EXTRACTING AND EVALUATING PRODUCT DATA FROM AUCHAN AND BIEDRONKA STORES IN GLOVO

Webscraping and Social Media Scraping Class Project

Jakub Tomczak and Paula Gwanchele

Glovo

Glovo - is scrapping this website legal?

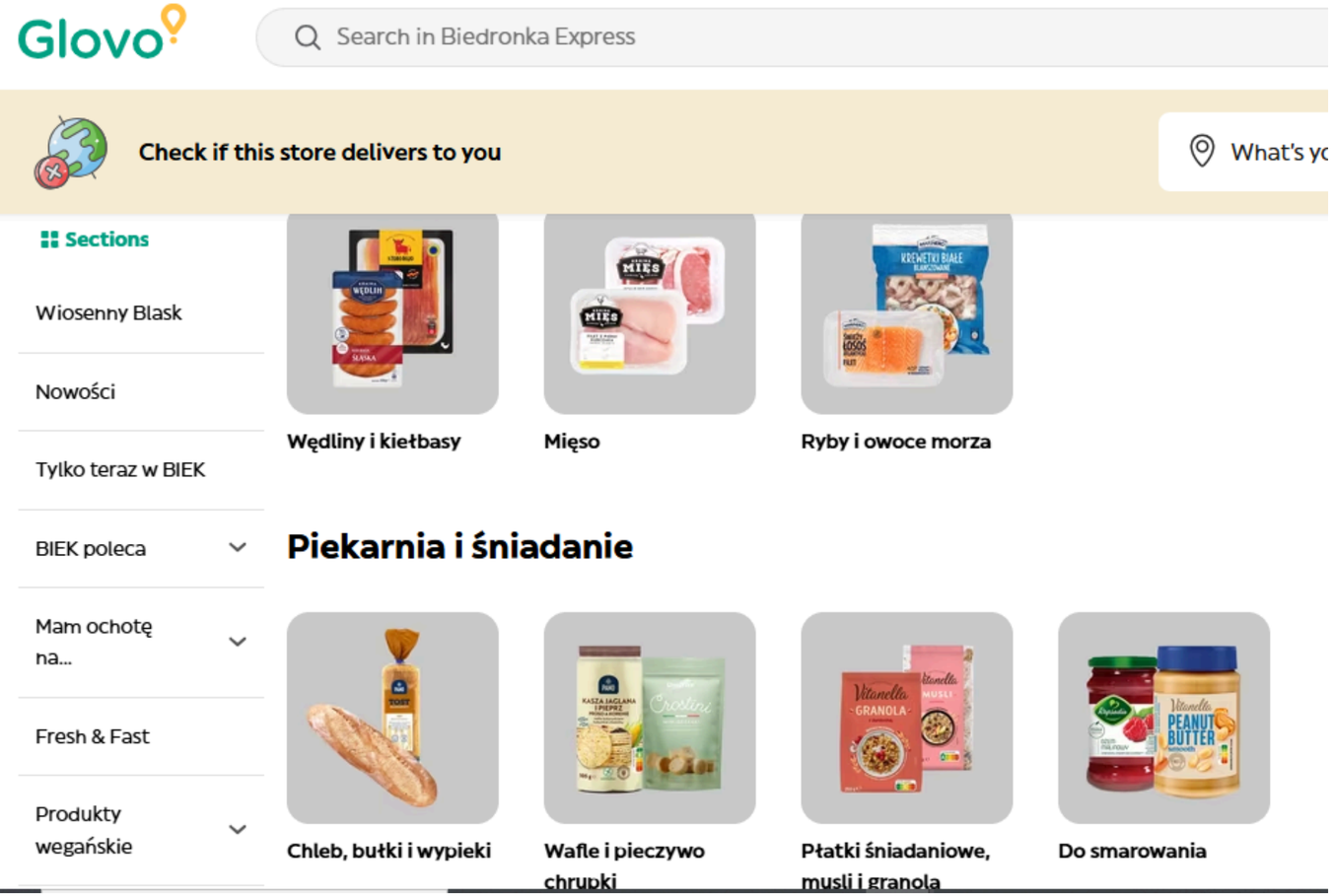
robot.txt

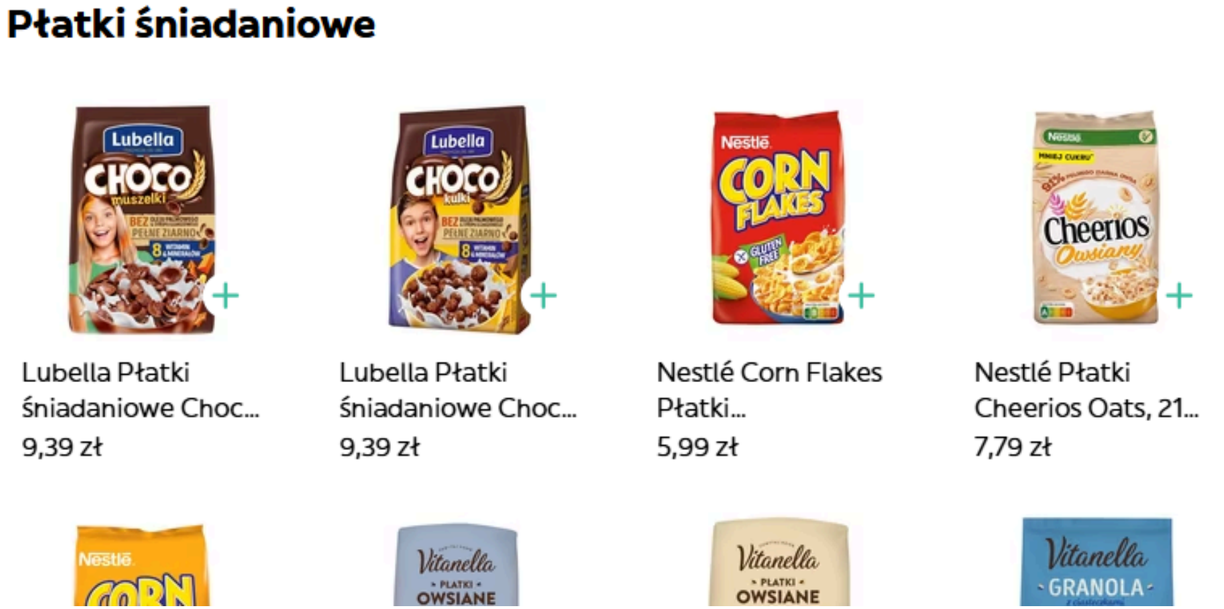
```
User-agent: PetalBot
Disallow: /

# All robots allowed
User-agent: *
Allow: /

# Sitemap files
Sitemap: https://glovoapp.com/sitemap-index.xml
```

Terms and Conditions

Deleting, bypassing or in any way tampering with the contents of the Glovo APP are all prohibited. In addition, modifying, copying, reusing, exploiting, reproducing, publicising, making second or subsequent publications of, uploading files, sending by post, transmitting, using, processing or distributing in any way all or some of the contents included in the Glovo APP for public or commercial purposes are also prohibited

Glovo - what do we scrapp?





Scrapping with BeautifulSoup

```
url = "https://glovoapp.com/pl/pl/warszawa/biedronka-express-waw/"
webpage_html = requests.get(url)
soup = BeautifulSoup(webpage_html.text, 'html.parser')

hrefs = []
for a in soup.find_all('a', {"data-test-id": "collection-link"}):
    hrefs.append(a['href'])

for i in range(len(hrefs)):
    hrefs[i] = 'https://glovoapp.com' + hrefs[i]

prices_full = []
category_dict = {}
images_full = []
```

Accessing the main page of the website

Collecting links for all product categories

Completing the links

Creating empty lists and dictionary to store data

Scrapping with BeautifulSoup

```
for url in hrefs:
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    for p in soup.find_all('span', class_ = 'product-price__effective product-price__effective--new-card'):
        pr = p.get_text(strip=True)
        pr = re.findall('\d{1,2},\d\d', pr)
        prices_full.append(pr)


    for i in soup.find_all('img', class_ = 'tile__image store-product-image'):
        img = i.attrs['src']
        images_full.append(img)


    for cat in soup.find_all('div', class_ = 'grid'):
        category = cat.find('h2', class_ = 'grid__title')
        category_name = category.get_text(strip = True)
        items = []
        for item in cat.find_all('span', class_='tile__description'):
                item_name = item.find('span')
                if item_name:
                    items.append(item_name.get_text(strip=True))
        category_dict[category_name] = items
```

Downloading the price of each product and adding it to the list after first changing the format to the appropriate one

Downloading link to every product image

Downloading the name of each product category and the names of the products assigned to it

