

University of Warsaw
Faculty of Economics Science

Zofia Dąbrowska
Index no.: 415129

Zuzanna Krypczyk
Index no.: 411143

Artur Skowroński
Index no.: 411423

**Wykop.pl – Webscrapping using BS4, Selenium,
Scrapy**

A credit-based paper
for the Webscrapping course
conducted by Dr Przemysław Kurek.

Warsaw, may 2022

1. Introduction

The site **wykop.pl** is considered (and rightly so) to be a forum where heckling and name-calling prevail. Nevertheless, the main page of wykop contains articles not only on current news from Poland but also the whole world. The site is updated practically every minute, so by spending 15-20 minutes of your evening, you can get reliable news of the day. The aim of this paper was to investigate which topics are currently trending on the site, which hashtags are most popular, and what people think about each post based on their responses. All that could be achieved with the use of popular web scraping libraries.

Unfortunately, due to the restrictive policy of the wykop.pl website, scraping a very large number of articles is infeasible. **Therefore, the scope of our analysis will be limited to scraping 120 pages of the site, where from each subpage we will only select the first article that is on the page.** Once we access the article, we would try to scrape the following things from it:

- title
- the nickname of the user who posted the article
- the number of likes under the article
- the number of dislikes
- number of views
- all the hashtags that have been placed under the article

Due to following limitations, we would like to **point out in advance**, the final number of records you receive may vary slightly from 120, but will certainly remain higher than the lower limit required to pass the course (i.e. 100). We have discovered two reasons for this:

- the first article may be an advertisement, which when clicked on, automatically leaves wykop.pl and goes to the advertiser's website
- the first article, may also contain content allowed only for persons over 18 years of age, which requires a prior account creation and a proof of age.

Furthermore, the site is **updated on a regular basis**, so running through the codes at different time intervals will **undoubtedly give us different results**. Nevertheless, below is a brief roadmap of how our tool should work in general.

wykop.pl Wykopaliszko 446 Hity Mikroblog

Ciekawostki Polska Covid19 Nauka Historia Rozrywka Sport Motoryzacja Świat Giełda Programowanie Technologia

STRONA GŁÓWNA: najnowsze aktywne ulubione

55 Piosenka z ruin Azovstału....
Poleca maro-grzechotnik wykopowe uniwsko

138 wykop Szeff Volkswagena chce żeby wojna się skończyła bo interes mu źle idzie .
@125procent forsal.pl #niemcy #ukraina +1 inny
Herbert Diess wezwał Ukrainę do wynegocjowania porozumienia z Moskwą, dzięki czemu będzie można znieść sankcje na Rosję, a niemiecka gospodarka uniknie szkód. Kontrowersyjne słowa szefa Volkswagena spotkały się z natychmiastową reakcją ukraińskich władz.
18 komentarzy opublikowany 3 min. temu

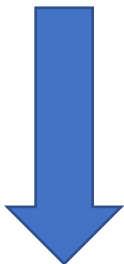
136 wykop Zelenski: Wojna się skończy, gdy Ukraina odzyska wszystkie swoje terytoria!
@Bryzoll twitter.com #swiat #wojna +2 inne
Prezydent Wołodymyr Zełenski powiedział, że jego chęć kontynuowania negocjacji maleje z dnia na dzień po zobaczeniu dowodów masakr i okrucieństw popełnionych przez Rosjan w całym kraju.
28 komentarzy opublikowany 9 min. temu

139 wykop Prywatny właściciel ogrodził plotem jezioro na Kaszubach.
@xallemorph onet.pl #polska #ekologia +2 inne
Nowy właściciel jeziora Wielkie Oczko ogrodził je ponaddwumetrowym plotem. Jezioro to znajduje się na terenie Wdzydzkiego Parku Krajobrazowego i obszaru Natura 2000. Nadzór budowlany nie widzi jednak problemu, gdyż plot nie przekracza wysokości 2,2 m, a więc nie wymaga zgody.
57 komentarzy opublikowany 14 min. temu

117 wykop Płonie niemal 800 hektarów. Ukraińcy mają problem z gaszeniem pożarów lasów
@szyna352 miz4.pl #europa #ukraina +3 inne
Prawie 800 hektarów lasów płonie w obwodzie chersońskim na południu Ukrainy. Walkę z żywiołem utrudnia okupacja tych terenów oraz działania bojowe – poinformował w środę doradca ministra ds. ochrony środowiska i zasobów naturalnych Serhij Własenko.
7 komentarzy opublikowany 18 min. temu

Select only the first article from each subpage.

