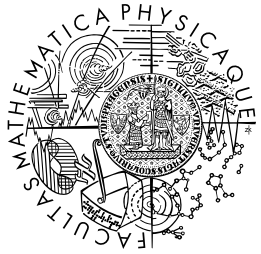

MATEMATICKO-FYZIKÁLNÍ
FAKULTA
Univerzita Karlova

**BAKALÁRSKA PRÁCA**

Lukáš Salak

# Detekcia anonymizovaných častí v PDF dokumentoch

Katedra softwaru a výuky informatiky

| | |
|---|---|
| Vedúci bakalárskej práce: | doc. RNDr. Elena Šikudová, Ph.D. |
| Študijný program: | Informatika |
| Študijný obor: | Počítačová grafika, vidění a vývoj her |

Praha 2023

Prohlašuji, že jsem tuto bakalársku práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne .....  .....

Podpis autora

Dedikácia. Nesmierne si cením podpory a pomoci, ktoré som dostal od mnohých ľudí, menovite od mojej rodiny, priateľky, priateľov, spolužiakov a kolegov. Najviac však by som chcel poďakovať vedúcej mojej bakalárskej práce, doc. RNDr. Elene Šikudovej, Ph.D., ktorá mi pomohla pri všetkom, čo som potreboval.

Názov práce: Detekcia anonymizovaných častí v PDF dokumentoch

Autor: Lukáš Salak

Katedra: Katedra softwaru a výuky informatiky

Vedúci bakalárskej práce: doc. RNDr. Elena Šikudová, Ph.D., Katedra softwaru a výuky informatiky

Abstrakt: You will need to submit both Czech and English abstract to the SIS, no matter what language you use in the thesis. If writing in English, translate the contents of `\AbstractEN` into this field. In case you do not speak czech, your supervisor should be able to help you with the translation.

Kľúčové slová: key words

Title: Detection of anonymized parts in PDFs

Author: Lukáš Salak

Department: Department of Software and Computer Science Education

Supervisor: doc. RNDr. Elena Šikudová, Ph.D., Department of Software and Computer Science Education

Abstract: Abstracts are an abstract form of art. Use the most precise, shortest sentences that state what problem the thesis addresses, how it is approached, pinpoint the exact result achieved, and describe the applications and significance of the results. Highlight anything novel that was discovered or improved by the thesis. Maximum length is 200 words, but try to fit into 120. Abstracts are often used for deciding if a reviewer will be suitable for the thesis; a well-written abstract thus increases the probability of getting a reviewer who will like the thesis.

Keywords: key words

# Obsah

# Úvod

Problém anonymizácie dát je dôležitý v rôznych oblastiach, napríklad v oblasti verejnej správy či v oblasti marketingu. Pod pojmom anonymizácie dokumentov si môžeme predstaviť vymazanie či skrytie údajov alebo iných citlivých informácií. Možností, ako pristupovať k anonymizácii dokumentov, je veľa.

Typickým miestom, kde sa stretávame s anonymizáciou dokumentov, je oblasť verejnej správy. V Českej republike majú organizácie verejnej správy povinnosť zverejňovať informácie o svojej činnosti, k čomu patrí aj zverejňovanie uzavrených zmlúv nad určitú čiastku do *registru zmlúv*, ktorý je verejne prístupný. Nachádzajú sa tu nielen informácie o predmete zmlúv, zmluvných stranách a cene, ale takisto všetky súbory, ktoré sú súčasťou zmlúv. Register zmlúv je významným nástrojom, ktorý zlepšuje transparentnosť; podstatou je kontrolovať a mať možnosť obmedziť korupciu a zneužívanie verejnej moci kvôli uzatváraniu nevýhodných zmlúv.

Aj napriek tomu, že zverejňovanie dát do registra zmlúv je právne vynútiteľné, nezabezpečuje to automaticky možnosť jednoduchého vyhľadávania či analýzy týchto dát. K tomu bol vytvorený projekt, webový portál *Hlídač smluv*, ktorý má za úlohu zlepšiť prístup k registru zmlúv. Neskôr, po skombinovaní ďalších verejne prístupných dát z registrov a databází, sa vytvoril projekt *Hlídač státu* [1], ktorý má za úlohu zlepšiť prístup k verejným informáciám. Poskytuje napríklad plnohodnotné vyhľadávanie v texte zmlúv.