Adding categories and product names to the dictionary

Downloading images

```
save_dir = r"C:\Users\kubas\OneDrive\Documents\Webscrapping\Project\product_images"
os.makedirs(save_dir, exist_ok=True)

for url in hrefs:
    response = requests.get(url)
    soup = BeautifulSoup(response.text, 'html.parser')

    for i in soup.find_all('img', class_='tile__image store-product-image'):
        p = i.get('src', '')
        match = re.search(r"(https://.*?\.(?:jpg|jpeg|png))", p)
        if match:
            img_d = match.group(1)
            response = requests.get(img_d)
            if response.status_code == 200:
                img_name = os.path.basename(img_d.split("?")[0])
                save_path = os.path.join(save_dir, img_name)

                with open(save_path, "wb") as f:
                    f.write(response.content)
```

Creating a folder where images will be saved

The regex matches links to images in .jpg, .jpeg and .png formats.

Downloading the actual image URL, downloading the image file from the URL and checking whether the download was successful

Removing additional URL parameters after ?, taking the file name itself from the URL and creating a full path where the file will be

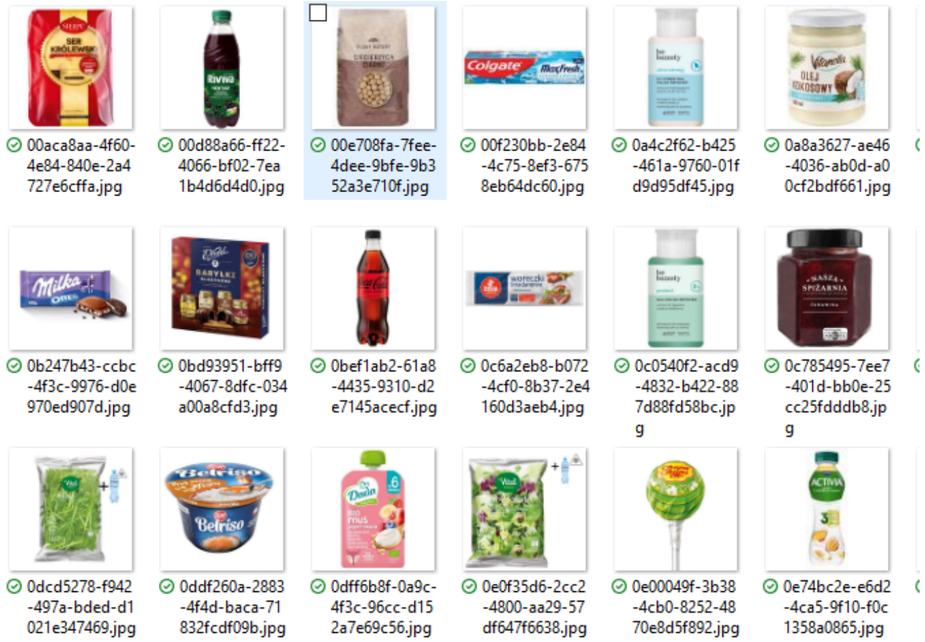
Data preparation and exporting to csv file

```
df[['Product name', 'Product weight/quantity']] = df['Product'].str.split(r', (?=\d)', n=1, expand=True)
df = df.drop(columns = ['Product'])
```

	Category	Price	Image	Product name	Product weight/quantity
0	Wiosenny Blask!	6.99	https://glovo.dhmedia.io/image/global-catalog-...	Mr Magic Ściereczki nawilżane do kuchni	60 szt.
1	Wiosenny Blask!	21.99	https://glovo.dhmedia.io/image/global-catalog-...	Cif Spray Power & Shine, przeciw kamieniowi	750 ml
2	Wiosenny Blask!	7.29	https://glovo.dhmedia.io/image/global-catalog-...	Sidolux Spray do szyb i luster Cytryna	500 ml
3	Wiosenny Blask!	21.99	https://glovo.dhmedia.io/image/global-catalog-...	Cif Spray Power & Shine, przeciw tłuszczowi	750 ml
4	Wiosenny Blask!	6.99	https://glovo.dhmedia.io/image/global-catalog-...	Mr Magic Ściereczki nawilżane do mebli	60 szt.