From both red boxes, scrape those things that will be useful for further analysis.



wykop.pl Wykopaliszko 447 Hity Mikroblog Szukaj Dodaj Zaloguj się

Wojna za naszą wschodnią granicą toczy się również w sieci. Atak Rosji na Ukrainę wywołał lawinę artykułów, informacji i komentarzy. Nie wszystkie doniesienia są jednak prawdziwe, część z nich ma za zadanie wywołać chaos i niepewność. Nie pozwólmy na to. Starajmy się wspólnie o wysoki poziom społecznej weryfikacji treści na naszym portalu. Sprawdzajmy i weryfikujmy informacje w rzetelnych oraz oficjalnych źródłach. Wszelkie próby rozpowszechniania informacji nieprawdziwych raportujemy moderacji.

182 wykop Szeff Volkswagena chce żeby wojna się skończyła bo interes mu źle idzie .
@125procent forsal.pl #niemcy #ukraina #rosja
Herbert Diess wezwał Ukrainę do wynegocjowania porozumienia z Moskwą, dzięki czemu będzie można znieść sankcje na Rosję, a niemiecka gospodarka uniknie szkód. Kontrowersyjne słowa szefa Volkswagena spotkały się z natychmiastową reakcją ukraińskich władz.

Dodany 1 godz. temu przez:
125procent dołączył 12 lat 8 mies. temu
182 wykopał 6 zakopał <1 tys. wyświetleń

OSTATNIO POPULARNE

Warszawa: ambasador Rosji obłany farbą. Bójka przed cmentarzem 4234

Mam dość monopolu Orlenu i wysokich cen paliw! 4158

2. Description of coding parts

2.1. BeautifulSoup

[Script name BS_final.py.]

At the very beginning we put two parameters to change. The first one *START_100_SCRAPE* is boolean and setting it to True, will make our scraper start retrieving information from the next 120 pages. When it comes to parameter *SET_PAGES_TO_SCRAPE_PER_ONE_LOOP*, we can control how many links to articles our scraper should scrape. Default parameter is equal to 1, but it can be set to a maximum of 35. However, setting this parameter this high, would result in 4200 requests, which would cause our connection to get broken by the anti-scraping program.

Our scraper starts by creating an empty list where the results will be stored. Next, we create a loop which goes through the following subpages of the wykop.pl/strona/{number}. Inside the function we save to the links variable all articles found in a given tab. After that, we are looping over the values of the links variable which opens a link to each article from the subpage. Once we are inside the article, we try to scrape our 6 variables of interest.

Finally, the results are stored in the *post_summary* dictionary, which is then added to the *list_of_result*, containing all the results scraped so far. We also include a 10 second pause between subpage changes.

Below, we have included an example of the result for our scraper. In order to demonstrate how our artificial log works, we have changed the range of scraped pages to 2 on one occasion and the range of scraped articles per page to 10.

Figure 1. Example output from tool based on beautifulsoup library.

```
ARTICLE 19 FROM PAGE 1 CONTAINS CONTENT ONLY FOR ADULTS. CANNOT BE SCRAPPED.
CURRENT NON-SCRAPPED ARTICLES: 1

Article: 1 from subpage: 1 has been successfully scraped !
Article: 2, SUBPAGE: 1 - ACCESS DENIED: THIS IS ADVERTISEMENT !!!
Article: 3 from subpage: 1 has been successfully scraped !
Article: 4 from subpage: 1 has been successfully scraped !
Article: 5 from subpage: 1 has been successfully scraped !
Article: 6 from subpage: 1 has been successfully scraped !
Article: 7 from subpage: 1 has been successfully scraped !
Article: 8 from subpage: 1 has been successfully scraped !
Article: 9 from subpage: 1 has been successfully scraped !
Article: 10 from subpage: 1 has been successfully scraped !
The subpage will be changed in 2 seconds.

Article: 1 from subpage: 2 has been successfully scraped !
Article: 2, SUBPAGE: 2 - ACCESS DENIED: THIS IS ADVERTISEMENT !!!
Article: 3 from subpage: 2 has been successfully scraped !
Article: 4 from subpage: 2 has been successfully scraped !
Article: 5 from subpage: 2 has been successfully scraped !
Article: 6 from subpage: 2 has been successfully scraped !
Article: 7 from subpage: 2 has been successfully scraped !
Article: 8 from subpage: 2 has been successfully scraped !
Article: 9 from subpage: 2 has been successfully scraped !
Article: 10 from subpage: 2 has been successfully scraped !
The subpage will be changed in 2 seconds.

Below, you can find your final output, number of scraped articles is equal to 18
[{'title': 'Prezes Głapiński: od polskiego "cudu gospodarczego" do galopującej inflacji', 'usern
```

