Datasets
- Media reports

Studying digital media in politics: Methods & Approaches

The slides + all examples are available at [*github.com/MisterXY89/polAna*](https://github.com/MisterXY89/polAna)

Different types of sources

Original/direct sources

APIs from the outlet itself, they often come with internal data other APIs/3rd party tools can not deliver.

- New York Times API
- Die ZEIT API

News-Crawler

Crawler can be used for a collection of sites. They can parse articles and often follow links on the site

- news-please
- mercury

3rd party archives

APIs/archives from 3rd parties possibly containing articles from many sources, e.g.:

- News API
- LexiNexis
- MediaCloud

Original/direct sources

Advantages

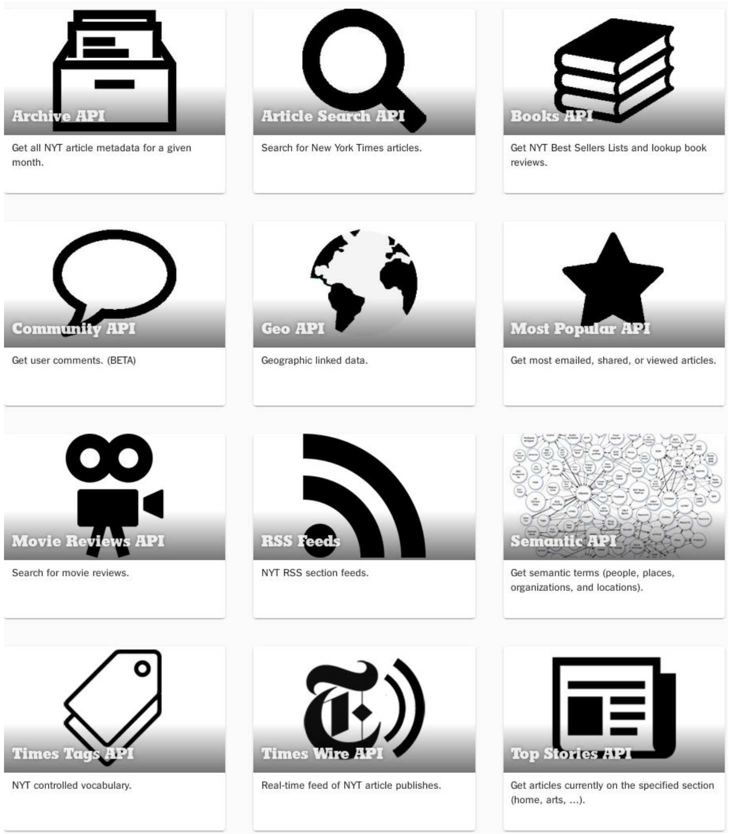
- Reliability
- Possibly more details
- Internal data

Disadvantages

- Only one source with one tool

New York Times API

- Getting started: developer.nytimes.com
- Well documented, but no API wrapper
- What kind of data can I get?
  Example: developer.nytimes.com/docs/top-stories-product



Archive API
Get all NYT article metadata for a given month.

Article Search API
Search for New York Times articles.

Books API
Get NYT Best Sellers Lists and lookup book reviews.

Community API
Get user comments. (BETA)

Geo API
Geographic linked data.

Most Popular API
Get most emailed, shared, or viewed articles.

Movie Reviews API
Search for movie reviews.

RSS Feeds
NYT RSS section feeds.

Semantic API
Get semantic terms (people, places, organizations, and locations).

Times Tags API
NYT controlled vocabulary.

Times Wire API
Real-time feed of NYT article publishes.

Top Stories API
Get articles currently on the specified section (home, arts, ...).

https://developer.nytimes.com/apis

How to access the data?

- The NYT itself does not provide any client
- NYTimesArticleAPInew “is a is a fully-functional Python wrapper for the New York Times Article Search API.”
- For the other features, such as the top story, you have to write your own wrapper or search for existing ones

News-Crawler

Advantages

- Data of multiple sources, defined by the user
- Recursive search
- Regular updates / change tracking