Separating Products column into two columns - with product name and product/ weight quantity, using regex

	A	B	C	D	E
1	Category	Price	Image	Product name	Product weight/quantity
2	Wiosenny Blask!	6.99	https://glovo.dhmedia.io/imag	Mr Magic Ściereczki nawilżane do kuchni	60 szt.
3	Wiosenny Blask!	21.99	https://glovo.dhmedia.io/imag	Cif Spray Power & Shine, przeciw kamieniowi	750 ml
4	Wiosenny Blask!	7.29	https://glovo.dhmedia.io/imag	Sidolux Spray do szyb i luster Cytryna	500 ml
5	Wiosenny Blask!	21.99	https://glovo.dhmedia.io/imag	Cif Spray Power & Shine, przeciw tłuszczowi	750 ml
6	Wiosenny Blask!	6.99	https://glovo.dhmedia.io/imag	Mr Magic Ściereczki nawilżane do mebli	60 szt.
7	Wiosenny Blask!	7.29	https://glovo.dhmedia.io/imag	Sidolux Spray do szyb i luster Arctic	500 ml
8	Wiosenny Blask!	8.49	https://glovo.dhmedia.io/imag	Ajax Płyn uniwersalny do mycia Floral Fiesta	1 L
9	Wiosenny Blask!	6.99	https://glovo.dhmedia.io/imag	Mr Magic Ściereczki nawilżane do łazienki	60 szt.
10	Czysto i świeżo!	17.99	https://glovo.dhmedia.io/imag	Bref Zawieszka do WC Color Aktiv+ Eukaliptus	2x50 g
11	Czysto i świeżo!	9.99	https://glovo.dhmedia.io/imag	Agent Max Kostka żelowa do WC Mango	75 ml
12	Czysto i świeżo!	9.99	https://glovo.dhmedia.io/imag	Agent Max Kostka żelowa do WC Limonka	75 ml
13	Czysto i świeżo!	17.99	https://glovo.dhmedia.io/imag	Bref Zawieszka do WC Color Aktiv+ Świeże Kwiaty	2x50 g
14	Czysto i świeżo!	49.99	https://glovo.dhmedia.io/imag	Finish Kapsułki do zmywarki Quantum All-in-1 l	46 szt.
15	Czysto i świeżo!	17.99	https://glovo.dhmedia.io/imag	Bref Zawieszka do WC Color Aktiv+ Cytryna	2x50 g



Scrapping with Selenium

Step 1 : Open the Auchan store on Glovo website with Selenium.

```
# Define Website URL
website = "https://glovoapp.com/pl/en/warsaw/auchan-waw"

# Initialize Selenium WebDriver
service_chrome = Service(ChromeDriverManager().install())
options_chrome = webdriver.ChromeOptions()
driver_chrome = webdriver.Chrome(service = service_chrome, options = options_chrome)

driver_chrome.maximize_window()
driver_chrome.get(website) #opens the website

# Handle Cookies
cookies_button_xpath = '''//button[@id='onetrust-accept-btn-handler']'''
try:
    WebDriverWait(driver_chrome, 10).until(
        EC.element_to_be_clickable((By.XPATH, cookies_button_xpath))
    ).click()
    print("Cookies accepted.")
except:
    print("No cookies banner found or already accepted.")
```

```
Cookies accepted.
```

Step 2: Scale the procedure of collecting links to subpages with products.

```
start = time.time()
# Find all product links
tags = driver_chrome.find_elements(By.XPATH, "//a[@data-test-id='collecti

# Collect product page links from the 'tags'
product_links = []

for tag in tags:
            href = tag.get_attribute("href")
            if (href not in product_links): # here we handle duplicates
                product_links.append(href)

print(f"✅ Collected {len(product_links)} product links.")
end = time.time()
print(end-start)
```

131 links

Step 3 : Access the collected links to extract product data and store data

```
all_products = []
all_prices = []
all_images = []

for link in product_links:

    try:
        driver_chrome.get(link)
        time.sleep(np.random.chisquare(1)+3)

        # Extract product details
        product_elements = driver_chrome.find_elements(By.CLASS_NAME, "tile__description")
        product_prices = driver_chrome.find_elements(By.CLASS_NAME, 'tile__price')
        product_images = driver_chrome.find_elements(By.XPATH, "//img[contains(@class, 'tile__image')]")



        all_products.extend([product.text.strip() for product in product_elements if product.text.strip()])
        all_prices.extend([price.text.strip() for price in product_prices if price.text.strip()])
        all_images.extend([img.get_attribute("src") for img in product_images if img.get_attribute("src")])


    except:
        continue

print(all_products)
print(all_prices)
print(all_images)
```

