Automating document anonymization & pseudonymization

October 2020

PDFTRON

What is redaction?

The process of censoring, deleting or obscuring confidential text and images in a document.

PDFTRON

Why do we need redaction?

Redaction use cases

- Redaction of PII (Personally Identifiable Information)
- Removal of proprietary data from documents shared during financial transactions such as M&A
- Redaction of PHI (Protected Health Information) from medical, health records and clinical trials
- Removing any bias from applications, resumes
- Removing any sensitive information for the release of data for training ML, users, employees
- To stay compliant with GDPR, California Privacy Act and LGPD to avoid fines and continue operating your business

Different approaches to redaction

Printing and redacting manually

Printing out the document, physically marking it up and scanning back in.

Drawbacks:

- Creation of multiple versions
- User has to download the document
- Quality of scanned images is affected (crooked, misaligned and low DPI)
- Redacted content can still be seen by changing the contrast values

Image based redaction

Image based redaction involves converting PDF to an image and blanking out the pixels.

Drawbacks:

- The document is no longer searchable
- The size might increase
- The quality is now lost at large zoom

True redaction by removing underlying elements

Removing elements marked for redaction underneath the coordinates directly in a PDF.

Benefits

- Removal of text, images, paths and other elements
- Maintain the original format
- Maintain searchability and original metadata

What are some redaction workflows?

Manual identification of sensitive content

The user has to manually scan through the content and then mark up content for redaction.

Drawbacks:

- The user might miss key content to be redacted
- The process is often time-consuming, especially with longer documents

Searching text for keywords

The user can perform search for specific keywords or pattern matching to find the content to be redacted by using various tools.

Drawbacks:

- Depending on the search engine used or text extraction algorithm might not yield the necessary results.
- The keyword could change from user to user and vary from what is actually inside of the document
- Check out the search results returned by PDF.js: https://youtu.be/-1QmlvIhtzs

Use of third-party services and tools

Users will fall back on improvised solutions where a formal workflow is unavailable.

Drawbacks:

- Downloading a local copy of the document
- Use of third-party tools that are not compliant or do not have retention policy
- Uploading sensitive data to REST API servers across multiple countries can lead to a higher risk of data loss or leaks
- Inefficient workflows lead to unnecessary time lost
- Barrier to enhancement

Redaction failures

Redaction failures

- Lawyers for former Trump campaign chair Paul Manafort failed to properly redact pleadings they filed in the federal court in 2019.
- Copy/paste revealed incriminating information underneath because content was only obscured with a black highlight; the text was still underneath.
- PDF documents released recently by the US government as part of the Jeffrey Epstein investigation were also reported to be improperly redacted, enabling public access to sensitive content underneath

Improving the redaction flow

Using true redaction

- Ensure you are using true redaction by removing all underlying elements
- Remove any PII in metadata or files attached inside of a PDF
- Adjust the cropbox of a PDF page to ensure there is no content outside of it that is not visible to the user
- Automate processes to increase accuracy and speed of review
- When verifying whether redaction took place, test by:
    - Select text in redacted areas
    - Attempt to copy/paste underlying content
    - Try to remove the redaction annotations
    - Use a low-level PDF inspector to confirm redacted elements have been removed

Avoid user download

- Provide users a workflow built into your application to avoid download and creation of multiple conflicting versions
- Ensure compliance and avoid document being leaked due to improper email exchange or another party getting access
- Gain full control over document lifecycle and transition over network

Leverage client-side

Advantages of client-side redaction:

- Minimized security risks
- Less server infrastructure and cost-efficient
- Reduced network traffic
- Easier to deploy and scale
- All of processing is offloaded to the client
- Save a new copy of redacted document or use right away in the next step of the workflow

Anonymization vs. Pseudonymization

Anonymization

The anonymization ensures that the person or data cannot be identified.

Both CCPA and GDPR allow sale of anonymized data.

SALE AND PURCHASE AGREEMENT

THIS AGREEMENT is made this 23rd day of August, 2018.

BETWEEN

(1) [redacted] (hereinafter referred to as "the Vendor"); and

(2) [redacted] a company incorporated in Canada and having its registered office at [redacted] (hereinafter referred to as "the Bank" which expression shall include its successors and assigns).

WHEREAS

(A) In consideration for the sum of [redacted] [plus the applicable Goods and Services Tax ("GST")], the Vendor has agreed to sell, and the Bank has agreed to purchase, the Goods described on the Schedule hereto ("the Goods"), in accordance with the terms and conditions.

(B) The Vendor agrees to transfer all his legal and beneficial title to the Bank absolutely, free from all charges liens and other encumbrance, upon the execution of this Agreement.

Redacted document readability

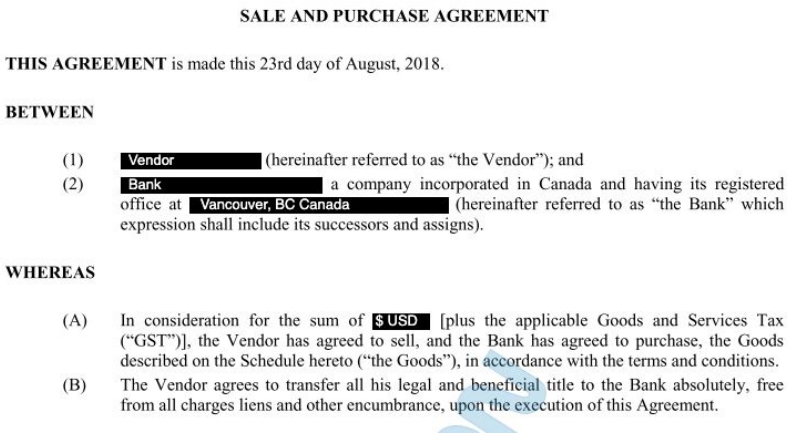
If the document is heavily redacted, it might lose a lot of contextual information necessary to connect the dots.

Therefore, many redactors add labels over top excised areas to identify what type of information was removed (pseudonymization - de-identification process):

- In case of Date of Birth, replace it with a date range (25-40 years of age)
- Addresses can be replaced with more general geographical location
- Amounts can be replaced with approximate range

Pseudonymization

The document did not lose context, and any identifying data have been removed.



SALE AND PURCHASE AGREEMENT

THIS AGREEMENT is made this 23rd day of August, 2018.

BETWEEN

(1) [redacted] (hereinafter referred to as "the Vendor"); and

(2) [redacted] a company incorporated in Canada and having its registered office at [redacted] (hereinafter referred to as "the Bank" which expression shall include its successors and assigns).

WHEREAS

(A) In consideration for the sum of [redacted] [plus the applicable Goods and Services Tax ("GST")], the Vendor has agreed to sell, and the Bank has agreed to purchase, the Goods described on the Schedule hereto ("the Goods"), in accordance with the terms and conditions.

(B) The Vendor agrees to transfer all his legal and beneficial title to the Bank absolutely, free from all charges liens and other encumbrance, upon the execution of this Agreement.

Using ML training, tagged structure of PDF the accuracy of replacements can be greatly improved.

Pseudonymization next steps

How can we improve the accuracy and provide meaningful context?

- Tags embedded in the PDF can give us insights
- Using Generative Pre-trained Transformer 3 from OpenAI to recognize human-like text from the search term provided to perform accurate redactions

Question

How many credit-card transaction data points does it take to uniquely identify a person?

- 2
- 4
- 6
- 10

Answer

90% of people can be identified from **four** samples out of 30 days of credit card transaction data from 1.1 million people

Questions