Workshop on the use of data from the SciELO database

Report slides and Python/R notebooks with the performed analyses

Danilo J. S. Bellini

Abstract

Knowledge from data science or contemporary statistics can be used to perform analyses and inferences on large datasets including hundreds of thousands of entries. The exploratory data analysis of a dataset in a research aiming to get information from it might include steps like data acquiring, cleaning, normalization, interpretation, grouping, description and visualization. The goal of this work is to share techniques, methodologies and tools for accessing and exploring data from the SciELO database through its own open access interfaces like SciELO Analytics' reports, SciELO Ratchet, and SciELO ArticleMeta (JSON API and Python software package), as well as from 4 external sources: Web of Science (SciELO Citation Index), Dimensions, SCImagoJR and Scopus. Using either Python (IPython/Jupyter, Numpy, Pandas, Matplotlib, Seaborn, Scipy, NetworkX) or R (R Studio, dplyr) as the programming languages, several analyses had been performed with their open source code included, aiming the reproducibility of the results.

Keywords

Python, R, Data science, Statistics, SciELO, H5, FCR, SJR, Citations, Open access, Open source, Exploratory data analysis

Source code repository

https://github.com/scieloorg/scielo20gt6/