Building Lula x Bolsonaro dataset

by Eduardo Leite

- The news venue was chosen: Folha de S. Paulo, the biggest news vehicle in Brazil;

- The period is chosen: Six months (From 01/05/2022 to 30/10/2022);

- Keyword used to scrape news pieces from the website: Eleições;

- We used FolhaR2[1] R script to scrape every news piece the vehicle published on its website which contained our chosen keyword;

- The script returned, then, a table containing timestamp, section, title, subtitle and URL from every news piece containing “eleições”. There were 4.524 of them;

- We excluded all the opinion pieces from our sample because they’re mostly not accompanied by photographs;

- We then searched through the table for every news piece containing the keywords “Lula” and/or “Bolsonaro”. Every news piece which didn’t cite any of the two candidates was excluded from the sample;

- Finally, we searched through the table for news pieces containing the keyword “Bolsonaro” but didn’t refer to Brazil’s president back then. Some of the pieces cited his sons, who are also involved in politics. Those pieces were also excluded from our sample;

- After refining our news pieces sample, we used Hexomatic[2] automations to scrape images URLs from every news piece on the sample;

---

[1] https://github.com/tomasbarcellos/folhar2
[2] https://hexomatic.com/#login

- Hexomatic’s output was a table containg both the page URL and the images URLs in most of the pieces. The errors were fixed manually;

- Finally, we used Tab save to download all images from the given URLs;

- Our dataset has 1903 photographs, some of them appearing multiple times because the same photographs have been used in different news pieces.