Webscraping and Social Media Scraping | project description

Katarzyna Piotrowska, Ismayil Ismayilov

May 14, 2022

Contents

Contents 2

1 Topic and web page description 3

2 Scraper mechanics 3

3 Output 3
3.1 Technical description of the output . . . 3
3.2 Elementary data analysis . . . 3
3.3 Analysis of collected data and Consistency of the obtained results . . . 3
3.4 Comparison of scraper performance . . . 4

4 Work devision among group participants 4

1 Topic and web page description

Short description of the topic and the web page.
!!TO WRITE!!

2 Scraper mechanics

Three scrapers have been written, and each of them scrape the same piece of information from the rossmann.pl domain.
Below is a description of each scraper's mechanics, focusing on the technical side.\

- **scraper using Beautiful Soup** !!TO WRITE!!
- **scraper using Scrapy** !!TO WRITE!!
- **scraper using Selenium** !!TO WRITE!!

3 Output

3.1 Technical description of the output

In case of each program, a csv file is obtained as a result, which contains scraped information about a certain number of products that are currently on sale in Rossmann on the Polish market. When the program opens a link to the next products, it retrieves 11 pieces of information and stores them as one row in a table. Below are listed the scraped information, for each variable its name fully explains what information about the promotional product it stores.

In addition to the csv file, a log.txt file is created for each program, which contains information about its status and/or execution time.

```
## Rows: 1,008
## Columns: 11
## $ Name            <chr> "NUMEE", "NUMEE", "NUMEE", "NEUTROGENA Hydro Boost", "~
## $ RegularPrice    <dbl> 29.99, 29.99, 29.99, 56.99, 56.99, 44.99, 26.99, 33.99~
## $ PromoPrice      <dbl> 17.99, 17.99, 17.99, 42.99, 42.99, 33.99, 19.99, 24.99~
## $ Rate            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ NumberOfReviews <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ Gender          <fct> NA, NA, NA, kobieta, kobieta, NA, NA, kobieta, kobieta~
## $ Categories      <chr> "Twarz/Pielęgnacja twarzy/Serum", "Twarz/Pielęgnacja t~
## $ Availability    <fct> Available online, Available online, Available online, ~
## $ Description     <chr> "serum złuszczające do twarzy30 ml, nr kat. 393360", "~
## $ Link            <chr> "https://www.rossmann.pl/Produkt/Serum/Numee-serum-zlu~
## $ Image           <chr> "https://www.ros.net.pl/GalleryImages/product_photos/1~
```

3.2 Elementary data analysis

3.3 Analysis of collected data and Consistency of the obtained results

!!TO WRITE!!

BS RESULTS:

Variable	max	mean	min	N	NA.
Discount	43.35	29.96350	22.74	240	0
NumberOfReviews	57.00	15.35294	5.00	240	223
PromoPrice	66.99	23.30292	4.49	240	0
Rate	5.00	4.44118	3.00	240	223
RegularPrice	89.99	33.23583	5.99	240	0

!!TO WRITE!!
SCRAPY RESULTS:

Variable	max	mean	min	N	NA.
Discount	43.35	29.96350	22.74	240	0
NumberOfReviews	57.00	15.35294	5.00	240	223
PromoPrice	66.99	23.30292	4.49	240	0
Rate	5.00	4.44118	3.00	240	223
RegularPrice	89.99	33.23583	5.99	240	0

!!TO WRITE!!

Percentage of discount by product category:

Categories	mean	min	max
Makijaż/Twarz/Kremy bb i cc	36.13000	36.13	36.13
Twarz/Oczyszczanie i demakijaż/Toniki do twarzy	32.20778	29.45	37.16
Twarz/Oczyszczanie i demakijaż/Żele i pianki do twarzy	31.95636	29.17	37.16
Twarz/Oczyszczanie i demakijaż/Peelingi do twarzy	30.82200	29.75	35.02
Twarz/Pielęgnacja twarzy/Kremy do twarzy	29.99991	22.86	43.35
Twarz/Oczyszczanie i demakijaż/Płyny micelarne	29.94417	22.74	33.35
Twarz/Pielęgnacja twarzy/Kremy pod oczy	29.68538	24.25	43.35
Makijaż/Twarz/Pudry	29.40500	29.19	29.62
Twarz/Pielęgnacja twarzy/Serum	29.15957	22.86	40.01
Twarz/Pielęgnacja twarzy/Maseczki	25.02750	25.01	25.04

!!TO WRITE!!
Some plots need to be added

3.4 Comparison of scraper performance

!!TO WRITE!!

#SCRAPED_PRODUCTS	BS [SEC]	SCRAPY [SEC]	BS [MINS]	SCRAPY [MINS]
120	87	3	1.4583	0.0500
240	110	5	1.8358	0.0833
480	246	10	4.0941	0.1667

4 Work devision among group participants

Team members:

- Katarzyna Piotrowska || studentID 397061
- Ismayil Ismayilov || studentID XXXXXX

!!TO WRITE!!