V registri zmlúv sú dokumenty z rôznych oblastí, napríklad z oblasti zdravotníctva, školstva, realitných služieb alebo IT projektov. V prípade, že dokumenty obsahujú citlivé údaje, sú častokrát anonymizované. V súčasnej dobe neexistuje štatistický nástroj, ktorý by znázorňoval koľko percent v takýchto dokumentoch je zanonymizovaných.

V tejto práci sa budeme zaoberať anonymizovanými PDF dokumentmi a budeme sa snažiť vytvoriť nástroj, ktorý bude schopný detekovať anonymizované časti dokumentu, využiť metódy strojového učenia a ďalších algoritmov na spracovanie obrazu a navrhnúť tak systém, ktorý umožní na základe dostupných dát vyhodnotiť percento anonymizácie jednotlivých zmlúv pri použití konkrétnych implementačných metód. Túto implementáciu potom bude možné nasadiť na

[1] `https://hlidacstatu.cz`

webový portál Hlídače státu.

Pri implementácii je nutné uvedomiť si rozličné spôsoby anonymizovania týchto dát, z ktorých najčastejšími sú:

- prekrytie časti dokumentu čiernou farbou
- prekrytie časti dokumentu bielou farbou
- zašumenie časti dokumentu

Jednou z ďalších komplikácií je fakt, že neexistuje sada zmlúv, kde je známe či je dokument anonymizovaný a ak, tak aké percento dokumentu je anonymizované.

Hlavným prínosom práce je porovnanie jednotlivých odvetví, ktoré zverejňujú zmluvy vzhľadom ku percentu anonymizácie a tvorba štatistiky vzhľadom na anonymizáciu dokumentov relatívne k jednotlivým oblastiam.

Vďaka tomuto procesu sme našli najvýhodnejšie metódy, ktoré sa môžu použiť na detekciu anonymizovaných častí dokumentov.

Práca je štruktúrovaná následovne. 1. kapitola je venovaná popisu registru zmlúv v Česku a medzinárodnému porovnaniu vzhľadom k anonymizácii dokumentov. V 2. kapitole je špecifikovaný konkrétny problém a zadanie, ktorému sa v práci venujeme. 3. kapitola je venovaná popisu vstupných dát, ich obsah a štruktúra. Takisto je tu popísaný proces získavania dát a ich príprava na ďalšie spracovanie. V 4. kapitole je popísaný proces detekcie anonymizovaných častí dokumentov. V 5. kapitole je popísaný proces vytvorenia štatistík a ich vizualizácie. Kapitola 6 je venovaná diskusii a záveru práce.

# Kapitola 1

# Register zmlúv, zverejňovanie a anonymizácia dokumentov

# Kapitola 2

# Špecifikácia problému

# Kapitola 3

# Popis vstupných dát a ich spracovanie

# Kapitola 4

# Proces detekcie anonymizovaných častí dokumentov

# Kapitola 5

# Štatistika a vizualizácia získaných dát

# Kapitola 6

# Implementácia riešenia

# Závěr

In the conclusion, you should summarize what was achieved by the thesis. In a few paragraphs, try to answer the following:

- Was the problem stated in the introduction solved? (Ideally include a list of successfully achieved goals.)
- What is the quality of the result? Is the problem solved for good and the mankind does not need to ever think about it again, or just partially improved upon? (Is the incompleteness caused by overwhelming problem complexity that would be out of thesis scope, or any theoretical reasons, such as computational hardness?)

  This is quite common.
- Does the result have any practical applications that improve upon something realistic?
- Is there any good future development or research direction that could further improve the results of this thesis? (This is often summarized in a separate subsection called 'Future work'.)

# Zoznam použitej literatúry

# Příloha A

# Príloha: Implementácia riešenia