

Rapport - Sprint 3

In deze sprint hebben we ons vooral gefocust op het integreren van alle losse onderdelen.

Front-end To Back-end:

Terwijl we in de vorige sprints de forms hadden die ingevuld moesten worden om Admin users en RSS feeds toe te voegen, deden die forms op dat moment nog niets.

Nu hebben we ervoor gezorgd dat de entries die we daar invullen ook effectief in de Database komen te staan, op voorwaarde dat het valid entries zijn uiteraard.

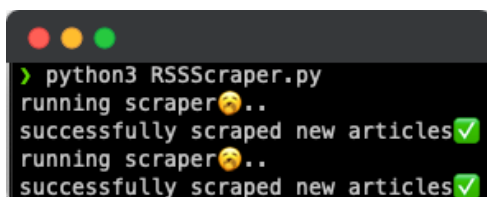
Infinite Scrolling:

Om het ophalen van artikelen sneller te maken, hebben we infinite scrolling geïmplementeerd. De fetch() functie bij onze artikel fetcher en de respectievelijke flask route die toegang naar de api geeft werden aangepast om een page argument aan te vaarden. Dat argument bepaalt welke pagina, of groep van een bepaald aantal artikelen (in dit geval 10) wordt gehaald uit de resultaten van de uitgevoerde query. Het scrollen naar de onderkant van de hoofdpagina neemt het pagina argument toe met 1 en roept de fetch() functie opnieuw.

Er werd gebruikgemaakt van de Infinite Scroller component uit de react-dom bibliotheek. Testcases:

We hebben een test-environment opgesteld om de implementaties en functies op de database te testen. Hiervoor moesten we een aparte omgeving voor maken om te voorkomen dat de inhoud van de database van de app zelf wordt beïnvloed.

Periodic RSS Scraper:

A terminal window with a dark background and three colored window control buttons (red, yellow, green) at the top left. The terminal displays the following text:

```
> python3 RSSScraper.py
running scraper😓..
successfully scraped new articles✅
running scraper😓..
successfully scraped new articles✅
```

In de vorige sprint werd de RSS scraper opnieuw opgeroepen elke keer dat de gebruiker de homepage refreshte. Dit is uiteraard niet zeer efficiënt. Vandaar dat we deze keer de RSS Scraper periodisch oproepen.

Als de gebruiker nu de homepage refresht, hoeven we de artikelen niet opnieuw te scrapen, we halen ze simpelweg uit de database. En met behulp van Infinite Scrolling zorgt dit voor een veel lagere laadtijd, en dus ook betere gebruiksvriendelijkheid, als ervoor.

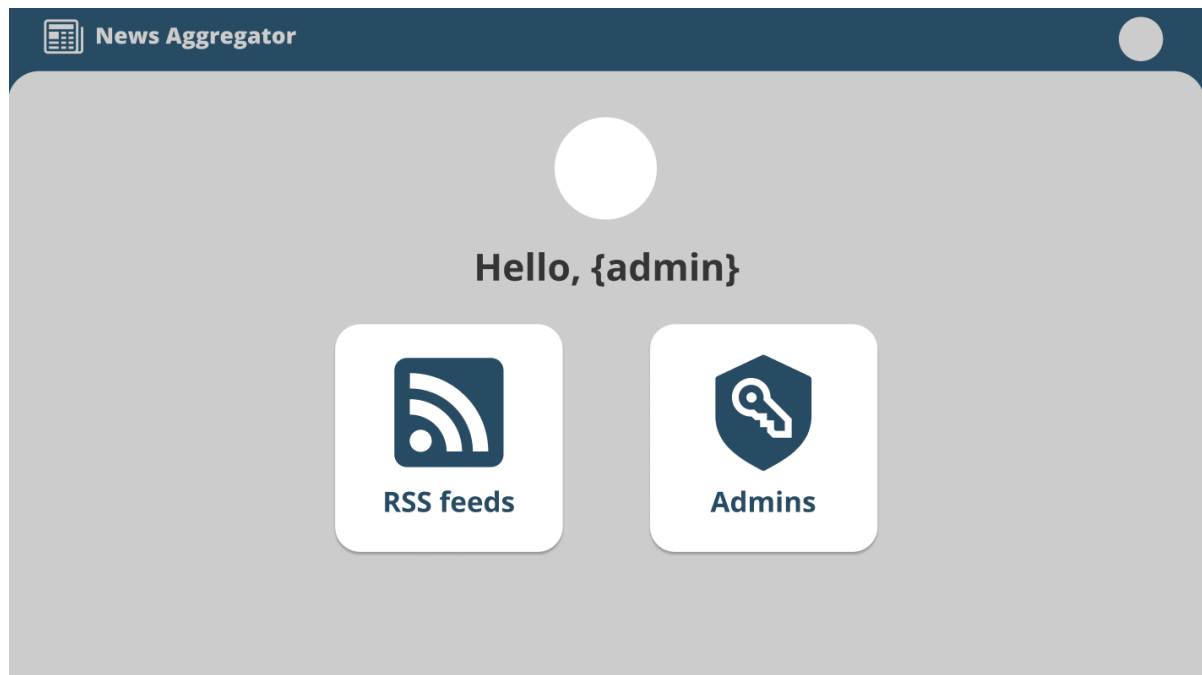
Ook haalt de scraper deze keer de RSS feeds effectief vanuit de database.

