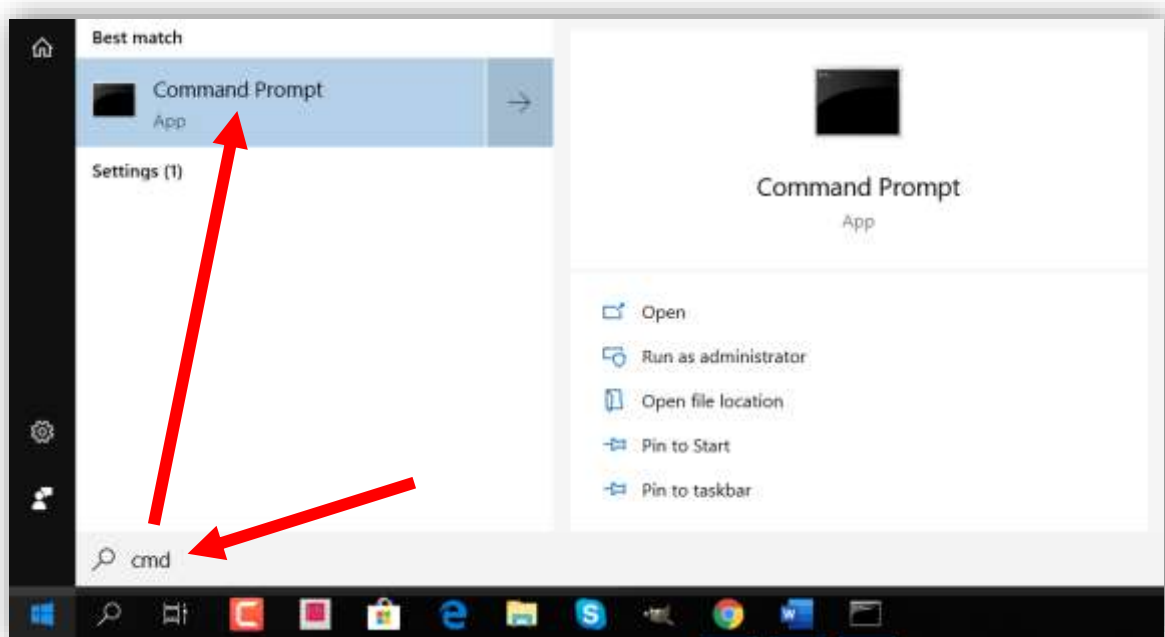
**Step 26:** Remove unused CSS from style.css file using PurgeCSS.

