IDA

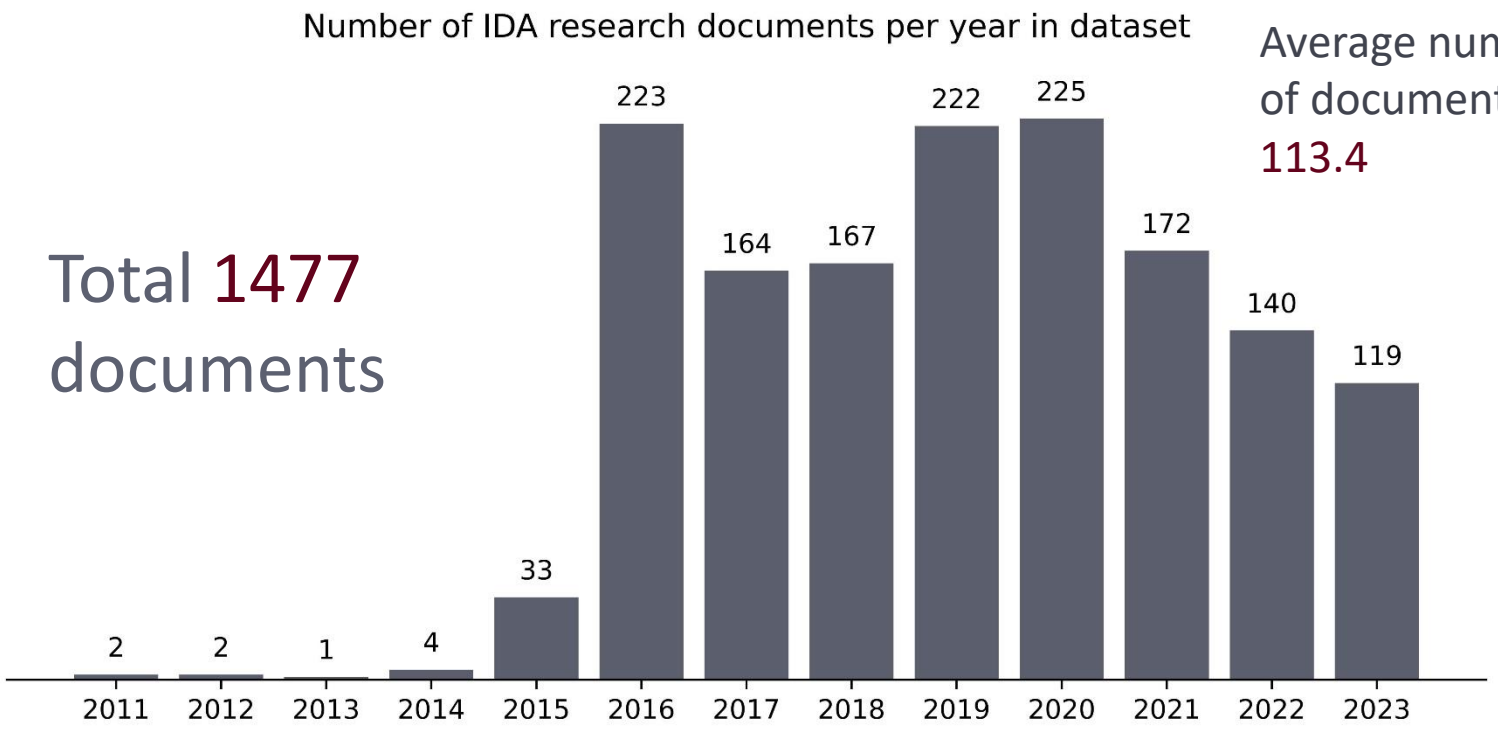
From Text to Metadata: Automated Product Tagging with Python and NLP

Aayushi Verma & Omar Agha Khan
Data Science Fellows

April 17, 2024
DATAWorks 2024

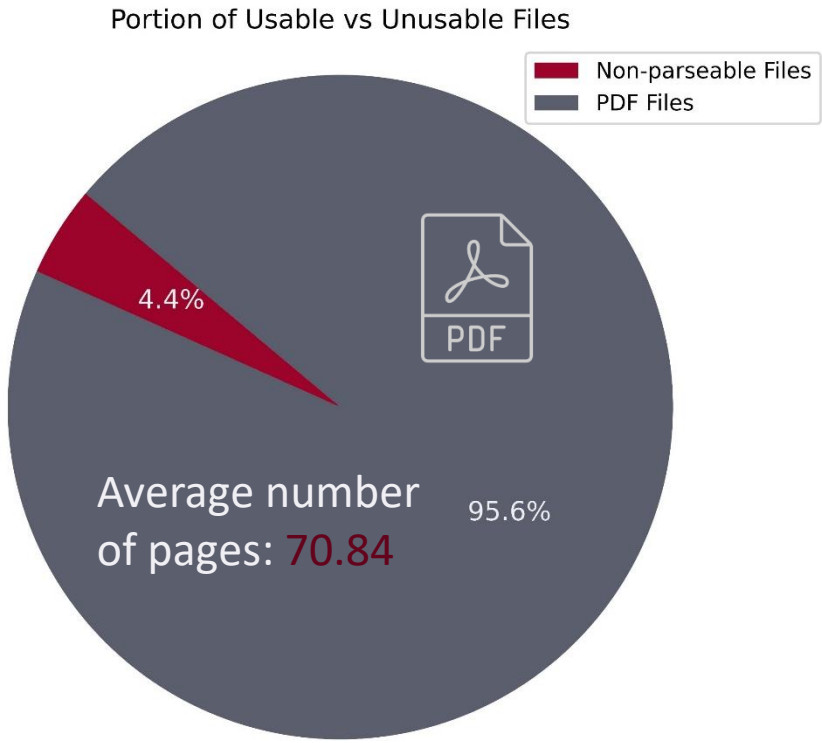
Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

IDA is a research institution and produces many research reports.



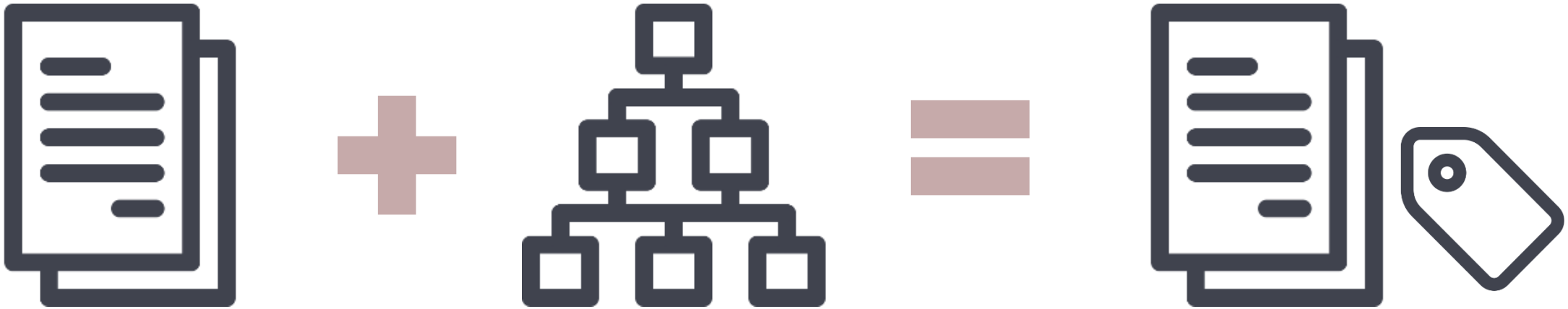
Total 1477 documents

Average number of documents: 113.4

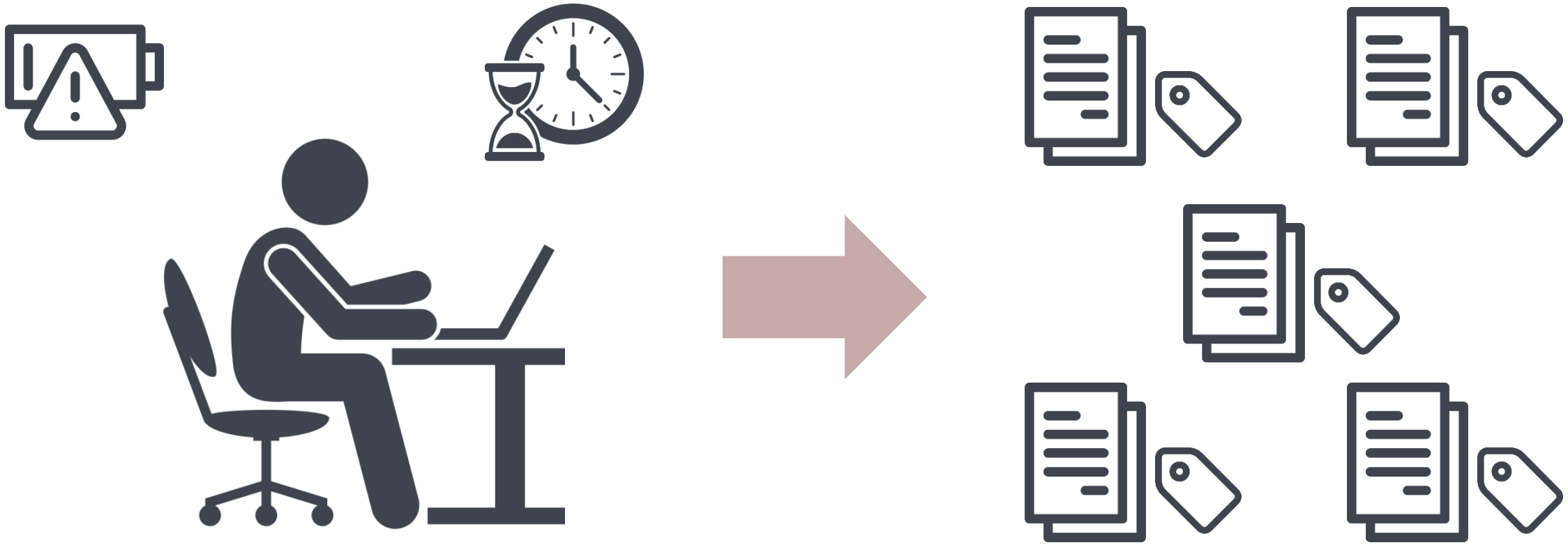


IDA has developed its own set of taxonomies to describe our research.

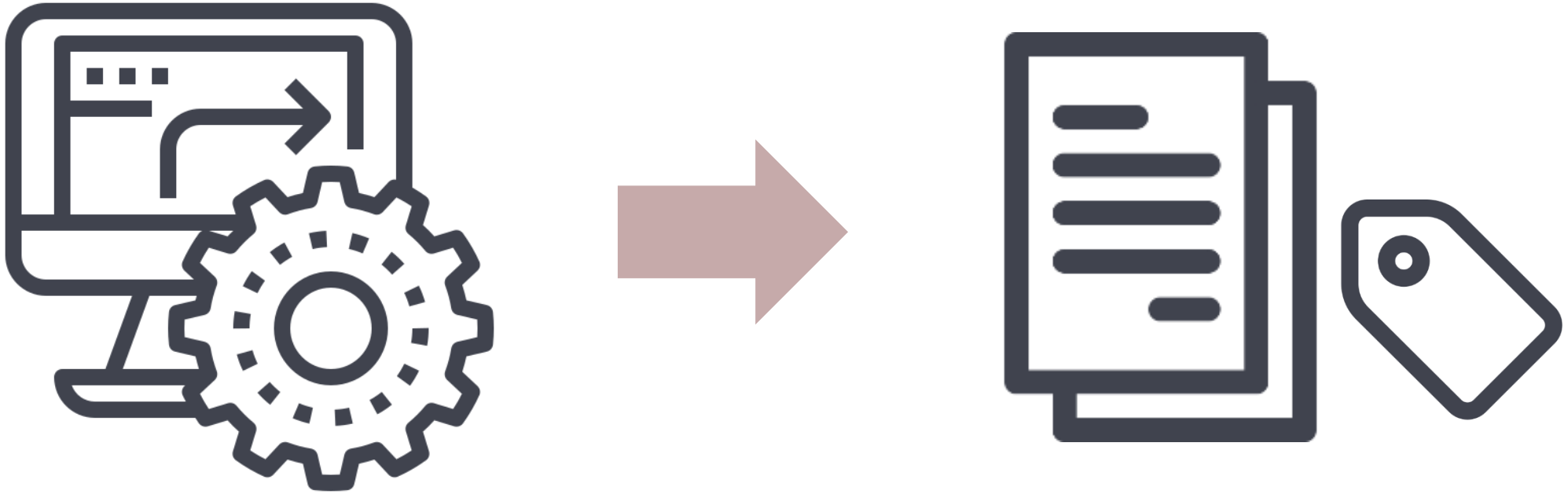
How can we quickly tag our research documents with taxonomy terms?



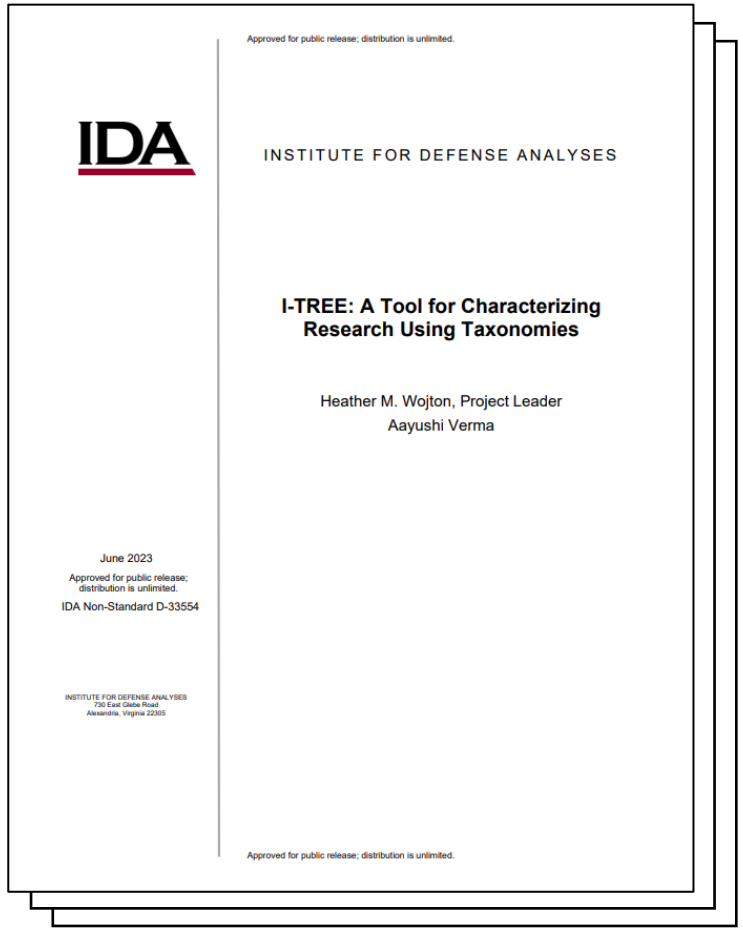
IDA researchers currently have to tag their documents themselves.



We developed a capability to automatically tag these documents with taxonomy terms using NLP!

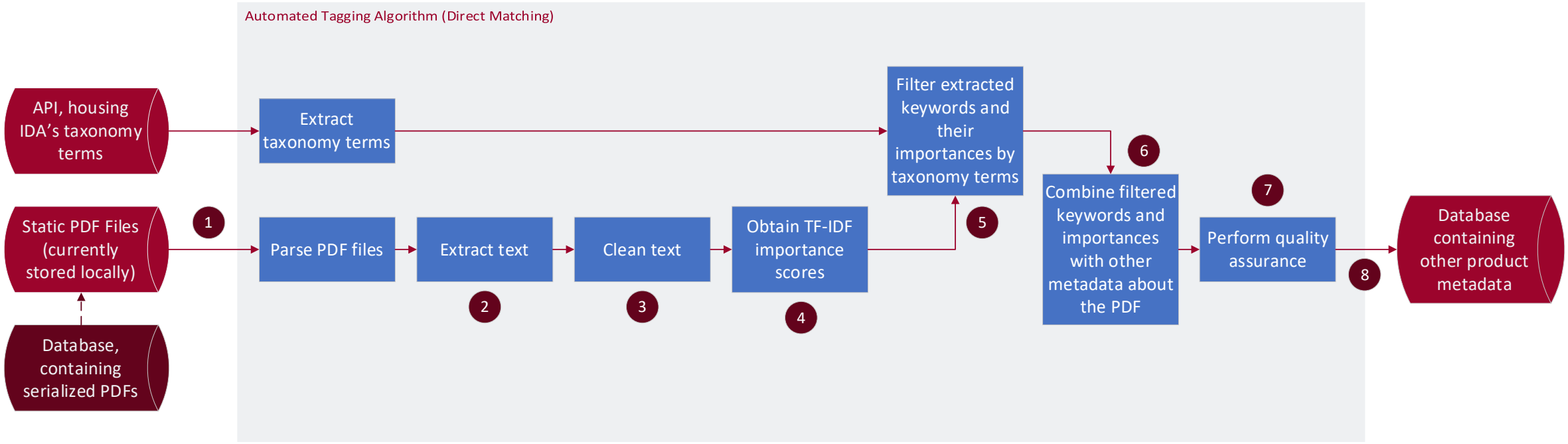


This is a visualization of what we have achieved!



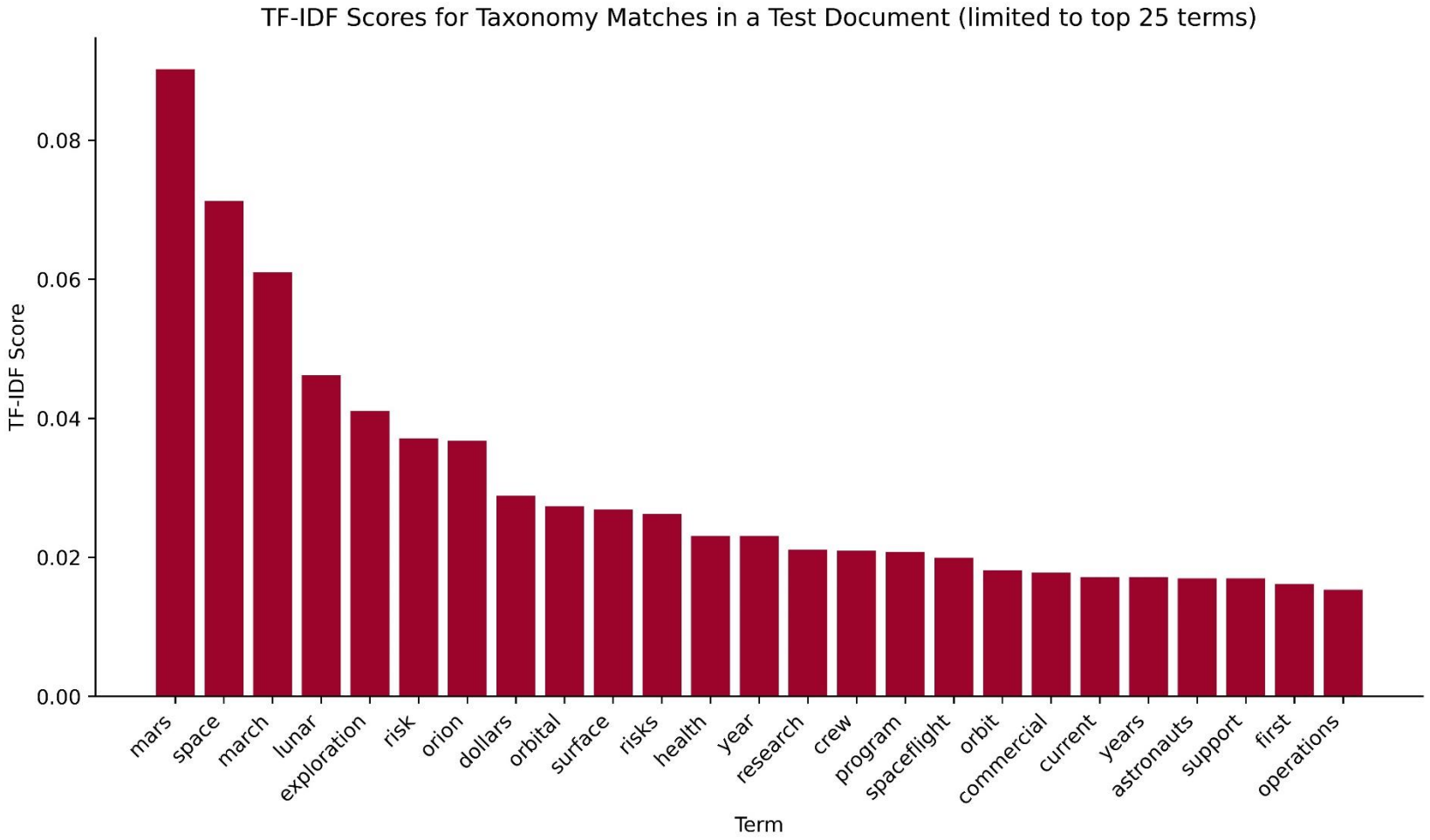
data management, data science, R, data visualization, interactive plots, quantitative methods, analytical methods, data analysis software, software tools, RStudio, data management maturity, information management, science and technology, methods for encouraging innovation

This is the pipeline we implemented!



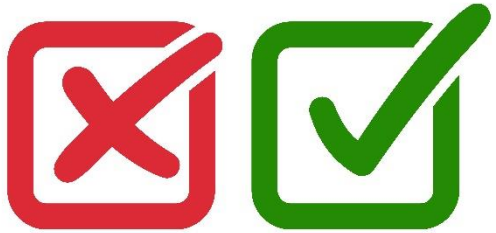
TF-IDF: Term Frequency-Inverse Document Frequency. A metric for determining how important a word is, relative to its frequency in this document and other documents.

Here's an example of TF-IDF scores for taxonomy terms in a test document, showing the 25 most relevant terms in the document.



We're also building quality assurance processes.

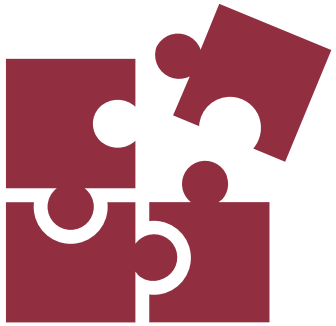
- Extracting keywords is easy!
- Need to consider more serious quality checks
  - Does the set of tags accurately describe the body of work?
  - Will a human tag this product with the same or similar terms?
  - Is the tagging process consistent across distinctive research?
  - How do we consider context of sentences, paragraphs, areas of research?
  - How do we deal with ambiguous terms like 'Space', 'Land', etc.?
- Aim is to present 'base' set of terms for a product that a human can verify



Why do we care about tagging products, anyway?

- Use document keywords as **metadata**
- Combine with other metadata to quantify research

- This project is an initiative as part of IDA's **Data Strategy**
  - Building numerous internal data-centric capabilities and systems for a resilient data infrastructure through **data management** and **governance** *(Verma, 2023)*.



References

- Verma, A. (2023). “I-TREE: A Tool for Characterizing Research Using Taxonomies.” *International Journal for Test & Evaluation*.
- Artifex Software, Inc. (2024). PyMuPDF (Version 1.23.26) [Computer software]. Available at https://github.com/pymupdf/PyMuPDF
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. *O'Reilly Media, Inc*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, 12, 2825–2830.
- Icons made by Freepik from www.flaticon.com.

Thank you!
Questions?

averma@ida.org

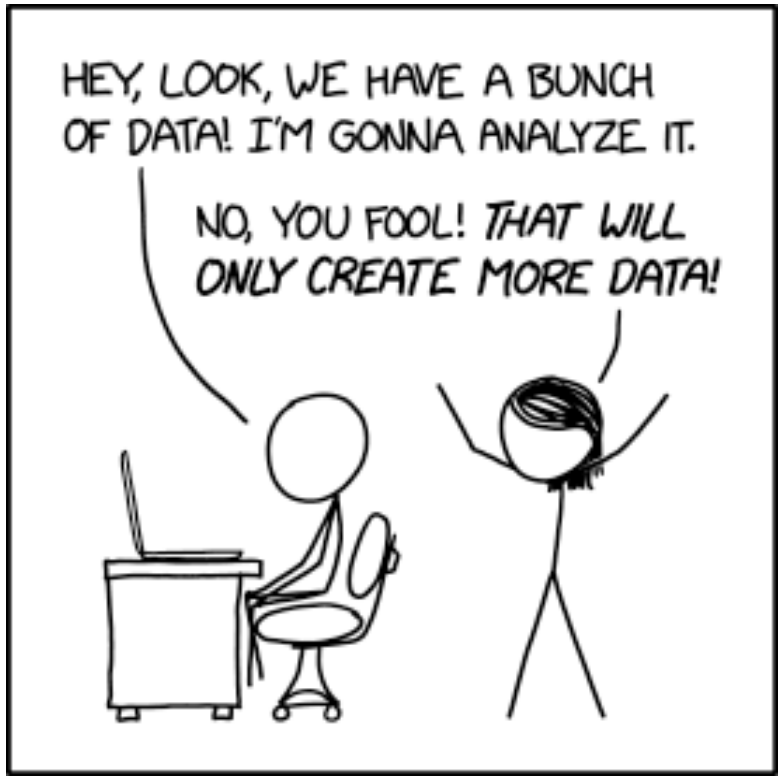
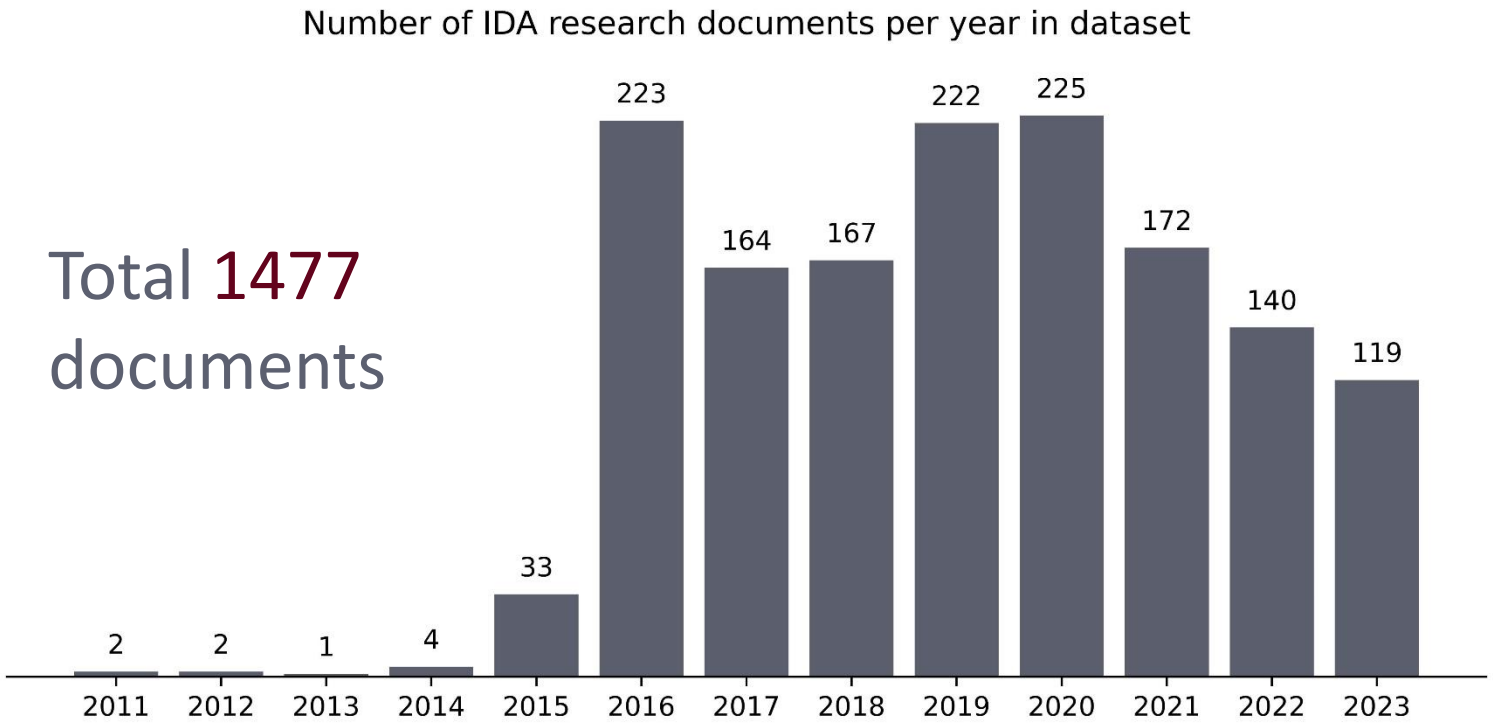


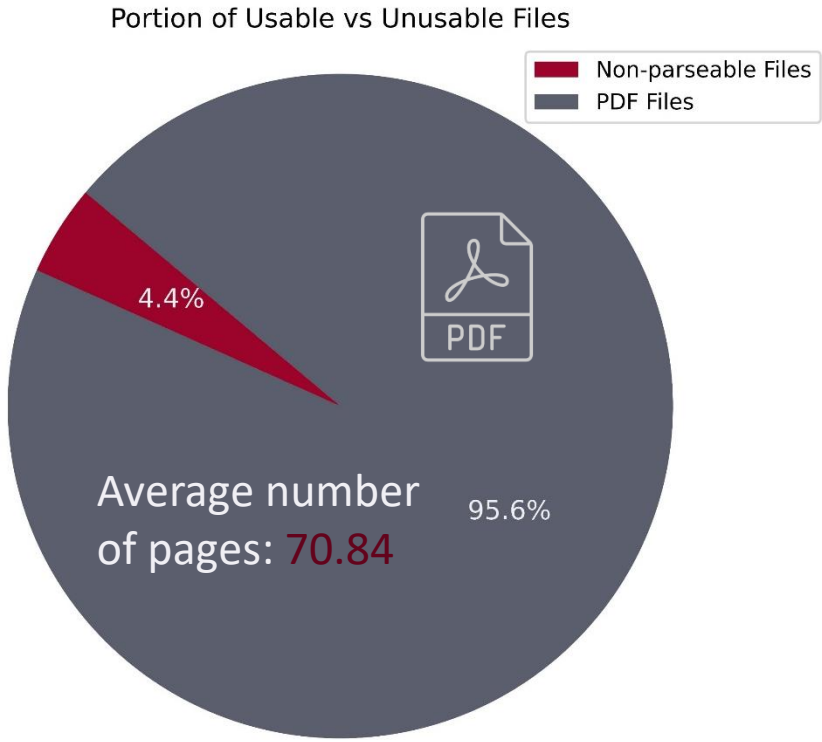
Image credit: https://xkcd.com/2582/

Backup Slides

We obtained a dataset of real IDA research documents.

Total 1477 documents





Case study: we made an interactive app which uses the tagged product data!

Scan to read more!

IDA-TREE | Authors for Tag | Tags for Author | Tags for Division | Authors for Division | Divisions for Tag | Taxonomy Browser

Who are the researchers most commonly producing formal products for the research topic Test and Evaluation and its narrower terms?

Taxonomy

Show Taxonomies

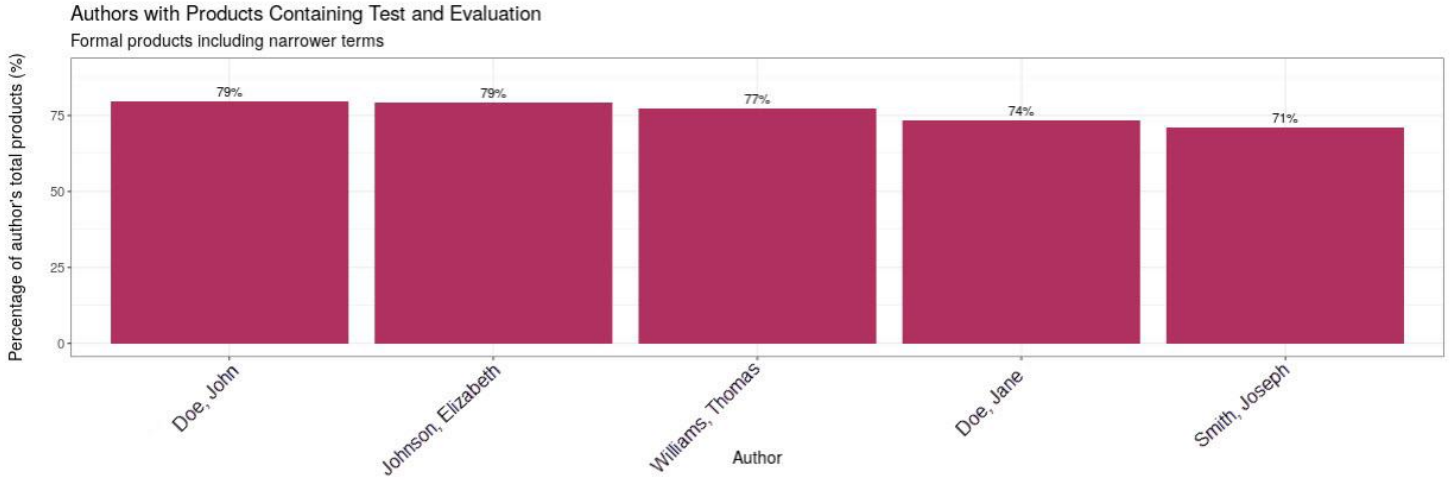
All

Click to select, expand, or collapse taxonomy labels.

- Analytical Methods
- Domain
- Mission Areas
  - Manpower and Personnel
  - Intelligence
  - Operations
  - Protection
  - Planning
  - Communications & Information
  - Sustainment & Logistics
  - Acquisition
    - Science and Technology
    - Research and Development
    - Test and Evaluation
    - Program Management and Strategy
    - Threat Emulation
- Military and Policy
- Science and Technology
- Weapon Systems

Select range of taxonomy terms to include:

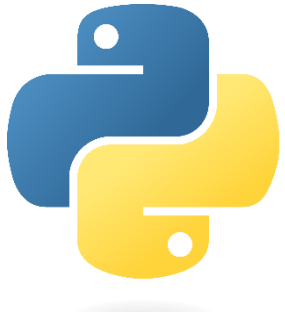
- (•) Narrower terms
- ( ) Only selected term



Show 10 entries Search:

	Author	Total Products with Selected Term for Author	Total Products for Author	% of Author's Products
1	Doe, John	31	39	79.49
2	Johnson, Elizabeth	38	48	79.17
3	Williams, Thomas	27	35	77.14
4	Doe, Jane	25	34	73.53
5	Smith, Joseph	22	31	70.97

Showing 1 to 5 of 5 entries Previous 1 Next




NLTK


Microsoft®
SQL Server®


scikit
learn


PyMuPDF