Tackling the US Government's PDF Mountain With Computer Vision

Originally published at https://www.unite.ai/tackling-the-us-governments-pdf-mountain-with-computer-vision/

Adobe's PDF format has entrenched itself so deeply in US government document pipelines that the number of state-issued documents currently in existence is conservatively estimated to be in the hundreds of millions. Often opaque and lacking metadata, these PDFs – many created by automated systems – collectively tell no stories or sagas; if you don't know exactly what you're looking for, you'll probably never find a pertinent document. And if you did know, you probably didn't need the search. However a new project is using computer vision and other machine learning approaches to change this almost unapproachable mountain of data into a valuable and explorable resource for researchers, historians, journalists and scholars.

When the US government discovered Adobe's ADBE 0.82% Portable Document Format (PDF) in the 1990s, it decided that it liked it. Unlike editable Word documents, PDFs could be 'baked' in a variety of ways that made them difficult or even impossible to amend later; fonts could be embedded, ensuring cross-platform compatibility; and printing, copying and even opening could all be controlled on a granular basis.

More importantly, these core features were available in some of the oldest 'baseline' specifications of the format, promising that archival material would not need to be reprocessed or revisited later to ensure accessibility. Nearly everything that government publishing needed was in place by 1996.

With blockchain provenance and NFT technologies decades away, the PDF was as near as the emergent digital age could get to a 'dead' analogue document, only a conceptual hiccup away from a fax. This was exactly what was wanted.

Internal Dissent About PDF

The extent to which PDFs are hermetic, intractable, and 'non-social' is characterized in the documentation on the format at the Library of Congress, which favors PDF as its 'preferred format':

> *'The primary purpose for the PDF/A format is to represent electronic documents in a manner that preserves their static visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files. To this end, PDF/A attempts to maximize device independence, self-containment, and self-documentation.'*

Ongoing enthusiasm for the PDF format, standards for accessibility, and requirements for a minimum version, all vary across US government departments. For instance, while the Environmental Protection Agency has stringent but supportive policies in this regard, the official US government website plainlanguage.gov acknowledges that *'users hate PDF'*, and even links directly to a 2020 Nielsen Norman Group report titled *PDF: Still Unfit for Human Consumption, 20 Years Later*.

Meanwhile irs.gov, created in 1995 specifically to transition the tax agency's documentation to digital, immediately adopted PDF and is still a keen advocate.

The Viral Spread of PDFs

Since the core specifications for PDF were released to open source by Adobe, a tranche of server-side processing tools and libraries have emerged, many now as venerable and entrenched as the 1996-era PDF specs, and as reliable and bug-resistant, while software vendors rushed to integrate PDF functionality into low-cost tools.

Consequently, loved or loathed by its host departments, PDFs remain ubiquitous in the communications and documentation frameworks across a huge number of US government departments.

In 2015 Adobe's VP Engineering for Document Cloud, Phil Ydens estimated that 2.5 trillion PDF documents exist in the world, while the format is believed to account for somewhere between 6-11% of all web content. In a tech culture addicted to disrupting old technologies, PDF has become ineradicable 'rust' – a central part of the structure that hosts it.

From 2018. There's scant evidence of a formidable challenger yet. Source: https://twitter.com/trbrtc/status/980407663690502145

According to a recent study from researchers at the University of Washington and the Library of Congress, *'hundreds of millions of unique U.S. Government documents posted to the web in PDF form have been archived by libraries to date'*.

Yet the researchers contend that this is just the 'tip of the iceberg'*:

> *'As leading digital history scholar Roy Rosenzweig had noted as early as 2003, when it comes to born-digital primary sources for scholarship, it is essential to develop methods and approaches that will scale to tens and hundreds of millions and even billions of digital [resources]. We have now arrived at the point where developing approaches for this scale is necessary.*
>
> *'As an example, the Library of Congress web archives now contain more than 20 billion individual digital resources.'*

PDFs: Resistant to Analysis

The Washington researchers' project applies a number of machine learning methods to a publicly available and annotated corpus of 1,000 select documents from the Library of Congress, with the intention of developing systems capable of lightning-fast, multimodal retrieval of text and image-based queries in frameworks that can scale up to the heights of current (and growing) PDF volumes, not only in government, but across a multiplicity of sectors.

