

Postępy prac projektu indywidualnego - cz. 1

Michał Banaszczyk

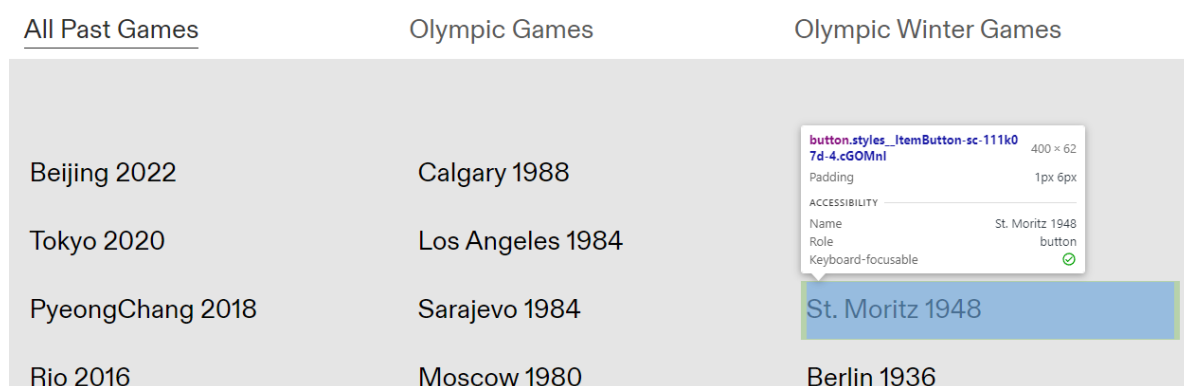
9 maja 2022

1 Napotkane problemy

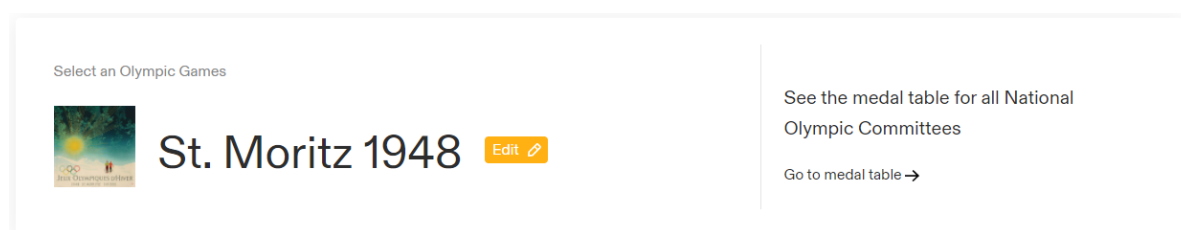
1.1 Dynamiczna zawartość i observer

Oficjalna strona Igrzysk Olimpijskich poświęcona wynikom wszystkich zawodów jest aplikacją webową, która dynamicznie zmienia zawartość strony przy każdej interakcji, co zaprezentowano poniżej.

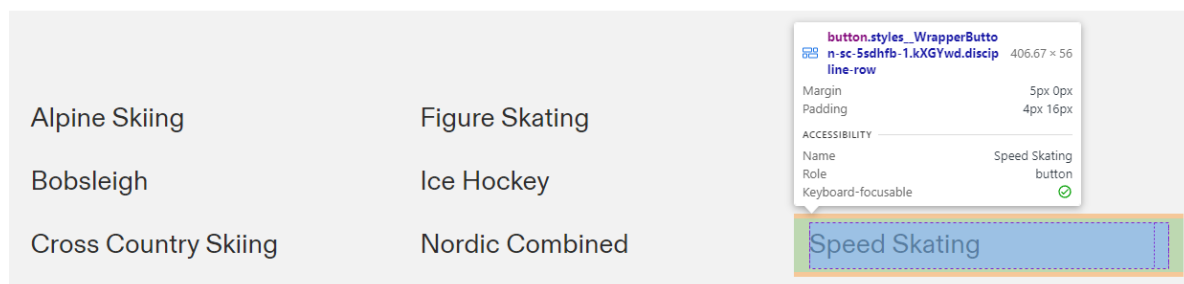
Select an Olympic Games



Zrzut ekranu 1: Widok aplikacji przed wybraniem konkretnych igrzysk



Select a Sport



Zrzut ekranu 2: Lista sportów po wybraniu konkretnych igrzysk

```
> document.querySelectorAll('[class^=styles__ItemButton]')
< NodeList(106) [button.styles__ItemButton-sc-111k07d-4.cGOMnI,
  button.styles__ItemButton-sc-111k07d-4.cGOMnI,
  button.styles__ItemButton-sc-111k07d-4.cGOMnI,
```

Zrzut ekranu 3: Lista elementów poszczególnych IO przed interakcją

```
> document.querySelectorAll('[class^=styles__ItemButton]')
< ▼ NodeList [] ⓘ
  length: 0
```

Zrzut ekranu 4: Lista elementów poszczególnych IO po interakcji

Guziki w poszczególnych listach aplikacji na szczęście są elementami innych klas, zatem po każdej interakcji wywoływane jest API observera, który czeka na załadowanie się oczekiwanej klasy guzików i dopiero wtedy kontynuowane jest scrapowanie. Puppeteer także posiada metodę obiektu strony `waitForSelector`, natomiast w naszym zastosowaniu czekanie odbywa się w środowisku przeglądarki, a nie w Node, zatem niemożliwym było skorzystanie z owej metody.