Disadvantages

- Data completeness
- No internal data

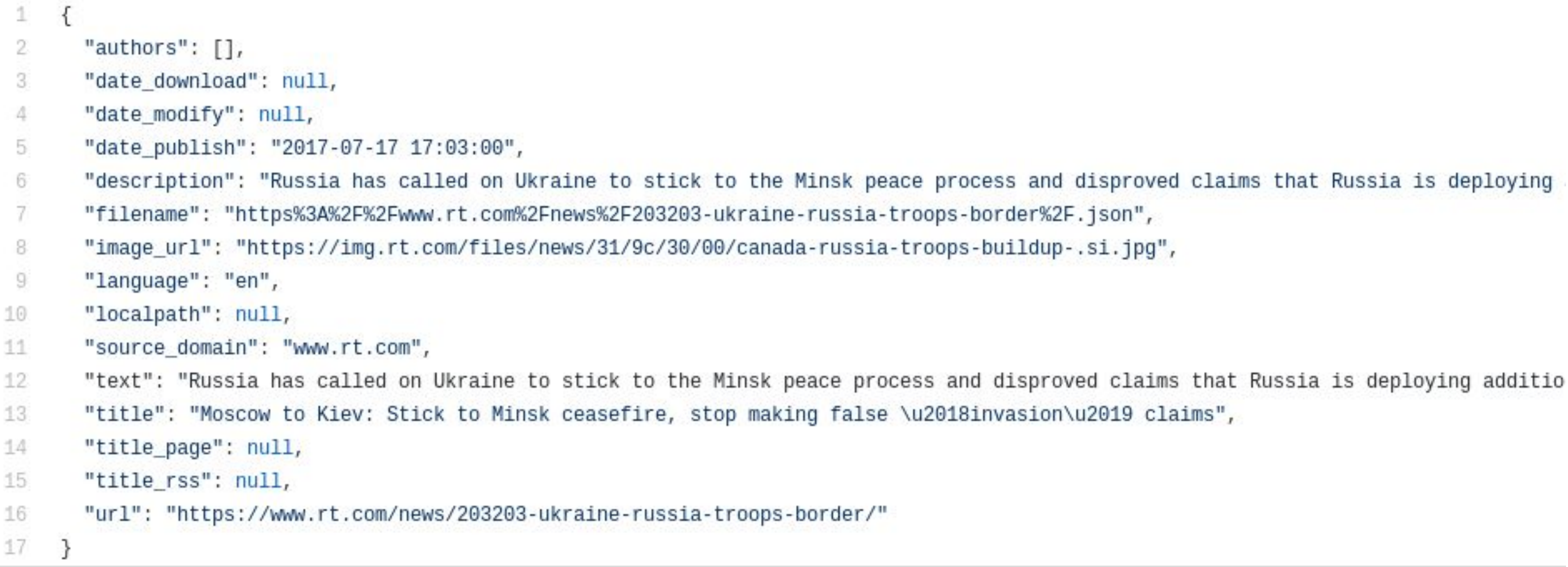
News-Please

“news-please is an open source, easy-to-use news crawler that extracts structured information from almost any news website”

- is build for and with python
- has 3 modes
  1. Library
  2. CLI (command line)
  3. News archive from commoncrawl.org (free-to-use archive of news articles)

https://github.com/fhamborg/news-please/

News-Please



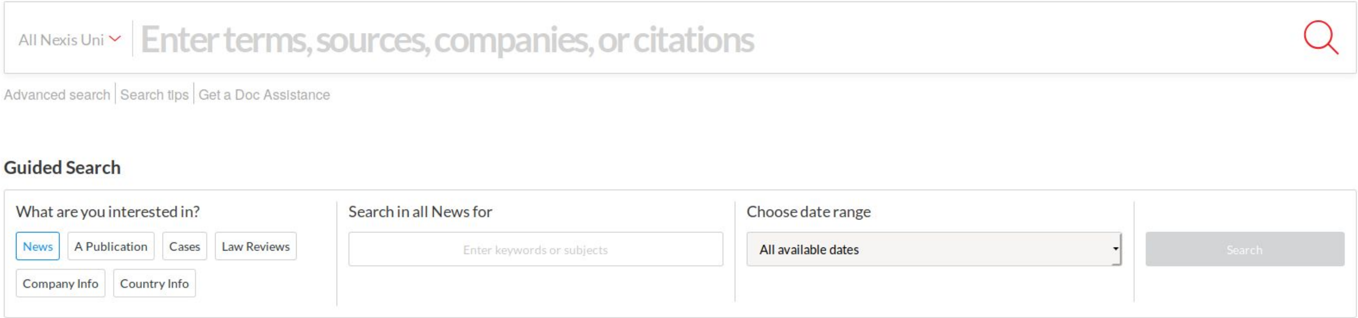
```
{
  "authors": [],
  "date_download": null,
  "date_modify": null,
  "date_publish": "2017-07-17 17:03:00",
  "description": "Russia has called on Ukraine to stick to the Minsk peace process and disproved claims that Russia is deploying
  "filename": "https%3A%2F%2Fwww.rt.com%2Fnews%2F203203-ukraine-russia-troops-border%2F.json",
  "image_url": "https://img.rt.com/files/news/31/9c/30/00/canada-russia-troops-buildup-.si.jpg",
  "language": "en",
  "localpath": null,
  "source_domain": "www.rt.com",
  "text": "Russia has called on Ukraine to stick to the Minsk peace process and disproved claims that Russia is deploying additio
  "title": "Moscow to Kiev: Stick to Minsk ceasefire, stop making false \u2018invasion\u2019 claims",
  "title_page": null,
  "title_rss": null,
  "url": "https://www.rt.com/news/203203-ukraine-russia-troops-border/"
}
```

https://github.com/fhamborg/news-please/blob/master/newsplease/examples/sample.json