2.2. Selenium

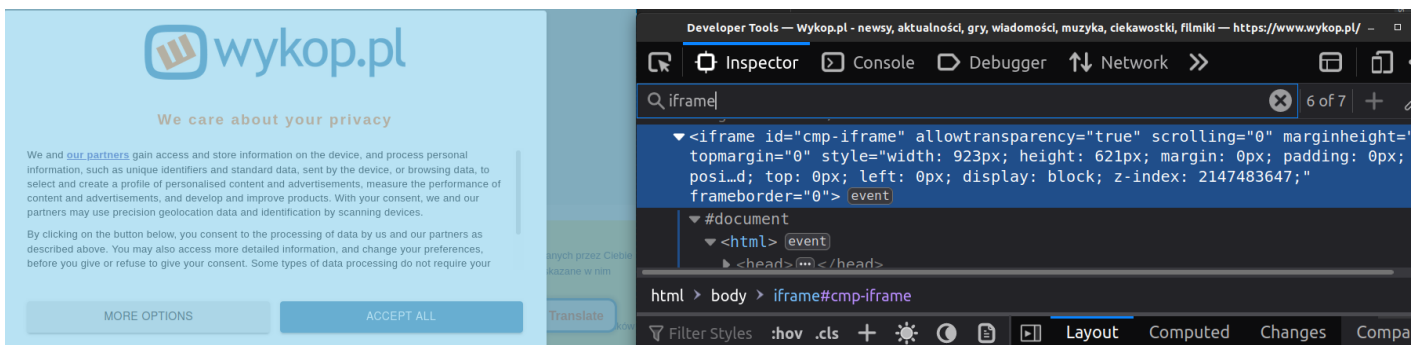
[Script name selenium_wykop.py.]

As with scrapping through the beautiful soup, we have entered the same two launch parameters (boolean: *START_100_SCRAPE*, integer: *SET_PAGES_TO_SCRAPE_PER_ONE_LOOP*) that define the number of scrapped pages and articles.

Initially, when we start firefox geckodriver, we get a pop-up about accepting cookies. It is vital to accept this, because the window is in a special "iframe". In a nutshell, the Iframe HTML element (<iframe>) represents a nested browsing context (en-US) by embedding another HTML page in the current one.

In our case, if you do not click on anything, the page does not allow you to continue browsing the site. This has forced us to change the format of the page to the "frame" structure, accept cookies and then return to the default character of the page (driver.switch_to.default_content()). After this step, we additionally accept privacy politics, located at the bottom of the page.

Figure 2. Wykop's iframe element.



Moving on to the scraping process itself, we have created a loop within a loop that goes through each article from a given tab and initially checks if the article is an advertisement (variable *check_adv_txt*). If not, the scraping process begins and the same things, indicated above, are retrieved; otherwise the article will be skipped. The obtained results, are written to the *post_summary* dictionary and then added to the *list_of_result*. After a 3-second pause, our tool returns to the previous page. If the tab's limit of scraped articles is reached, the program automatically clicks the "next" button and redirects the user to the next page after 3 seconds.

2.3. Scrapy

[Script name scrapy_wykop.py.]

Unlike the previous libraries, with scrapy we leave only one boolean control parameter which controls the number of scraped pages.

Firstly, we instantiate the *Wykop* class, that defines which parameters from the page we expect to be scrapped. Then we create another class - *WykopSpider*, that sends the request to subsequent pages of the wykop.pl/strona/ website.

Again, when we open each subpage, the program will select only the FIRST ARTICLE that appears on the page and then extract the information we are interested in. Obviously, we are aware that this first article may be an advertisement or an 18+ link (which will not be scrapped). That is why, the whole loop has been encapsulated in the try function and will return an empty field if the data cannot be scrapped. After calling the proper command (scrapy crawl wykop -o wykop.csv), we should get a csv file in our working directory, which in a later stage of our work was used for a short data analysis.

