medialab.sciencespo.fr

Software

Keywords:
webmining, scraping, crawling, python

DOI:
10.5281/zenodo.4564399

Software repository:
https://github.com/medialab/minet

Downloads: 340k
Contributors: 16
Versions: 122

*Author for correspondence. Email:
guillaume.plique@sciencespo.fr

Minet, A webmining CLI tool & library for python

Guillaume Plique,* Pauline Breteau, Jules Farjas, Héloïse Théro, Jean Descamps, Amélie Pellé, and Laura Miguel

médialab, SciencesPo

Abstract

minet is a webmining command line tool & library for python that can be used to collect and extract data from a large variety of web sources such as raw webpages, Facebook, CrowdTangle, YouTube, Twitter, Media Cloud etc.

It adopts a very simple approach to various webmining problems by letting its users perform a variety of actions from the comfort of the command line. It does not require any database nor advanced dependencies to function as raw CSV files should be sufficient to do most of the work.

minet has also been designed to be used programmatically as a python library so it can be adapted to most research-oriented use-cases.

Developed as an industrialization of SciencesPo's médialab engineering practices, it has been successfully used by many social science projects and is often used as a pedagogical tool in data science classes.

Since the rise of Internet, social sciences have been interested in studying how people used this new technology for at least the two following reasons:

1. one can find sometimes massive amounts of digital traces left by people online, and those can be leveraged to answer traditional social sciences questions
2. the Internet itself, and by extension the world wide web and social networks, can be thought of as a new field to work on

But to be able to study Internet data, one first needs to be able to collect and analyze it. And to do so properly, one must first understand Internet protocols and technologies enough to wield and retro-engineer them, which is, first and foremost, engineering work. One must therefore learn how to scrape websites, to use online Application Programming Interfaces (APIs), to design web crawlers, etc.

Following this impetus, research engineers from médialab SciencesPo have been doing webmining work for more than a decade now. Based on this experience they often ended up designing tools, intended for a larger audience, so that anybody might have a chance to use them to be able to collect data from the web without requiring advanced technical skills nor access to particular IT resources or people.

minet is one of those tools and quickly became the distillation of the webmining skills of the médialab as a whole. It takes the form of a command line tool one can use to perform complex webmining tasks such as scraping websites, crawling a web corpus, collecting data on social networks or other platforms such as Twitter and YouTube, aggregating urls in a meaningful way, etc. Developed using the Python programming language, it can also be used as an installable library exposing the webmining-related helpers used by the command line tool itself.

What's more, minet has been designed to be fault-tolerant, straightforward to use, low-tech and uses as few computer resources, such as disk space or computing power, as possible so it can be used on virtually any low-cost device or server.

Following the laboratory's philosophy, minet is Open Source and available freely. It has been successfully used in a variety of social sciences projects around the world such as Cointet et al. 2021. It is also frequently used in classrooms as a pedagogical tool, by master's degree or PhD students, by data journalists, by NGOs and also in the private sector.

SciencesPo
MÉDIALAB

Related works

- **gazouilloire**: a longitudinal data collection tool for Twitter
  https://github.com/medialab/gazouilloire
- **ural**: a python library full of url-related heuristics used by minet
  https://github.com/medialab/ural
- **Hyphe**: a web corpus crawler and curation tool for social sciences
  https://github.com/medialab/hyphe
- **4Cat**: a web data collection platform geared towards social sciences
  https://github.com/digitalmethodsinitiative/4cat
- **snscrape**: a scraper for social networking services (SNS)
  https://github.com/JustAnotherArchivist/snscrape
- **Twint**: Twitter intelligence tool
  https://github.com/twintproject/twint

References

Bounegru, Liliana, Jonathan Gray, Tommaso Venturini, and Michele Mauri. 2017. A field guide to fake news: a collection of recipes for those who love to cook with digital methods (chapters 1-3). *Public Data Lab, Research Report.*

Cointet, Jean-Philippe, Pedro Ramaciotti Morales, Dominique Cardon, Caterina Froio, Benjamin Ooghe, and Guillaume Plique. 2021. De quelle (s) couleur (s) sont les gilets jaunes? plonger des posts facebook dans un espace idéologique latent. *Statistique et Société.*

Jacomy, Mathieu, Paul Girard, Benjamin Ooghe-Tabanou, and Tommaso Venturini. 2016. Hyphe, a curation-oriented approach to web crawling for the social sciences. In *Tenth international aaai conference on web and social media.*

Ooghe-Tabanou, Benjamin, Mathieu Jacomy, Paul Girard, and Guillaume Plique. 2018. Hyperlink is not dead! In *Proceedings of the 2nd international conference on web studies,* 12–18.

Peeters, Stijn, and Sal Hagen. 2022. The 4cat capture and analysis toolkit: a modular tool for transparent and traceable social media research. *Computational Communication Research* 4 (2): 571–589.

Rogers, Richard, et al. 2017. Digital methods for cross-platform analysis. *The SAGE handbook of social media,* 91–110.