3rd party APIs/tools

Advantages	Disadvantages
● Data of multiple sources combined	● Data completeness
● Regular updates / change tracking	● No internal data

Lexis Nexis



https://advance.lexis.com/bisacademicresearchhome

- Extensive research tool for lots of sources
- Visit nexisuni.com

- Your device has to be connected to the network of the university
  1. Connect your device to the eduroam network
  2. Use the vpn of the university

MediaCloud

“Media Cloud is an open-source platform for media analysis.”

Explorer	**Topic Mapper**	**Source Manager**
● create an instant analysis of how digital news media covers your topic ● you can see attention to the issue, the language used, and the people and places mentioned	● everything from Explorer ● plus data about media ecosystems, social media shares, and influence networks	● View print, broadcast and digital news collections ● see what media sources we have and suggest additional sources

https://mediacloud.org/#/home

How to access the data?

- Via the User Interface on https://mediacloud.org/
- Great visualisation, possibility to export the data as csv

- Via API (docs)
- The Python client

Example

https://explorer.mediacloud.org/#/queries/search?qs=%5B%7B%22label%22%3A%22FC%20Bayern%22%2C%22q%22%3A%22FC%20Bayern%22%2C%22color%22%3A%22%231f77b4%22%2C%22startDate%22%3A%222019-01-01%22%2C%22endDate%22%3A%222019-12-09%22%2C%22sources%22%3A%5B%5D%2C%22collections%22%3A%5B38379825%5D%2C%22searches%22%3A%5B%5D%7D%2C%7B%22label%22%3A%22FC%20Bayern%22%2C%22q%22%3A%22FC%20Bayern%22%2C%22color%22%3A%22%23ff7f0e%22%2C%22startDate%22%3A%222019-01-01%22%2C%22endDate%22%3A%222019-12-09%22%2C%22sources%22%3A%5B%5D%2C%22collections%22%3A%5B38379817%5D%2C%22searches%22%3A%5B%5D%7D%5D

News API

What search parameters you can use?

- **Keyword or phrase.**
  find all articles containing a specific word
- **Date published.** Eg: yesterday.
- **Source name.** Eg: all articles by 'TechCrunch'
- **Source domain name.** Eg: find all articles published on nytimes.com.
- **Language.** Eg: find all articles written in English

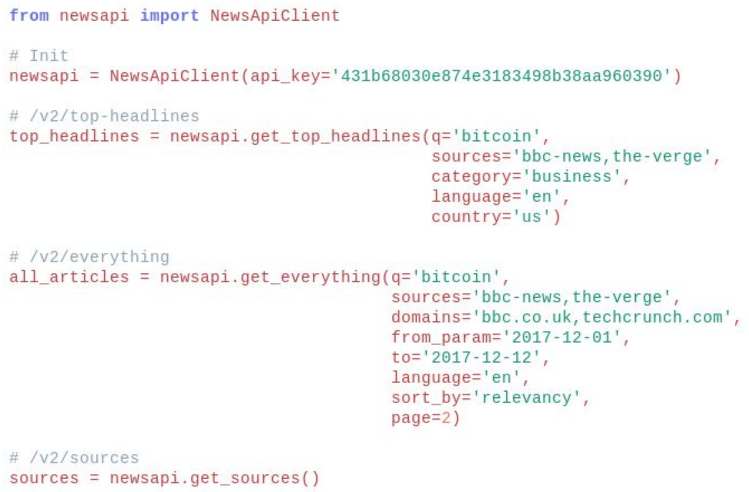
https://newsapi.org/docs

How to access the data?

Best way is to use the client libraries, available for:

- Python
- Ruby
- C#
- Node

But the documentation is not 100% correct



```python
from newsapi import NewsApiClient

# Init
newsapi = NewsApiClient(api_key='431b68030e874e3183498b38aa960390')

# /v2/top-headlines
top_headlines = newsapi.get_top_headlines(q='bitcoin',
                                          sources='bbc-news,the-verge',
                                          category='business',
                                          language='en',
                                          country='us')

# /v2/everything
all_articles = newsapi.get_everything(q='bitcoin',
                                      sources='bbc-news,the-verge',
                                      domains='bbc.co.uk,techcrunch.com',
                                      from_param='2017-12-01',
                                      to='2017-12-12',
                                      language='en',
                                      sort_by='relevancy',
                                      page=2)

# /v2/sources
sources = newsapi.get_sources()
```

https://newsapi.org/docs/client-libraries/python

Is there a best tool?