As the paper observes, the accelerating pace of digitization across a range of Balkanized US government departments in the 1990s led to diverging policies and practices, and frequently to the adoption of PDF publishing methods that did not contain the same quality of metadata that was once the gold standard of government library services – or even very basic native PDF metadata, which might have been of some help in making PDF collections more accessible and friendly to indexing.

Discussing this period of disruption, the authors note:

> *'These efforts led to an explosive growth of the quantity of government publications, which in turn resulted in a breakdown of the general approach by which consistent metadata were produced for such publications and by which Libraries acquired copies of them.'*

Consequently, a typical PDF mountain exists without any context except the URLs that link directly to it. Further, the documents in the mountain are enclosed, self-referential, and don't form part of any 'saga' or narrative that current search methodologies are likely to discern, even though such hidden connections undoubtedly exist.

At the scale under consideration, manual annotation or curation is an impossible prospect. The corpus of data from which the project's 1000 Library of Congress documents were derived contains over 40 million PDFs, which the researchers intend to make an addressable challenge in the near future.

Computer Vision for PDF Analysis

Most of the prior research the authors cite uses text-based methods to extract features and high-level concepts from PDF material; by contrast, their project centers on deriving features and trends by examining the PDFs at a visual level, in line with current research into multimodal analysis of news content.

Though machine learning has also been applied in this way to PDF analysis via sector-specific schemes such as Semantic Scholar, the authors aim to create more high-level extraction pipelines that are widely applicable across a range of publications, rather than tuned to the strictures of science publishing or of other equally narrow sectors.

Addressing Unbalanced Data

In creating a metrics schema, the researchers have had to consider how skewed the data is, at least in terms of size-per-item.

Of the 1000 PDFs in the select dataset (which the authors presume to be representative of the 40 million from which they were drawn), 33% are only a page long, and 39% are 2-5 pages long. This puts 72% of the documents at five pages or fewer.

After this, there's quite a leap: 18% of the remaining documents run at 6-20 pages, 6% at 20-100 pages and 3% at 100+ pages. This means that the longest documents comprise the majority of individual pages extracted, while a less granular approach which considers the documents alone would skew attention towards the much more numerous shorter documents.

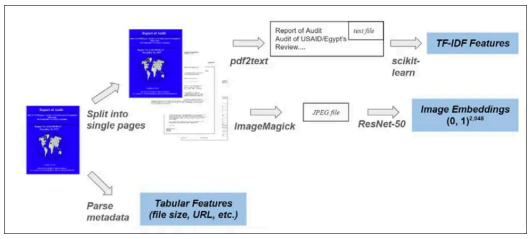
Nonetheless, these are insightful metrics, since single-page documents tend to be technical schematics or maps; 2-5 page documents tend to be press releases and forms; and the very long documents are generally book-length reports and publications, though, in terms of length, they're mixed in with vast automated data dumps that contain entirely different challenges for semantic interpretation.

Therefore, the researchers are treating this imbalance as a meaningful semantic property in itself. Nonetheless, the PDFs still need to be processed and quantified on a per-page basis.

Architecture

At the beginning of the process, the PDF's metadata is parsed into tabular data. This metadata is not going to be absent, because it consists of known quantities such as file size and the source URL.

The PDF is then split into pages, with each page converted to a JPEG format via ImageMagick. The image is then fed to a ResNet-50 network which derives a 2,048 dimensional vector from the second-to-last layer.



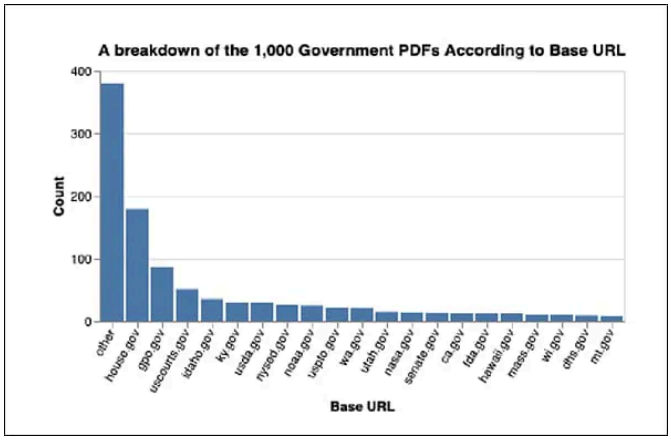
The pipeline for extraction from PDFs. Source: https://arxiv.org/ftp/arxiv/papers/2112/2112.02471.pdf

At the same time, the page is converted to a text file by pdf2text, and TF-IDF featurizations obtained via scikit-learn.

TF-IDF stands for Term Frequency Inverse Document Frequency, which measures the prevalence of each phrase within the document to its frequency throughout its host dataset, on a fine-grained scale of 0 to 1. The researchers have used single words (unigrams) as the smallest unit in the system's TF-IDF settings.

Though they acknowledge that machine learning has more sophisticated methods to offer than TF-IDF, the authors argue that anything more complex is unnecessary for the stated task.

The fact that each document has an associated source URL enables the system to determine the provenance of documents across the dataset.



This may seem trivial for a thousand documents, but it's going to be quite an eye-opener for 40 million+.

New Approaches to Text Search

One of the project's aims is to make search results for text-based queries more meaningful, allowing fruitful exploration without the need for excessive prior knowledge. The authors state:

> *'While keyword search is an intuitive and highly extensible method of search, it can also be limiting, as users are responsible for formulating keyword queries that retrieve relevant results.'*

Once the TF-IDF values are obtained, it's possible to calculate the most commonly featured words and estimate an 'average' document in the corpus. The researchers contend that since these cross-document keywords are usually meaningful, this process forms useful relationships for scholars to explore, which could not be obtained solely by individual indexing of the text of each document.

Visually, the process facilitates a 'mood board' of words emanating from various government departments:

Agency	Base URL	Top 10 TF-IDF Terms
United States Courts	uscourts.gov	court, RCJ, WGC, motion, CV, filed, case, district, united, states
United States Department of Agriculture	usda.gov	acres, none, farm, USDA, percent, sales, agriculture, farms, total, value
New York State Education Department	nysed.gov	students, school, indicator, target, disabilities, state, district, education, data, meets
National Oceanic and Atmospheric Administration	noaa.gov	NOAA, bay, forecast, sea, climate, species, N2O5, chart, marine, DLA
United States Patent and Trademark Office	uspto.gov	trademark, com, board, claims, appeal, patent, object, trial, virtual, john
National Aeronautics and Space Administration	nasa.gov	NASA, ingest, space, data, SABER, security, atmosphere, mission, spirit, prod
United States Senate	senate.gov	DTS, act, CME, treaty, senate, morgan, DOD, pool, industry, chairman
U.S. Food and Drug Administration	fda.gov	food, FDA, hearing, drug, lane, branch, irradiated, safety, studies, treatment

TF-IDF keywords for various US government departments, obtained by TF-IDF.

These extracted keywords and relationships can later be used to form dynamic matrices in search results, with the corpus of PDFs beginning to 'tell stories', and keyword relationships stringing together documents (possibly even over hundreds of years), to outline an explorable multi-part 'saga' for a topic or theme.