Step 4: Downloading product images

```
# Define the correct folder path
image_folder = r"C:\Users\Surface 4\OneDrive\Documents\Web&Social Media Scrapping\Projec
os.makedirs(image_folder, exist_ok=True)  # Ensure the folder exists

# Download and save images
for i, image_url in enumerate(product_links):
    try:
        response = requests.get(image_url, stream=True)# Stream the image
        response.raise_for_status()  # Check for request errors
        image_path = os.path.join(image_folder, f"image_{i+1}.jpg")
        with open(image_path, "wb") as file:
            for chunk in response.iter_content(1024):
                file.write(chunk)
            print(f"✅ Image {i+1} downloaded: {image_path}")

    except requests.exceptions.RequestException as e:
        print(f"⚠️ Failed to download {image_url}: {e}")

    else:
        print(f"⚠️ Skipping invalid URL: {image_url}")
```

Data preparation and exporting to csv file

```
# Process product details
category_names = []
product_names = []
weights = []

for product in all_products:
    category, name, weight = split_product_details(product)
    category_names.append(category)
    product_names.append(name)
    weights.append(weight)

# Create DataFrame
df = pd.DataFrame({
    "Category Name": category_names,
    "Product Name": product_names,
    "Weight/Size": weights,
    "Price": all_prices,
    "Image URL": all_images
})
```

✅ Data saved to glovo_auchan_products.csv successfully!

```
df.head()
```

	Category Name	Product Name	Weight/Size	Price	Image URL
0	Auchan	Ser mozzarella w zalewie solankowej	100 g	2,79 zł	https://glovo.dhmedia.io/image/global-catalog-...
1	Auchan	Ser gouda plastry	150 g	4,98 zł	https://glovo.dhmedia.io/image/global-catalog-...
2	Auchan	Ser Edamski	150 g	5,49 zł	https://glovo.dhmedia.io/image/global-catalog-...
3	Auchan	Ser gouda podpuszczkowy dojrzewający	250 g	6,66 zł	https://glovo.dhmedia.io/image/global-catalog-...
4	Auchan	Śmietana 18% homogenizowana	330 ml	3,38 zł	https://glovo.dhmedia.io/image/global-catalog-...

```
df.tail()
```

	Category Name	Product Name	Weight/Size	Price	Image UR
1854	Mlekovita	Ser Favita sałatkowo-kanapkowy bez laktozy	270 g	7,54 zł	https://glovo.dhmedia.io/image/global-catalog
1855	Auchan	Twaróg półtłusty bez laktozy	250 g	5,09 zł	https://glovo.dhmedia.io/image/global-catalog
1856	Mlekovita	Serek homogenizowany waniliowy bez laktozy	150 g	3,38 zł	https://glovo.dhmedia.io/image/global-catalog
1857	Vicenzi	Ciasteczka waniliowe z kremem cytrynowym bez g...	150 g	17,89 zł	https://glovo.dhmedia.io/image/global-catalog
1858	Arco	Sękacz śmietankowy bez laktozy	28 g	2,79 zł	https://glovo.dhmedia.io/image/global-catalog

The Analysis:

Diversity of Products:

- Auchan typically has a broader range of product categories since it’s a hypermarket. You can expect to find more diversity in terms of fresh produce, household items, and non-food products compared to Biedronka, which focuses more on groceries and essential items.
- Biedronka usually has a narrower but curated selection, focusing on budget-friendly groceries, snacks, and daily essentials.

Personal Experience from purchasing in both stores on Glovo:

Diversity = Auchan

Better or similar Prices to stores = Auchan

Offers and Promotions = Auchan

Good Packaging = Both

Quick Delivery = Both

Time management = Biedronka (due to less products 🙂)

Price Comparison:

- Biedronka is generally known for lower prices due to its discount retail model. You’ll often find cheaper prices on staple products like dairy, bread, and pantry items.
- Auchan, being a hypermarket, may have higher prices on certain items but also offers a wider range of brands, including premium and imported goods.
- On Glovo, both stores may show slightly higher prices compared to in-store due to the delivery markup.

Offers and Promotions:

- Biedronka frequently has promotions on bulk purchases and seasonal discounts, making it attractive for budget-conscious shoppers.
- Auchan might have promotions on specific brands or categories, especially during themed sales or special events.
- On Glovo, the promotional offers from both stores are usually highlighted, but Biedronka’s discounts tend to be more aggressive i.e "buy one, get one free" and bulk discounts

Thank you!