Below, we show the first 5 records of our exemplary desired output.

Figure 3. Scrapy csv output.

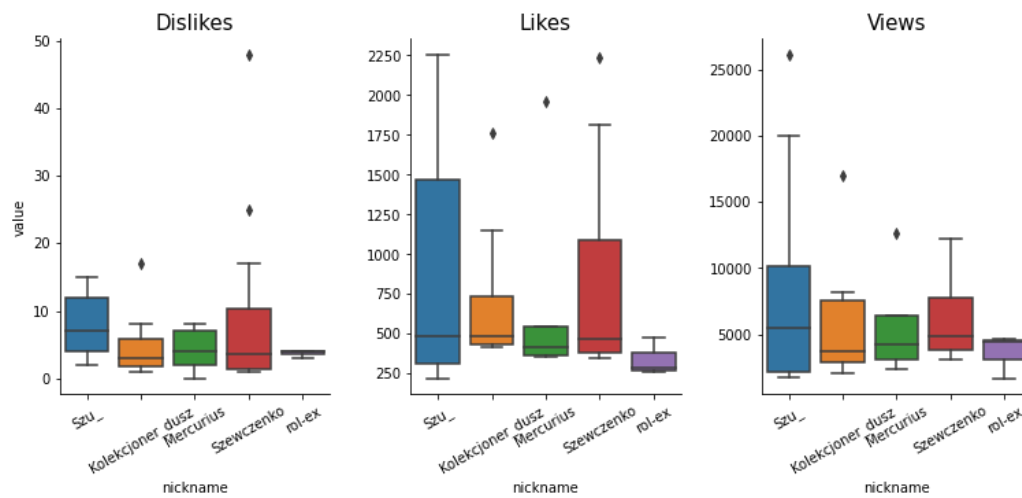
	dislikes	hashtags	likes	nickname	title	views
0	0	technologia,nauka,zainteresowania,kosmos,mars,...	186	elektryk91	NASA ma problem z Ingenuity, marsjańskim wirol...	2.4 tys.
1	0	rosja,historia,polityka,zbrodniekomunistyczne	86	Qtasus2Wielgus	Badacz zbrodni NKWD został wysłany do łagru	<1 tys.
2	14	polska,chlewoobsranygownem	269	BennyHarvey	Łódzkie. Kradł znicze wrzucając do automatu st...	12.5 tys.
3	23	polska,kultura,rozrywka	359	Wirtualnedia_pl	Polsat zrealizuje polską wersję „Sługi Narodu”	7.9 tys.
4	39	polska,technologia,amazon,afeta	476	Mirxar	Strach coś zamawiać na Amazonie. Polacy płacą ...	13.4 tys.

3. Exploratory Data Analysis

[Script name wykop_analysis.ipynb]

For this study, we analysed the data collected from 120 articles which are the first articles from 120 pages (as of 11.05.2022, 21.30). First step, aims to show the distribution of dislikes, likes and views per top 5 most frequent users.

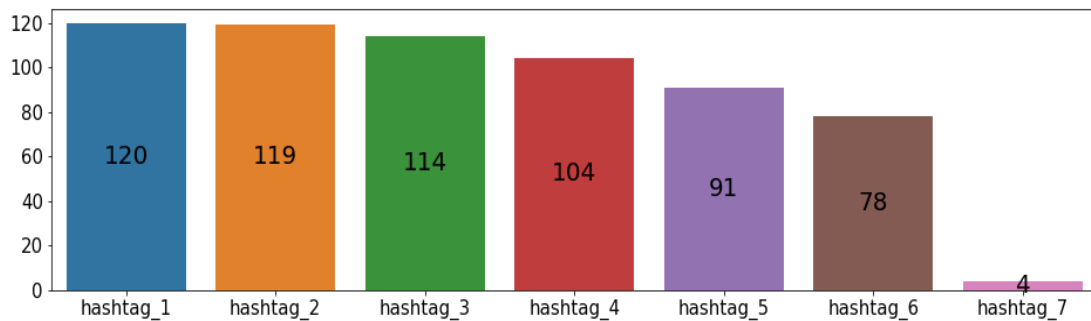
Figure 4. Boxplots for top 5 users.