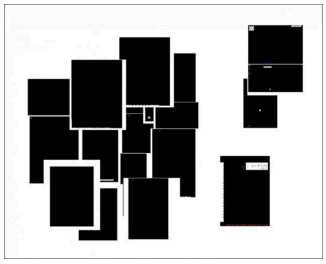
The researchers use k-means clustering to identify documents that are related, even where the documents don't share a common source. This enables the development of key-phrase metadata applicable across the dataset, which would manifest either as rankings for terms in a strict text search, or as nearby nodes in a more dynamic exploration environment:

Agency	Base URL	Descriptions of k-means Clusters
United States House of Representatives	house.gov (k=8)	Students & education; congressional press releases; healthcare; veterans
United States Courts	uscourts.gov	Bankruptcy court documents; RCJ-WGC documents from the Nevada United States District Court
United States Department of Agriculture	usda.gov	Tabular agricultural data; 2007 Census of Agriculture county reports
New York State Education Department	nysed.gov	Special education in New York State; assessments of schools ("school report cards")
U.S. Food and Drug Administration	fda.gov	Food labeling requirements; questionnaires regarding medical waivers for hearing aids

Visual Analysis

The true novelty of the Washington researchers' approach is to apply machine learning-based visual analysis techniques to the rasterized appearance of the PDFs in the dataset.

In this way, it's possible to generate a 'REDACTED' tag on a visual basis, where nothing in the text itself would necessarily provide a common enough basis.



A cluster of redacted PDF front pages identified by computer vision in the new project.

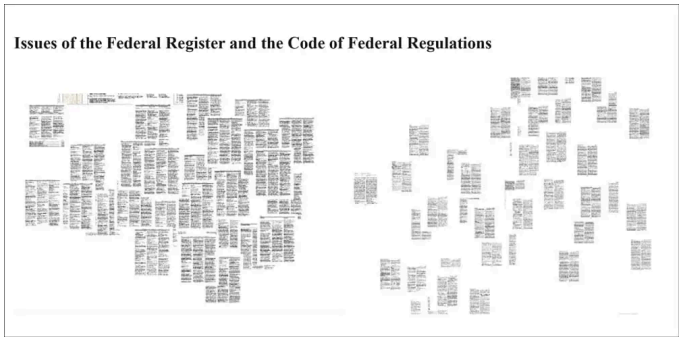
Furthermore, this method can derive such a tag even from government documents that have been rasterized, which is often the case with redacted material, making possible an exhaustive and comprehensive search for this practice.

Additionally, maps and schematics can be likewise identified and categorized, and the authors comment on this potential functionality:

> *'For scholars interested in disclosures of classified or otherwise sensitive information, it may well be of particular interest to isolate exactly this type of cluster of material for analysis and research.'*

The paper notes that a wide variety of visual indicators common to specific types of government PDF can likewise be used to classify documents and create 'sagas'. Such 'tokens' could be the Congressional seal, or other logos or recurrent visual features that have no semantic existence in a pure text search.

Further, documents which defy classification, or where the document comes from a non-common source, can be identified from their layout, such as columns, font types, and other distinctive facets.



Layout alone can afford groupings and classifications in a visual search space.

Though the authors have not neglected text, clearly the visual search space is what has driven this work.

> *'The ability to search and analyze PDFs according to their visual features is thus a capacious approach: it not only augments existing efforts surrounding textual analysis but also reimagines what search and analysis can be for born-digital content.'*

The authors intend to develop their framework to accommodate far, far larger datasets, including the *2008 End of Term Presidential Web Archive* dataset, which contains over 10 million items. Initially, however, they intend to scale up the system to address 'tens of thousands' of governmental PDFs.

The system is intended to be evaluated initially with real users, including librarians, archivists, lawyers, historians, and other scholars, and will evolve based on the feedback from these groups.

Grappling with the Scale of Born-Digital Government Publications: Toward Pipelines for Processing and Searching Millions of PDFs is written by Benjamin Charles Germain Lee (at the Paul G. Allen School for Computer Science & Engineering) and Trevor Owens, Public Historian in Residence and Head of Digital Content Management at the Library of Congress in Washington, D.C..

** My conversion of inline citations to hyperlinks.*

Originally published 28th December 2021

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More