**1.** Install Node.js



**2.** Open Command Prompt (cmd.exe)



**3.** To **Command Prompt Paste:** npm i -g purgecss
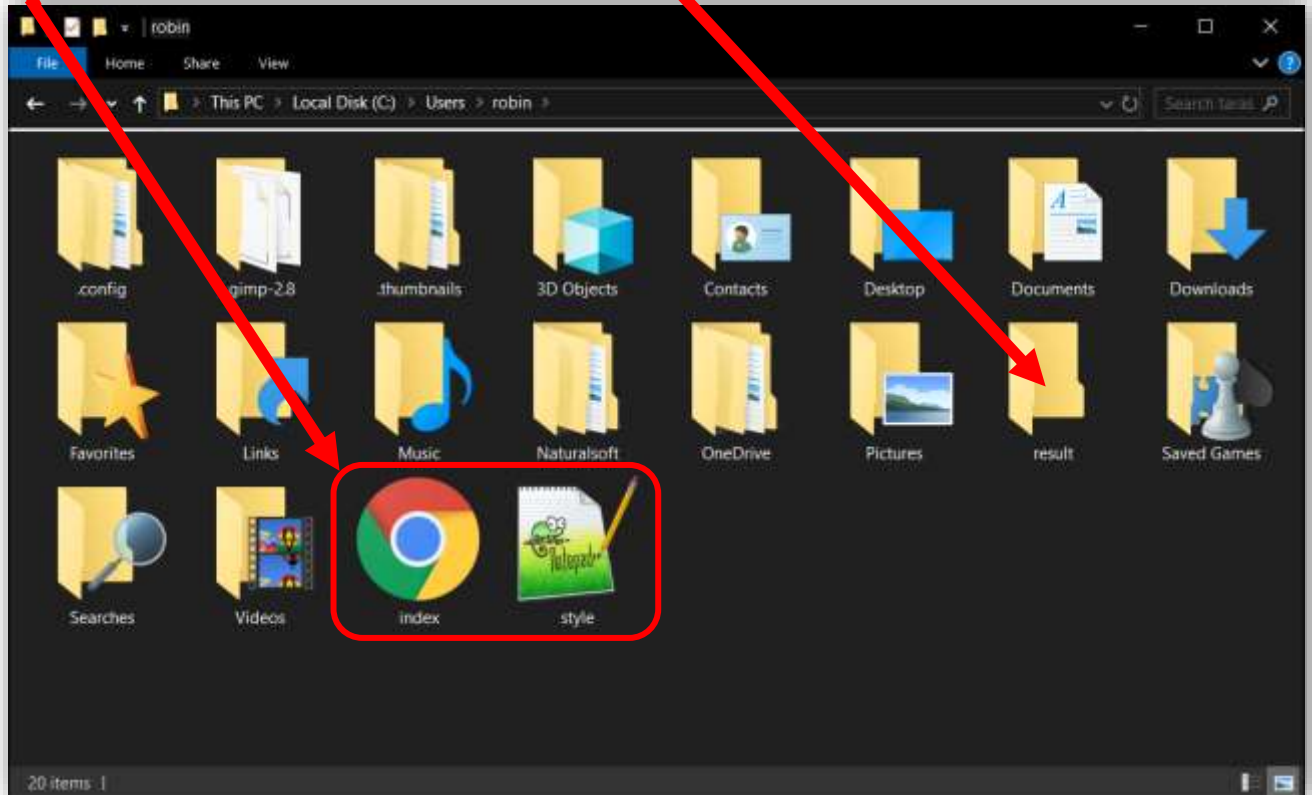and click on keyboard "Enter". This is what you should get:

**4.** Put your **css** and **html** files
to the location you see here,
in my case it's:
**C:\Users\robin**



**5.** Then create a folder called "result".



**6.** then **to Command Prompt Paste:** purgecss --css **style.css** --content **\*.html** --out **result**

and click on keyboard "Enter". Then go to **result** folder we created and replace old **style.css** file with this new smaller version.
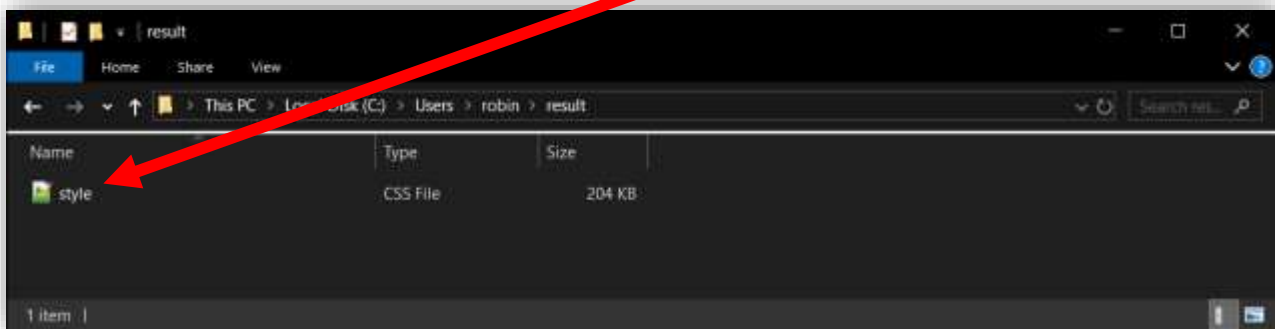
**Step 27:** Save all webpage images locally.  To quickly search image formats, use Notepad++" text editor. Highlight all text by Ctrl+A, then use Ctrl+F to open **Find** window. Fill out "Find what". Example: **\.svg**
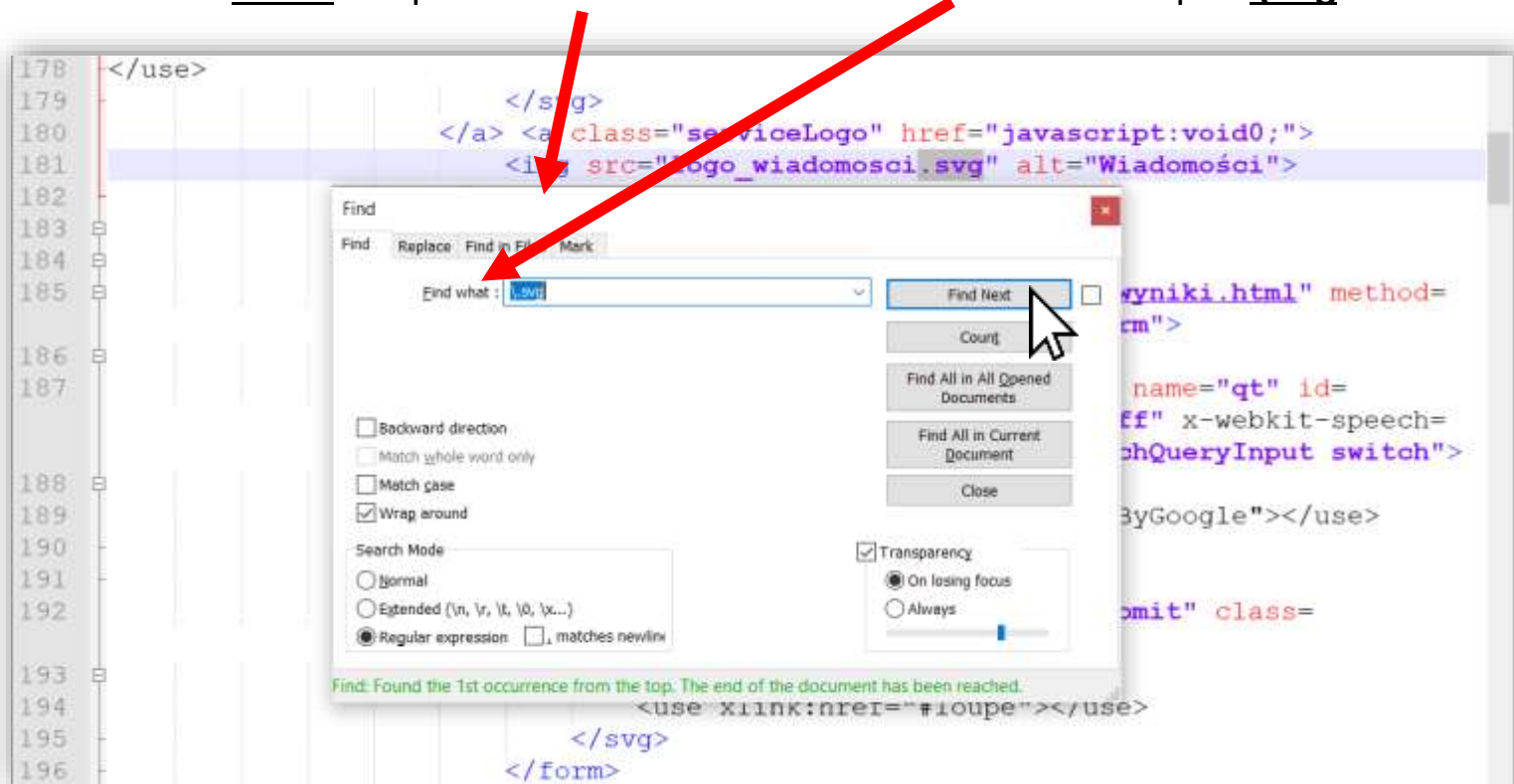


and Then press "Find Next". You have to check if index.html and style.css have any of these image formats one by one:
**\.png, \.jpeg, \.jpg, \.gif, \.svg, \.bmp**.

Save all those images locally and put them on the same folder with index.html. Remember to point all those images inside index.html and



style.css files to local directory as shown here:

**Step 28:** If you have followed this 28 step instruction you will get small in size HTML and CSS files which will weigh together without images approximately 100-300 kilobytes. After you create webpage copy, put files to .zip archive and attach it for review. **Inside a folder: "Example" I attached onet.pl webpage copy, to show how the end files should look like. Explore this webpage copy to get a good idea what you have to achieve.**

This is how ready Polish page copy will look like in my case example: