Authors

Jérôme Dockès, Kendra Oudyk, Jean-Baptiste Poline.

Title

From PubMed to a DataFrame: tools & resources for streamlining neuroimaging text-mining.

Body

Intro

With thousands of publications each year, the neuroimaging literature is a rich but challenging resource. To exploit it efficiently, systematic or (semi-)automated approaches for indexing, aggregating and summarising articles are necessary. Examples of literature analysis include large-scale meta-analyses (e.g. NeuroSynth, NeuroQuery), topic modelling, and more.

An important challenge to text-mining projects is the construction of an appropriate dataset (i.e. a list of articles). One must (automatically) download a large number of articles and extract the relevant text, metadata and often the stereotactic coordinates of neuroimaging results. Due to this difficulty, most projects rely on NeuroSynth or NeuroQuery infrastructures to obtain articles' data. This drastically limits possibilities: the text itself is not available (only text-frequency features or abstracts), recent or missing articles cannot be added, etc.

Here we introduce here two inter-operable tools that help collecting and labelling articles:

- pubget downloads and processes articles from PubMed Central,
- labelbuddy is a simple text labelling application.

With these tools, neuroimagers should be able to skip the tedious data collection step and jump to performing high-level analyses using familiar scientific software on a rich and user-friendly dataset.

Methods

Fig. 1 shows the stages of a text-mining project. The high-level analysis (in green) is the main objective – for example, plotting the evolution of neuroimaging study sample sizes through time and can be simply performed once the data are available in an appropriate format. However, previous steps (obtaining text, labelling sample sizes) will generally take a long time, even for a researcher with strong expertise.

pubget

pubget (https://neuroquery.github.io/pubget/) is a command-line tool for downloading and processing articles from PubMed Central. It builds upon the code used to create NeuroQuery. Given a search query or a list of PMCIDs, it provides the matching articles in their original XML format, in addition to CSV files containing: (i) metadata such as authors or publication year, (ii) the full text, and (iii) the activation coordinates. pubget can extract term-frequency features, and run NeuroQuery's or NeuroSynth's analyses. It can prepare a NiMARE (nimare.readthedocs.io) dataset, making a wide range of meta-analysis methods easy to apply. It can be extended with plugins.

labelbuddy

Most tools for text labelling - essential to establish ground truth in machine learning projects - are Web-based and incur an important set-up overhead for small research projects. labelbuddy is a simple and lightweight desktop application that manage annotations with a regular file (a SQLite database). pubget's output can directly be imported into labelbuddy. labelbuddy imports and exports its data to a simple JSON format, and offers a command-line interface, making it well-suited for projects organised around a Git repository. An example repository containing over 1,800 annotations can be found at https://neurodatascience.github.io/labelbuddy-annotations/.

Results

To illustrate the use of this ecosystem, we replicated and extended the investigation of sample sizes from Poldrack & al, "Scanning the horizon". We downloaded articles with pubget, designed a heuristic to extract participant counts and demographics, and validated it on 100 articles that we annotated with labelbuddy. As shown in Fig. 1.A, the median sample size continued to increase since 2015. We also show the distribution of participant's ages. The mode for healthy participants is between 20 and 30 years, suggesting that the declaimed practice of recruiting mostly students may be widely followed.

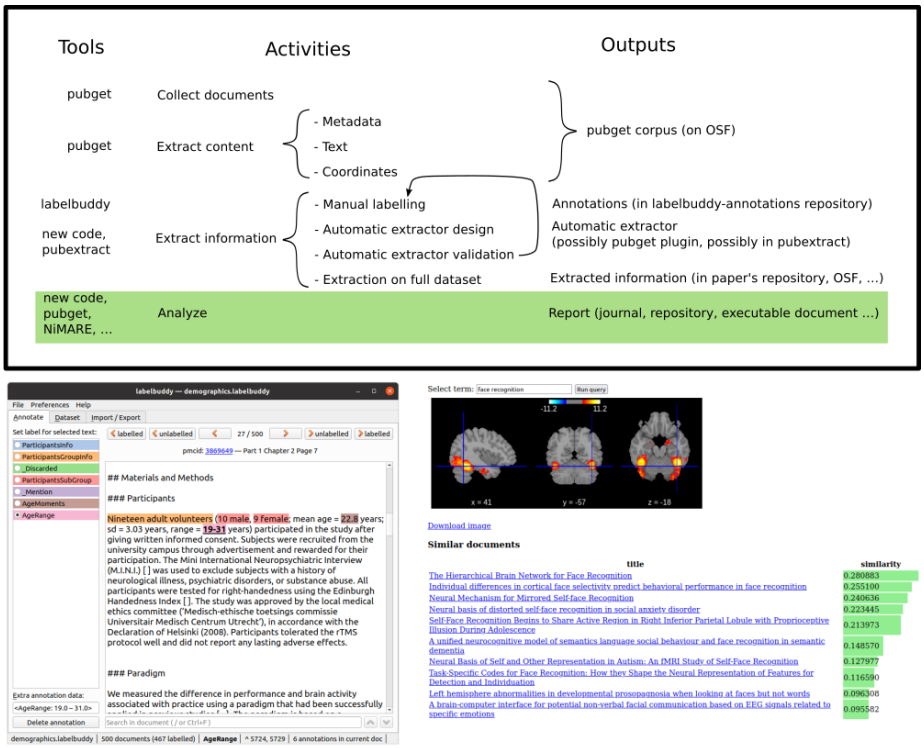
We also ran pubget for a query matching a larger number of articles (over 9K). In Fig 1.B, we show meta-analytic maps obtained with pubget's –fit_neurosynth option (top) and from neurosynth.org (bottom). Results are similar for frequent terms, but for rare terms, pubget's use of the full text produces more powerful analyses.

Conclusions

We facilitate downloading, annotating and preparing articles for analysis. This can be key for the understanding of an ever increasing body of literature. This eco-system of tools is likely to make feasible projects that would otherwise have required too much resources, from a small quantitative section in a review or

meta-science paper, to the development of new large-scale meta-analysis methods. As the tools are still evolving, we hope that discussions at the OHBM 2023 meeting will help us tailor them to the needs of the neuroimaging community.

Figures

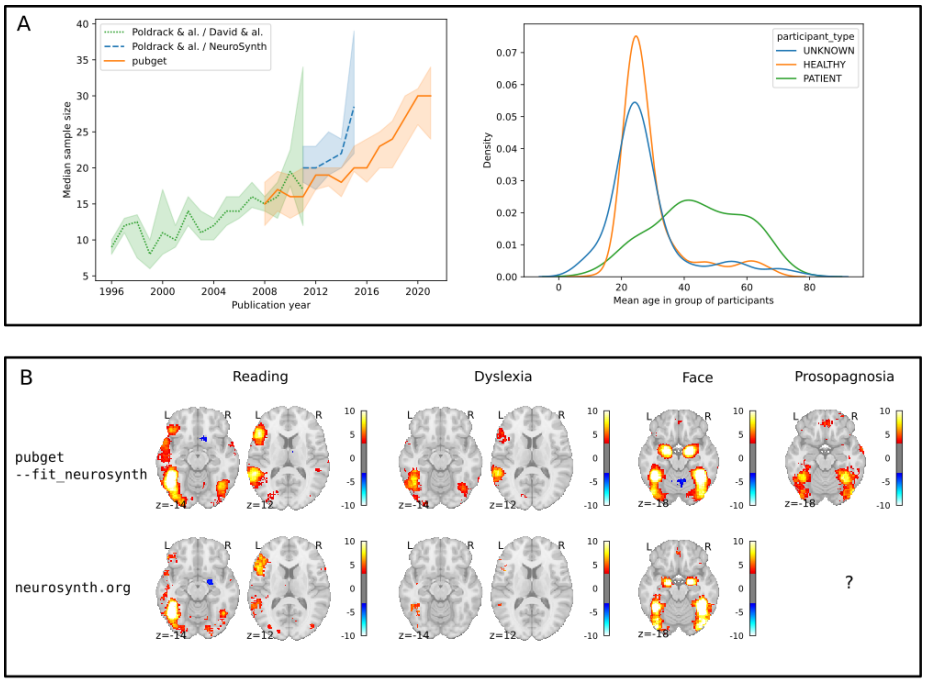


Top: overview of a text-mining project. The center column shows the main steps, from data collection to the final analysis. The left column shows tools useful for the corresponding step. "new code" is written for a particular project. pubextract will be a Python package in which we plan to collect information extraction modules for specific aspects such as the number of participants. The right column shows the useful outputs produced by the project, and in parentheses suggestions of where they can be stored online.Bottom left: screenshot of labelbuddy annotating an article downloaded with pubget. Bottom right: screenshot of a small local web application displaying the results of a NeuroSynth-like meta-analysis, that is part of pubget's output when the "--fit_neurosynth" option is used.

Figure 1: Methods

Top: overview of a text-mining project. The center column shows the main steps, from data collection to the final analysis. The left column shows tools useful for the corresponding step. “new code” is written for a particular project. pubextract will be a Python package in which we plan to collect information extraction modules for specific aspects such as the number of participants. The right column shows the useful outputs produced by the project, and in parentheses suggestions of where they can be stored online.

Bottom left: screenshot of labelbuddy annotating an article downloaded with pubget. Bottom right: screenshot of a small local web application displaying the results of a NeuroSynth-like meta-analysis, that is part of pubget's output when the "–fit_neurosynth" option is used.



Top left: evolution of the median sample size through time. The green and blue curves represent data from Poldrack & al. The orange curve represents sample sizes automatically extracted from articles obtained with pubget and the query "fMRI[Abstract]". We notice that the blue curve seems slightly higer. These data were obtained from the NeuroSynth dataset, and thus may be slightly biased because it only contains abstracts. Indeed, we noted that articles with larger sample sizes are more likely to report the sample size in their abstract.

Top right: distribution of (automatically extracted) participant ages by category in the articles downloaded with pubget. We note that the healthy populations display a peak around the typical age of university students.

Bottom: maps obtained with NeuroSynth's independence test, as obtained with pubget and the --fit_neurosynth option (first row) and from neurosynth.org (second row). For frequent terms such as "reading", results are similar, but for less frequent terms the increased statistical power afforded by the full-text becomes important. "Prosopagnosia" is absent from neurosynth.org's vocabulary (too rare).

Figure 2: Results

Top left: evolution of the median sample size through time. The green and blue curves represent data from Poldrack & al. The orange curve represents sample sizes automatically extracted from articles obtained with pubget and the query "fMRI[Abstract]". We notice that the blue curve seems slightly higher. These data were obtained from the NeuroSynth dataset, and thus may be slightly biased because it only contains abstracts. Indeed, we noted that articles with larger sample sizes are more likely to report the sample size in their abstract.

Top right: distribution of (automatically extracted) participant ages by category

in the articles downloaded with pubget. We note that the healthy populations display a peak around the typical age of university students.

Bottom: maps obtained with NeuroSynth's independence test, as obtained with pubget and the –fit_neurosynth option (first row) and from neurosynth.org (second row). For frequent terms such as "reading", results are similar, but for less frequent terms the increased statistical power afforded by the full-text becomes important. "Prosopagnosia" is absent from neurosynth.org's vocabulary (too rare).