Details on data collection

Asura Enkhbayar & Juan Pablo Alperin, ScholCommLab, Simon Fraser University, Canada
January 26, 2022

The following is an abridged version of the documentation how this dataset was created. Full details, including the source code used to collect the sources, was written by Asura Enkhbayar and can be found here: https://github.com/ScholCommLab/altmetric-news-quality .

The data collection used a combination of RSS and Twitter feeds, depending on what was available from each publication.

Sources

8 publications were selected to be collected through two possible distribution channels: RSS for those that maintain functioning feeds and Twitter for the rest.

Publication	URL	Channel	Details
New York Times – Science	https://www.nytimes.com/section/science	RSS	https://rss.nytimes.com/services/xml/rss/nyt/Science.xml
The Guardian – Science	https://www.theguardian.com/science	RSS	https://www.theguardian.com/science/rss
Wired – Science	https://www.wired.com/category/science/	RSS	https://www.wired.com/feed/category/science/latest/rss
Popular Science	https://www.popsci.com/	Twitter	https://twitter.com/PopSci
IFLScience	https://www.iflscience.com/	Twitter	https://twitter.com/IFLScience
HealthDay	https://consumer.healthday.com	RSS	https://consumer.healthday.com/feeds/feed.rss
News Medical	https://www.news-medical.net/	RSS	http://www.news-medical.net/syndication.axd?format=rss
MedPageToday	https://www.medpagetoday.com	RSS	https://www.medpagetoday.com/rss/headlines.xml

Data processing

Downloading and identifying correct metadata from articles required a combination of meta tags on the publishers pages, custom HTML parsers adjusted to individual sources, and NLP-processing. As with all web scraping, this has some caveats that need to be considered for each field due to the limitation and challenge of comparing various classifications used in different ways by news sources. These include:

- published: It is unclear how “published date” is used by some sources (published vs last-updated).
- modified: Not available for three sources.
- section: Difficult to compare across sources as each source uses their classifications differently. While some are quite comparable (guardian, nyt, wired), in the case of newsmed, the two available sections derive from their particular model of publication (i.e., “Medical News” & “Life Science News”).
- keywords: Again, similar challenges to keywords. However, for some sources the keywords were replaced by tags for a lack of other keywords. Once again, each source might be using tags and keywords differently in their own contexts. ifls and healthday did not provide any keywords or tags, however, using `newspaper` we could derive keywords from the text bodies.

Postprocessing

After downloading and identifying metadata for every article, the collected articles were then filtered. Articles were excluded if they met any of the following exclusion criteria:

- they were published before Mar 1 or after Apr 30
- they are in Spanish
- they have been used in previous samples (that we used for developing the codebook)
- they were not successfully parsed by the *newspaper* python library.