Leffe Circuit - de scraper

Een toelichting voor het gebruik en onderhoud van de webscraper voor https://MijnKNLTB.toernooi.nl/

Versiebeheer

Auteur(s)	Wijzigingen	Datum	Status	Versie
Merijn Schreijen	Initiële versie	18/12/23	concept	0.1
Merijn Schreijen	Scraper Modules stappenplan	16/01/24	Concept	0.2

Distributiebeheer

Ontvangers	Datum	Versie
Jelle Huibregtse, Jeroen Vos	19/12/23	0.1

Inhoudsopgave

Wat is de scraper?

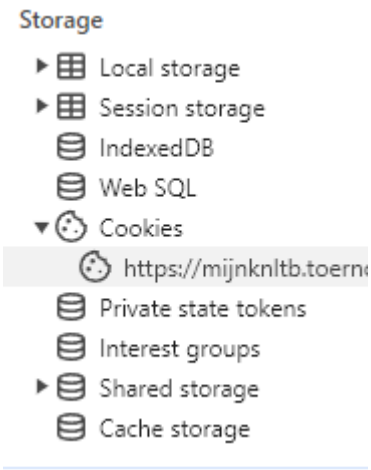
De scraper is een tool in de back-end van het Leffe Circuit systeem. Deze tool haalt de benodigde informatie van de webpagina's van de MijnKNLTB website om de scores van toernooien te berekenen.

Configuratie

Cookie

Voor het werken van de scraper is het nodig dat er een cookie wordt geconfigureerd in de software. De volgende stappen moeten hiervoor worden ondernomen:

1. Ga naar https://mijnknltb.toernooi.nl/.
2. Open de Ontwikkelaars Tools van je webbrowser. Voor Chrome en vergelijkbare browsers kan dit door naar het menu rechts boven te navigeren (drie puntjes), en onder het menu 'Meer tools' te klikken op 'Ontwikkelaars tools'.
3. Er opent nu onderaan de webbrowser een nieuw scherm. Ga in dit scherm naar het tabblad 'Applicatie'.
4. In de lijst links, onder het kopje 'opslag', vouw je het menu 'Cookies' open en klik je op de knop met de mijnknltb link.



5. Er is nu een lijst met cookies te zien. Degene genaamd 'st' hebben we nodig. Dubbelklik bij deze cookie op het hokje onder 'Value' en kopieer deze volledig.
6. We zijn nu klaar met de webbrowser, en gaan naar de back-end software van het Leffe Circuit systeem.
7. In de bestanden van het Leffe Circuit systeem, open het bestand genaamd `.env`
8. Op één van de regels van dit bestand is een variabele genaamt `MIJNKNLTB_COOKIE` te vinden. Achter het = teken plak je de waarde die we in de eerdere stap hebben gekopieerd tussen twee aanhalingstekens.

```
5   MIJNKNLTB_COOKIE="st=1=1043&exp=45601.6042241551&c=1&cp=31&s=0"
```

Onderhoud

Nieuwe data toevoegen

Als er nieuwe data weergegeven moet worden op de website die vanaf de mijnknltb komt, zal er ook een nieuwe module voor de scraper moeten worden aangemaakt.

Voorwaarden

- Het Prisma schema is geüpdatet met de nieuwe velden binnen ScoreSchemas, PlayerSchemaScores, of Tournament
- In het geval het de berekening van scores betreft, is deze functionaliteit binnen de schemas.service.ts bestand ook aangepast.
- De DTO's zijn aangepast met de juiste velden.
- De nieuwe data uit de DTO’s wordt naar de database gestuurd met prisma.

Stap 1: Update `scraper.result.ts`

De nieuwe data is onderdeel van de data die je uit de mijnknltb wilt scrapen. Tijdens het scrapen wordt het TournamentScraperResult in dit bestand ingevuld met de data. Voor de nieuwe data moeten dus ook nieuwe velden worden aangemaakt in dit object.

Waarvoor is de data?	Op welke pagina van mijnknltb kan ik de data vinden?	Waar hoort het veld thuis in `scraper.result.ts`:
Tournament	Elke pagina - algemene informatie staat altijd bovenaan	TournamentScraperResult
Schema	De lijst met alle schema's	SchemaScraperResult
schema	De pagina van het schema	SchemaContentScraperResult
Speler	De lijst van spelers	PlayerScraperResult
Speler	De pagina van de speler	PlayerContentScraperResult

Aan de hand van bovenstaande tabel kan je de velden aanmaken in het object

Stap 2: Update `persist.scraper-result.ts`

Dit bestand is te vinden in het `scripts` mapje van de scraper.
Als het goed is, is er voor de nieuwe data nu een veld aangemaakt in de DTO's, en in de corresponderende objecten in de `scraper.result.ts` objecten. Zorg er nu voor dat in de functies voor het persisteren van de data uit `scraper.result.ts` de data juist in de DTO's wordt gezet. Hieronder is te zien waar de data naartoe moet aan de hand van waar de data voor bedoeld is:

Tournament → UpdateTournamentDto
Schemas → CreateSchemaScoresDto
Spelers → CreatePlayerDto

Stap 3: Voeg een nieuwe scraper module toe

In het `scrapermodules` mapje moet nu een nieuwe module worden toegevoegd om de data van de mijnknltb website af te halen. Maak een nieuw bestand aan, waarbij de naam het format heeft van de scope, gevolgd door een punt en daarna de naam van het veld. Als de scope dat van een speler is, en ik zijn knltb id ophaal, zal het dus de naam `player.knltbid.ts` hebben.

In het bestand maak je een Class aan die deze naam ook heeft, in ons voorbeeld dus `PlayerKnltbIdScraper`.
Deze class implementeert altijd een type scraper. Aan de hand van onderstaande tabel is te zien welke scraper er nodig is:

Waarvoor is de data?	Op welke pagina kan ik de data vinden?	Welk Scraper interface wordt geïmplementeerd?
Tournament	Elke pagina - algemene informatie staat altijd bovenaan	TournamentScraper
Schema	De lijst met alle schema's	SchemaScraper
Schema	De pagina van het schema	SchemaContentScraper
Speler	De lijst van spelers	PlayerScraper
Speler	De pagina van de speler	PlayerContentScraper

Voeg aan de klasse de volgende twee velden toe:

- `for`: de naam van het veld waar de data hoort als string (in ons voorbeeld dus "knltb_id")
- `type`: een enum van het type ScraperLevel. Dit staat gelijk aan de interface benaming. (SchemaContentScraper heeft als level SchemaContent, etc).

Nu kan je beginnen met het scrapen van de html door middel van cheerio in de `scrape` functie. De functie krijgt in de request parameter de html meegestuurd.
Ook zijn er in twee gevallen de `options` parameter meegestuurd.

- Gaat het om informatie uit de lijst met spelers, wordt altijd `options.player_identifier` meegestuurd.
- Gaat het om informatie uit de lijst met schema's, zal `options.schema_number` er erbij zitten. Dit om de mogelijkheid te geven om de juiste rij uit de lijst te kunnen kiezen.

Voor het scrapen gebruiken wij Cheerio. Dit is een library voor Node. (https://cheerio.js.org/)
Cheerio fungeert als een uitbreiding op jQuery. Je kan HTML inladen via de cheerio.load(), wat een cheerioelement teruggeeft. Daarna kan je via css selectors het document doorlopen en attributen of tekst uit elementen halen.

Nu kan aan het eind van de functie de gevonden data voor het veld in de return statement gezet worden. Deze wordt dan automatisch door de rest van het scraper systeem doorgezet naar het TournamentScraperResult object doorgezet, en daarna gepersisteerd.