The graphs above show that two users' posts stand out from the rest in terms of the number of responses they received. However, for all users, the median scores across all variables are similar.

The second step was to check the frequency of hashtag usage per article.

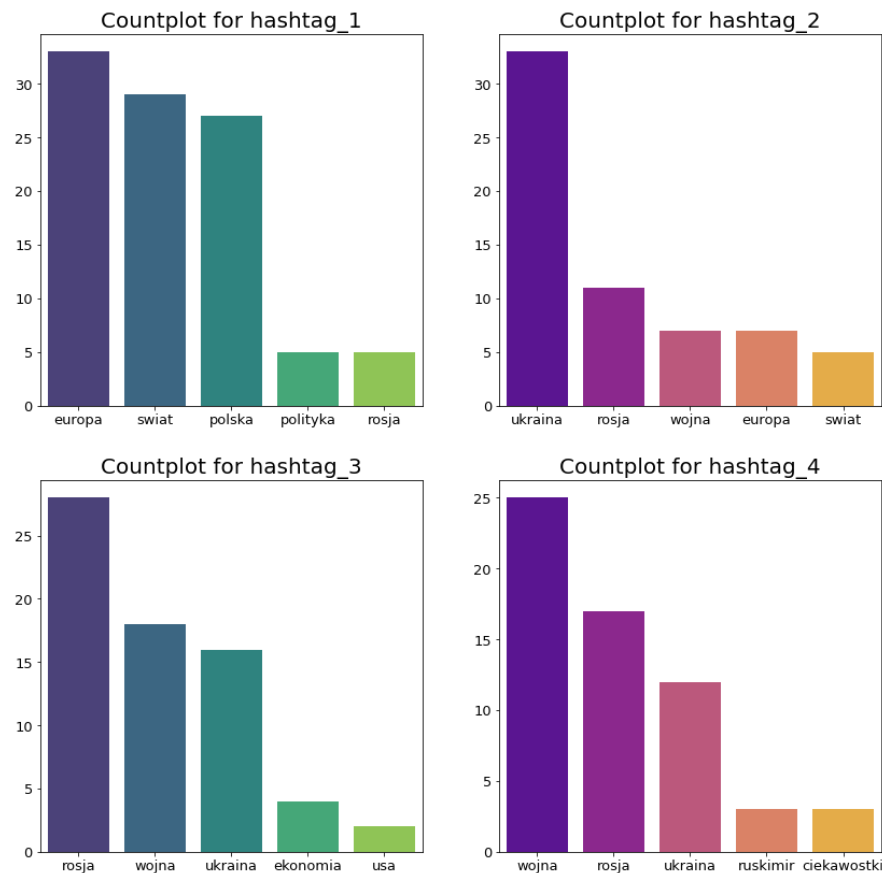
Figure 5. Distribution of hashtags used.



From the graph above, we can see that at least one hashtag was used for all articles. Therefore, we can conclude that this is a portal limitation. Interestingly, for the scraped articles, the average hashtag used is around 5.

To determine what topics are currently trending, we examined the relevance of the top 5 words for 4 hashtags.

Figure 6. Top five hashtags.



The graph above shows that the majority of the articles were about Europe, specifically the conflict between Russia and Ukraine.

4. Conclusions

In our work, we assessed articles found on wykop.pl. For this study, we analysed the data collected from the site and used three different libraries. We believe that the aim of the work, i.e. to retrieve information from at least 100 consecutive pages, has been achieved. The quality of the data is good enough for a short analysis to be carried out, which allowed us to find what news from the world are currently trending.

Should you wish to continue the work, we suggest you also download descriptions for each article, which could be used to build an NLP model. Another step worth considering is to try circumvent the anti-scrap system of the wykop.pl website (in the good spirit of the law, of course). This would make it possible, without any restrictions, to scrape every article from the site.

5. Distribution of work

Full name:	Coding part	Report
Zofia Dąbrowska Zofia Dąbrowska Zofia Dąbrowska	BeautifulSoup	BeautifulSoup part Data Analysis description Conclusions
Artur Skowroński Artur Skowroński Artur Skowroński	Selenium Data Analysis "Cleanliness" of codes	Selenium part
Zuzanna Krypczyk Zuzanna Krypczyk Zuzanna Krypczyk	Scrapy	Scrapy part Introduction Report formatting