De RSS feeds die we reeds hebben getest:

RSS Link	Name
https://www.vrt.be/vrtnws/nl.rss.articles.xml	VRT Nieuws
https://www.hln.be/home/rss.xml	Het Laatste Nieuws
https://www.gva.be/rss/section/ca750cdf-3d1e-4621-90ef-a3260118e21c	Gazet Van Antwerpen
https://www.nieuwsblad.be/rss/section/55178e67-15a8-4ddd-a3d8-bfe5708f8932	Nieuwsblad
https://www.demorgen.be/in-het-nieuws/rss.xml	De Morgen
https://sporza.be/nl.rss.xml	Sporza
https://www.thebulletin.be/rss.xml	The Bulletin
https://www.standaard.be/rss/section/1f2838d4-99ea-49f0-9102-138784c7ea7c	De Standaard
https://www.hbvl.be/rss/section/D1618839-F921-43CC-AF6A-A2B200A962DC	Het Belang Van Limburg

Front-End Design:

Het design van de admin pagina hebben we verder uitgebreid. Ervoor was het simpelweg een pagina waar je RSS feeds en admins kan toevoegen (en verwijderen). Nu hebben we extra paginas om dan een lijst te hebben van de RSS feeds en Admins elk op hun eigen pagina.





Artikelen linken:

Wanneer de artikelen in de database zijn ingeladen, kan je 'link_articles()' runnen. Deze functie gebruikt het TF-IDF algoritme zoals geïmplementeerd bij de vorige sprint. Wanneer het algoritme een score hoger dan 0.8 teruggeeft, zullen de titels als gelijk worden beschouwd. Merk op: een artikel wordt ook vergeleken met zichzelf en zal een score van 1.0 of hoger teruggeven. Dit is echter geen probleem aangezien dit opgelost wordt door een andere functionaliteit.

Sommige nieuws-outlets zullen heel gelijkaardige titels hebben, zoals 'CLUBNIEUWS APRIL. De belangrijkste nieuwtjes over de Limburgse basketbalclubs.' en 'CLUBNIEUWS APRIL. De belangrijkste nieuwtjes over de Limburgse voetbalclubs.'. Dit zijn uiteraard twee heel verschillende artikelen die wel gematcht worden. Het is niet verkeerd om er van uit te gaan dat dezelfde site nooit twee keer hetzelfde artikel zal plaatsen, dus checken we of de url hetzelfde is met behulp van een reguliere expressie: 'https:\\V([0-z]+\\.)+[0-z]*V'. Hierdoor zullen artikelen ook niet als duplicaat van zichzelf in de database worden toegevoegd.

Inloggen en Registreren:

Een gebruiker kan zich nu ook registreren op onze site. We houden dan een hashed versie van zijn wachtwoord bij.

Zo kan de gebruiker zich inloggen met de username en het wachtwoord dat hij heeft gekozen bij het registreren.