1.2 Skrypty ewaluacyjne

Największą niedogodnością okazało się usuwanie elementów przez aplikację po każdej interakcji, zamiast np. chowania ich. W związku z tym, niemożliwe było tworzenie handlerów elementów za pomocą metody `evaluateHandle` i obsługiwanie ich z poziomu Node'a. Zamiast tego, całą logikę obsługi aplikacji trzeba było wstrzyknąć do zawartości strony jako skrypt ewaluacyjny. Z tego samego powodu koniecznym było wywołanie metody `querySelectorAll` przy każdym przejściu przez drzewo dyscyplin, gdyż elementy pobrane w poprzednim przejściu nie istniały już na stronie w trakcie następnego.

2 Postępy








2.1 Zbiór linków z danymi z poszczególnych dyscyplin sportowych

Po wykonaniu wyżej opisano scrapowania otrzymano zbiór wszystkich statycznych stron z wynikami poszczególnych dyscyplin, zapisany jako plik CSV. W poniższej tabeli zaprezentowano pierwsze 20 linijek tego pliku (z 6575).

Host	Year	Type	Sport	Event	Link
Beijing	2022	Winter	Alpine Skiing	Men's Downhill	../alpine-skiing/men-downh..
Beijing	2022	Winter	Alpine Skiing	Men's Giant Slalom	../alpine-skiing/men-giant..
Beijing	2022	Winter	Alpine Skiing	Men's Super-G	../alpine-skiing/men-super..
Beijing	2022	Winter	Alpine Skiing	Men's Slalom	../alpine-skiing/men-slalo..
Beijing	2022	Winter	Alpine Skiing	Mixed Team Parallel	../alpine-skiing/mixed-tea..
Beijing	2022	Winter	Alpine Skiing	Men's Alpine Combined	../alpine-skiing/men-alpin..
Beijing	2022	Winter	Alpine Skiing	Women's Giant Slalom	../alpine-skiing/women-gia..
Beijing	2022	Winter	Alpine Skiing	Women's Alpine Combined	../alpine-skiing/women-alp..
Beijing	2022	Winter	Alpine Skiing	Women's Super-G	../alpine-skiing/women-sup..
Beijing	2022	Winter	Alpine Skiing	Women's Downhill	../alpine-skiing/women-dow..
Beijing	2022	Winter	Biathlon	Men's 12.5km Pursuit	../biathlon/men-12-5km-pur..
Beijing	2022	Winter	Biathlon	Men's 10km Sprint	../biathlon/men-10km-sprin..
Beijing	2022	Winter	Alpine Skiing	Women's Slalom	../alpine-skiing/women-sla..
Beijing	2022	Winter	Biathlon	Men's 15km Mass Start	../biathlon/men-15km-mass-..
Beijing	2022	Winter	Biathlon	Mixed Relay 4x6km (W+M)	../biathlon/mixed-relay-4x..
Beijing	2022	Winter	Biathlon	Men's 4x7.5km Relay	../biathlon/men-4x7-5km-re..
Beijing	2022	Winter	Biathlon	Men's 20km Individual	../biathlon/men-20km-indiv..
Beijing	2022	Winter	Biathlon	Women's 12.5km Mass Start	../biathlon/women-12-5km-m..
Beijing	2022	Winter	Biathlon	Women's 10km Pursuit	../biathlon/women-10km-pur..
Beijing	2022	Winter	Biathlon	Women's 15km Individual	../biathlon/women-15km-ind..
...

Każda z tych stron zawiera zbiór divów klasy `styles__Row-*`. W nim z kolei dane, które chcemy uzyskać zapisane są w przedstawiony niżej sposób. Scrapowanie zatem powinno pójść bezproblemowo.

Summary

Rank	Team	Participant	Results	Notes
	 AUT	 Johannes Strolz	2:31.43	
	 SUI	 Justin Murisier	2:32.29	
5	 AUT	Schwarz	2:32.71	
6	 SLO	as Szollos	2:33.18	
7	 AUT	napnael Haaser	2:33.27	

Zrzut ekranu 5: klasy poszczególnych danych