Posudek oponenta bakalářské práce

Jméno a příjmení:	Jaroslav Malát
Název tématu:	Zabezpečené zpracování medicínských obrazových dat
Vedoucí práce:	Ing. Petr Včelák (NTIS)

Hlavním cílem práce bylo seznámení se s požadavky bezpečnostními a legislativními a analýza existujících metod a řešení anonymizace obrazových medicínských dat. Na základě analýzy měl student navrhnout a implementovat vlastní řešení, které zlepší dosavadní výsledky. Současně také bylo cílem dosažené výsledky ověřit a diskutovat.

Obsah práce

Formální úroveň

Po formální stránce patří práce mezi slabší neboť obsahuje řadu překlepů, neúplných vět, chybných konstrukcí věty a nedodržuje typografická pravidla. Dle metodiky KIV je rozsah dokumentu necelých 41 normo stran. Práce obsahuje seznam literatury, tabulek, obrázků, zkratek, dvě přílohy (jedna z nich ukazuje jeden anonymizovaný snímek) a CD-ROM.

Logická struktura

Práce není jasně logicky členěna. V první kapitole student seznamuje čtenáře s bezpečnostní problematikou formou uvedení celkem sedmi vyhlášek nebo zákonů a dále popisuje formát DICOM. Druhá kapitola s názvem „*Analýza dostupných metod anonymizace a návrh algoritmu*“, kde popisuje jen čtyři existující nástroje, z nichž se k tématu práce vztahuje pouze jediný. Na konci druhé kapitoly (str. 24) student přilepil odstavec označený návrh algoritmu. Třetí kapitola opět označená jako „*Analýza anonymizace obrazových dat*“ s analýzou nemá nic společného. Spíše jde o popis návrhu vlastního řešení. Ostatně ve třetí kapitole se vyskytuje podkapitola 3.2 s popisem použitých nástrojů, které dle mého patří spíše do implementace. V analýze by dával smysl popis existujících nástrojů, ale nikoliv jen těch opravdu použitých. Snad pouze kapitola „*3.1 Metody hledání textových řetězců*“ by v této kapitole dávala smysl, ale to by se student nesměl spokojit s jediným algoritmem „hrubé síly“, jak jej sám označuje a zamýšlí použít pro hledání bezpečných/bílých slov.

Ve čtvrté kapitole je popsána implementace všech celkem tří vytvořených tříd a jejich metod následovaná popisem předpřípravy obrazových souborů, zmínkou o trénování nástroje tesseract-ocr a použití knihovny jSoup. Pátou kapitolou už je pouze hodnocení výsledků a text uzavírá kapitola závěr.

Obsah

Celkově obsah a forma studentem předkládaného dokumentu neodpovídá pravidlům a zvyklostem pro psaní odborných textů. Texty jsou velmi obecné, povrchní a místy až bulvární. Tvrzení (a mnohdy dosti silná) nejsou podložena žádnými zdroji, která by je dokládala. Nedostatečné studium a seznámení se s existujícími nástroji pro požadovanou anonymizaci medicínských obrazových dat je zřejmé i z druhé kapitoly. Stejně tak student vůbec nebere v potaz legislativu platnou v zahraničí.

Od počátku student používá termín „citlivá data“, ale nikde explicitně nedefinuje co si pod nimi představuje, takže lze jen odhadovat.

V teoretické části student v podstatě opisuje vyhlášky a zákony ČR, ale bez hlubšího uvedení kontextu a detailů i bez přínosu k tématu práce. Uvedená tvrzení jsou nepodložená, např. „*Strany zastávající úpravu NZIS tvrdí...*“, „*Kritici tvrdí, že v zákoně chybí účelné a přesně stanovené postupy zpracování citlivých údajů...*“ (str. 6, poslední odstavec) a při bližším pohledu se jedná ze strany autora o **plagiátorství**, protože text není jeho původní a necituje zdroj odpovídajícím způsobem.

Pro *Zákon č. 181/2014Sb. o kybernetické bezpečnosti* nebo *Vyhláška č. 317/2017 Sb. o významných informačních systémech a jejich určujících kritériích*, vzdáleně s tématem souviset může, ale

v uvedeném textu postrádám uvedení kontextu, proč autor zákon zmiňuje v souvislosti s anonymizací medicínských dat a jejich zabezpečeného zpracování.

