**Verzamelen van Nu.nl links en semi-automatisch scrapen van comments**

**Wat heb je nodig of is handig om te hebben?**

1. Chromedriver. Ik heb de meest recente versie al in de map staan (voor Windows). Als je een Mac/Linux hebt, moet je deze downloaden via <https://chromedriver.chromium.org/downloads> . Je kan hem dan ook in de *Liesje*-map zetten, maar moet dan wel een kleine aanpassing doen in het NuCommentScraper.py script: line 89 “chromedriver.exe” wordt “chromedriver”

2. DB Browser for SQLite: niet essentieel om te hebben, maar wel handig. Ik sla de links van Nu.nl op in een SQLite-bestand (URLlist.db) en als je deze in wil zien of aan wil passen, dan is dit wel handig om te hebben. <https://sqlitebrowser.org/>

3. Notepad++ (of iets dergelijks, niet beschikbaar voor Mac geloof ik. Ik gebruikte toen Brackets): Ik sla de comments op als JSON bestand, en met Notepad++ zijn deze mooi geformatteerd als je ze opent (<https://notepad-plus-plus.org/downloads/>)

4. Een Google\_Profile: ook dit is niet absoluut nodig, maar met scrapen is het wel netjes en ook handig om een custom Google Chrome profiel te hebben. Ik heb de mijne al bijgevoegd in de map (Google\_Profile). Je kan ook je eigen profiel aanmaken. Heb ik lang niet gedaan, maar ik geloof dat <https://support.leapwork.com/hc/en-us/articles/360020888292-How-to-create-and-use-Chrome-Profile-in-LEAPWORK> dit werkt? Als je hier geen zin in hebt kun je line 97 van NuCommentScraper.py wegcommenten.

5. De volgende Python libraries:

- BeautifulSoup (<https://pypi.org/project/beautifulsoup4/>)

- Selenium (<https://pypi.org/project/selenium/>)

- Regex (<https://pypi.org/project/regex/>)

- Multiprocessing (<https://pypi.org/project/multiprocessing/>)

- Pyperclip (<https://pypi.org/project/pyperclip/>)

- Dateutil (<https://pypi.org/project/python-dateutil/>)

**Nu.nl linkscraper**

Dit gedeelte is niet zo ingewikkeld:

1. Draai **NuLinkScraper.py**

Dit script verzamelt alle links van de RSS feeds van de Nu.nl subpagina’s (tenminste, alle subpagina’s die je vindt onder “Meer” bovenaan. Als je meer subpagina’s wil toevoegen moet je de RSS pagina toevoegen aan Line 13.

Met dit script verzamel je de 10 meest recente artikelen per onderwerp (want de RSS feeds bevatten alleen maar de 10 meest recente artikelen). Als je wil dat hij geen artikelen mist moet je hem dus regelmatig draaien. Je zou een wrapper kunnen schrijven dat hij ieder uur nieuwe artikelen ophaalt, bijvoorbeeld.

De artikelen slaat hij op in **URLlist.db**, met (bijna) alle informatie van de RSS feed. Ik stuur alvast eentje mee met het resultaat van één keer draaien een paar dagen geleden. Je zou al die rijen in de database weg kunnen halen om met een schone lei te kunnen beginnen.

**Nu.nl commentscraper**

1. Draai **NuCommentScraper.py**

2. Als het goed is opent het script automatisch in Chrome een Nu.nl artikel.

3. Klik op de knop om de reacties op te halen, en op alle “Toon meer reacties” en “Laad meer reacties”.

4. Nu wil je het juiste element kopiëren. Dat is het element dat begint met **div.talk-tab-content.TabContent** (zie **Welk element.png**). Mijn strategie om deze te vinden is om **op de bovenste comment Rechtermuisknop --> Inspecteren** te doen, en dan in die lijst omhoog scrollen totdat ik het goede element gevonden heb (alle comments zijn gehighlight als je op dit element klikt). Vervolgens doe je **Rechtermuisknop op dit element --> Copy --> Copy element**.

5. Het script verwerkt nu dit element en het artikel en slaat het op in de **ArticleJson** map, en laadt automatisch het volgende artikel zodat je heerlijk kan doorwerken. 😉

Ik heb twee Jsonfiles in de ArticleJson gezet om te kijken of alles werkt en hoe het eruitziet. Alleen was ik hier te lui om alle comments te laden. Dus deze kun je beter weggooien (en als je met de door mij verzamelde links wil werken moet je daarna ook de X bij “IsCollected” van de URLlist.db bij de eerste twee rijen weghalen).

Dit script verzamelt trouwens alleen artikelen die nog niet verzameld zijn (als er iets mis is gegaan moet je in de URLlist bij “IsCollected” de X verwijderen, en ook het bijbehorende JSON bestand). En het verzamelt alleen artikelen die meer dan 24 uur geleden zijn gepost. Verder verzamelt de scraper best wel wat metadata naast de comments (kijk maar naar een JSON-bestand). Want het leek mij interessante data voor allerlei onderzoek.

**Succes!**