Empowering social scientists with web mining tools

FOSDEM 2020
Open Research Tools and Technologies Devroom
Guillaume Plique, SciencesPo médialab
Why and how to enable researchers to perform complex web mining tasks?
Guillaume Plique, a.k.a. Yomguithereal

SciencesPo
MÉDIALAB

INVESTIR L’AVENIR
What is web mining?
Crawling
Collecting data from APIs

Search Tweets

Basics
Accounts and users
Tweets

Introduction
The Twitter API platform offers three tiers of search APIs:

- **Standard**: This search API searches against a sampling of recent Tweets published in the past 7 days. Part of the 'public' set of APIs.
- **Premium**: Free and paid access to either the last 30 days of Tweets or access to Tweets from as early as 2006. Built on the reliability and full-fidelity of our enterprise data APIs, provides the opportunity to upgrade your access as your app and business grow.
- **Enterprise**: Paid (and managed) access to either the last 30 days of Tweets or access to Tweets from as early as 2006. Provides full-fidelity data, direct account management support, and dedicated technical support to help with integration strategy.

Feature summary

<table>
<thead>
<tr>
<th>Category</th>
<th>Product name</th>
<th>Supported history</th>
<th>Query capability</th>
<th>Counts endpoint</th>
<th>Data fidelity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Standard Search API</td>
<td>7 days</td>
<td>Standard operators</td>
<td>Not available</td>
<td>Incomplete</td>
</tr>
<tr>
<td>Premium</td>
<td>Search Tweets: 30-day endpoint</td>
<td>30 days</td>
<td>Premium operators</td>
<td>Available</td>
<td>Full</td>
</tr>
<tr>
<td>Premium</td>
<td>Search Tweets: Full-archive endpoint</td>
<td>Tweets from as early as 2006</td>
<td>Premium operators</td>
<td>Available</td>
<td>Full</td>
</tr>
</tbody>
</table>
But why is this useful to [social] sciences?
Bad take

01. Every social sciences data collection is biased (i.e. observer's paradox)
02. People express themselves without being asked to, on the Internet
03. What's more they are not being observed (lol, I know...)
04. Web mining is therefore a superior source of data for social sciences!
01. Internet data comes with its own biases that you should be aware of
02. Apply media studies and STS without moderation
03. Still is another very interesting and large data source!
Web mining is hard

You need to know The Web™:

DNS  HTTP  HTML  CSS  JS  DOM  AJAX  SSR  CSR  XPATH  ...
How do you teach researchers web technologies

01. The same as anyone else really (CSS as sushi plates anyone?)
02. What most consider as an easy layer of technologies really ISN'T
03. We really are standing on the shoulders of giants
Teaching researchers how to scrape

01. Fighting the platforms and their APIs
02. Legal issues in some countries
03. Sometimes forbidden to teach it (~lock picking)
04. Publication wiggles (the monkey army)
Jupyterizing researchers is not a solution

01. Some researchers don't have the time nor the will to learn python and web stuff.
02. We should be OK with that!
Web mining is **HARD**

It really is a craftsmanship.
Internet is a dirty, dirty place

Browsers truly are heuristical wonders!
Multithreading, parallelization, throttling etc.

Once we cut access to Google to our whole university!
Complex spidering, scalability, storage, indexing, recombobulation, steam engines, fancy boats, unionization, agility, upper management, Peters syndrom, eXtreme programming

Most of it is irrelevant and made up but you get the point...
How do we empower researchers then?

By designing tools suited to their research questions
SciencesPo's médialab

01. Social Science Researchers
02. Designers
03. Engineers
A brief guided tour of tools we designed

01. artoo.js
02. minet
03. Hyphe
04. (Gazouilloire)
Parasitizing web browsers instead of emulating them!

Demo Time!
Leveraging bookmarklets to empower researchers

arto.js The client-side scraping companion.

Bookmarklet Generator

This handy bookmarklet generator lets you create a custom standalone artoo.js bookmarklet using the provided snippet of code:

Your bookmarklet name

Paste your code here...

Generate
But can we scale up?

Not-contractual logo - Jules Farjas ©
Handling the pesky details for you

01. Multithreaded, memory-efficient **fetching** from the web.
02. Multithreaded, scalable **crawling** using a comfy DSL.
03. Multiprocessed raw text **content extraction** from HTML pages.
04. Multiprocessed **scraping** from HTML pages using a comfy DSL.
05. **URL-related heuristics** utilities such as normalization and matching.
06. Data collection from various **APIs** such as CrowdTangle.
The Unix philosophy

Do one thing well

xsv search -s url urls.csv | minet fetch url -d html > result.txt
The low-fi approach
Relocalizing data collection

01. Sometimes you don't need a server
02. We are rarely doing BigData™
03. Let's put the researcher at the center so they can control their data
A programmatic API

Jupyter's back y'all!

```python
from minet import multithreaded_fetch

for result in multithreaded_fetch(urls_iterator):
    print(result.status)
```
How to enable researchers to crawl the Web?
HYPHE

Hyphe is a web corpus curation tool featuring a research-driven web crawler
A dedicated interface
Serving a robust methodology

- The node has been crawled (IN)
- Another node has been crawled
- You can discard nodes manually
- Add and crawl nodes manually
Non-trivial technical challenges
Trade-off between scalability & usability

We need to be able to **design** user paths.
The future!

What about a GUI for minet?
Thank you for listening!