Toelichting bij repository
Utrecht Data School / Planbureau voor de Leefomgeving

In deze mappen tref je een selectie scripts, queries en samples van datasets behorende bij het onderzoek dat Utrecht Data School in 2021 opdracht van het PBL uitvoerde. Het doel van deze repository is om inzicht te geven in de computationele methodes, die zijn ingezet om het online debat omtrent twee onderzoeksthema's – de circulaire economie en de verduurzaming van woningen – in kaart te brengen. Deze repository is niet allesomvattend, omdat sommige van de gebruikte tools zich niet lenen voor het delen van bestanden i.v.m. dependencies, plugins et cetera.

We geven hieronder een overzicht van de bestanden in deze repository, alsmede de gebruikte onderzoekstools. We volgen hierbij de mappenstructuur van de repository.

1. Dataverzameling

De map 'Dataverzameling' bevat een Excel-bestand waarin de gebruikte zoektermen per onderzoeksthema zijn opgenomen. De keuze voor deze zoektermen wordt verder toegelicht in het onderzoeksrapport. Het PDF-bestand 'Queries Twitter API' bevat de verschillende zoekopdrachten die zijn gebruikt om data op te halen bij de academische API van Twitter. We maken hiervoor gebruik van twarc2, een command line tool voor Python. We halen hiermee de data binnen en zetten het om in een CSV-bestand dat verder kan worden opgeschoond in R.

Voor de dataverzameling van blogs en nieuwssites maken we gebruik van de commerciële data-aggregator OBI4wan. De YouTube-data verzamelen we aan de hand van de YouTube Data Tools van het Digital Methods Initiative (DMI).

2. Data samples

Deze map bevat kleine samples van de verschillende datasets. We treffen hier data zoals deze door OBI4wan wordt geleverd (blogs_sample.csv), de 'ruwe' Twitter-dataset van twarc2 (twitter_raw_sample.csv), de opgeschoonde versie van die dataset (twitter_clean_sample.csv) en de YouTube-dataset van het DMI (youtube_sample.csv). Het doel van deze samples is om inzicht te geven in de structuur van de datasets die voor dit onderzoek zijn gebruikt, zodat de R-scripts (zie volgende map) begrijpelijk zijn.

3. R scripts

In de map 'R scripts' vinden we de scripts die zijn gebruikt in verschillende stadia van het onderzoek, waaronder het opschonen van de datasets, analyses, en de voorbereiding op verdere verwerking in andere tools. Bovenaan elk script is een beknopte omschrijving van zijn functie toegevoegd. Niet alle scripts zijn volledig uit te voeren met de data die in deze repository aanwezig is, omdat ze soms afhankelijk zijn van verdere verwerking in Gephi of Tableau. In de meeste gevallen wordt dit vernoemd als commentaar in het script zelf.

4. Gephi

Tot slot de map 'Gephi'. Hier treffen we een bestand (ce_subset.gephi) dat kan worden geopend in het open-source netwerkanalyseprogramma Gephi. Het bestand bevat netwerkdata van het Twitter-debat dat interactief kan worden geëxploreerd. Online zijn uitstekende Gephi-tutorials te vinden om snel wegwijs te worden in het programma. Ter illustratie is in deze map een van de resulterende netwerkvisualisaties die is gecreëerd op basis van dit netwerk toegevoegd.

Gebruikte software

Voor onze analyses maken we gebruik van verschillende programma's. De belangrijkste zijn:

a. **RStudio** voor het opschonen en verwerken van de datasets. De gebruikte libraries zijn vindbaar in de bijgevoegde scripts.
b. **Gephi** voor het analyseren en verkennen van netwerkdata. We maken hierbij gebruik van de Fieldnotes plugin voor het documenteren van ons proces.
c. **Tableau** voor verdere analyse van de data, bijvoorbeeld frequentieanalyses en tijdlijnvisualisaties, alsmede intuïtieve exploratie van de datasets.

Met de resultaten van de verschillende tools verrijken we continu de oorspronkelijke datasets. De uitgebreide datasets gebruiken we vervolgens om kwalitatieve analyses uit te voeren. Willen we bijvoorbeeld weten waardoor in een bepaalde tijdperiode veel activiteit in het debat plaatsvindt, dan kunnen we deze data in Tableau selecteren en zien we hierbij direct de resultaten van onze netwerkanalyses, geautomatiseerde tekstanalyses etc. Het verkennen van online debatten is vrijwel altijd een exploratief proces, waarin we ons laten leiden door opvallende patronen of uitschieters in de data.