Safe PDF text redaction tool

Supervisor: Maarten Marx (maartenmarx@uva.nl)

Safe PDF redaction tool

Text redactie (zwart- of weg-lakken in het NL) wordt heel erg veel gedaan door de overheid op stukken die openbaar gemaakt worden. Dit kan gaan om persoonsgegevens, of gevoelige informatie. Zwart-lakken is een informatica onderwerp dat heel simpel aan 2 eisen moet voldoen

1. wat wegge"lakt" wordt moet ook echt weg zijn (**safety**). Het moet echt niet terug te halen zijn, ook niet met geavanceerde slimme technieken. Dus echt nergens in de PDF niet (niet in de meta-data, de inhoudsopgave, de tekst natuurlijk, allicht in plaatjes, weet ik het...)
2. wat niet weggelakt hoeft te worden wordt ook niet veranderd of aangetast.

Het lukt de NL overheid in heel veel gevallen niet om hieraan te voldoen. We zijn vooral dat een enorme focus of de *safety* zorgt voor een grote aantasting van de niet weg te lakken tekst. In het ergste geval wordt gewoon alle tekst weggehaald en worden alleen afbeeldingen van PDF documenten openbaar gemaakt.

In dit project laat je zien dat er echt wel een simpele redactie tool voor PDF te maken valt die wel aan de 2 esien voldoet, iets wat hevig ontkend wordt door "IT specialisten" bij de NL overheid. Pymupdf
lijkt een zeer geschikte tool, in ieder geval een mooi begin.

Zie ook

* X-ray: <https://github.com/freelawproject/x-ray>
* Bland et al, 2022: <https://arxiv.org/abs/2206.02285>

Project offered to study/studies: Bachelor Informatica (computer science)

Max number of students: 1