Popis formátu DICOM je velmi obecný a na stranách 12 18 poskytuje pouze přehled částí standardu, nikoliv však konkrétní popis možností pro uložení snímku vyšetření na modalitě, který pak má být použit pro anonymizaci textu. V kapitole 1.3 věnované PACS pak přidává další informace k DICOMu, opět **téměř doslovně převzaté a bez správné citace.**

Dále student v teoretické části prezentuje čtyři nástroje, z nichž s tématem souvisí pouze jediný (Zorro Anonymizace) a ostatní tři se věnuji rozpoznání obličeje (Facepixelizer, Lionytics Image Anoymizer, PhotoHide) a jeho skrytí nebo rozostření. Existenci skutečných nástrojů pro práci s formátem DICOM a jejich anonymizací tedy ignoruje.

Kvalita řešení a dosažených výsledků

Zdrojové kódy přiložené na CD-ROM spočívají ve třech Java souborech s celkovým počtem 475 řádek včetně komentářů (původně bylo 285 řádek v jednom Java souboru a bez komentářů). Zdrojové kódy i jejich komentáře jsou v českém jazyce.

U zdrojových souborů jsou uloženy soubory *manifest*, překladový shellový skript *skript.sh*, soubor *bilaSlova.txt* se 7 povolenými slovy), soubor *properties.properties* s nastavením tří hodnot (název souboru k anonymizaci, cesta k programu a soubor s povolenými slovy). Přiložena je potřebná knihovna *JSoup*.

Aplikace pouze volá externí *convert*, pro zvětšení snímku, a OCR nástroj *tesseract*, který provede nalezení textu ve snímku metodou OCR. Kvalita zdrojového kódu se mírně zlepšila. Seznam slov, jež mají být při anonymizaci ignorovány, již nejsou uvedeny jako jednotlivé podmínky, ale jsou v souboru *bilaSlova.txt* (obsahuje 6 slov). Oproti minulému odevzdání, text zmínku o Hammingově vzdálenosti ani regulárních výrazech neobsahuje. Bohužel ani ve zdrojových kódech ke zlepšení v tomto směru nedošlo a veškerá logika je pevně zakódována. Název souboru pro anonymizaci je nutné zapsat do properties souboru, který je argumentem aplikace. O uživatelské přívětivosti tohoto řešení pro zpracování velkého množství dat nemůže být řeč.

V textu student informace o trénování použitého nástroje Tesseract OCR doplnil, ale soubor s natrénovanými daty přiložen není. Ve zdrojových kódech není o takové možnosti žádné stopy. Zdrojový kód obsahuje napevno nastavený anglický jazyk, bez možnosti změny volbou argumentů v používaném properties souboru.

Autor v 5. kapitole tvrdí, že z 86 souborů jich bylo použitelných pouze 45, protože ostatní neobsahovaly text. Dle mého však při automatickém zpracování snímků nehraje roli, zda ve snímku text skutečně je či nikoliv, ale zda bude nějaký text provedením OCR nalezen a jak bude následně vyhodnocen. Je nutné se vypořádat i se situací, kdy je nalezen neexistující text. Ve skutečnosti vzorek obsahoval pouze 22 souborů zcela bez textu, ostatní obsahovaly vždy alespoň jeden znak.

Jediný předložený výsledek je částečně anonymizovaný snímek v příloze bakalářské práce. Práce neobsahuje kapitolu diskuze nebo porovnání s jinými autory. Fáze testování, ověření a zhodnocení výsledků v dokumentu chybí, resp. na str. 45 (4. odstavec) uvádí pouze text: „*Na závěr bych chtěl podotknout, ze výsledky programu, který úspěšně anonymizuje citlivá medicínská data nalezené v dodaných obrazových souborech části DICOM, mě samotného překvapily*“. Také opět uvádí 80% úspěšnost rozeznávání znaků OCR nástrojem Tessearact, ale nikoliv jak zlepšil proces anonymizace oproti již existujícím nástrojům jak bylo v zadání požadováno.

Práce s literaturou

Autorova práce s literaturou neodpovídá požadavkům na kvalifikační práci, pravidlům pro citování zdrojů v odborné publikaci a seznam zdrojů neodpovídá používané citační normě.

Seznam literatury obsahuje 22 elektronických zdrojů, které částečně odpovídají popisované problematice (vyhlášky, zákony, DICOM), ale existující nástroje a metody anonymizace popsány nejsou. Zahraniční literatura není dostatečně zastoupena a zcela chybí zahraniční literatura pro oblast legislativy a existující nástroje pro anonymizaci DICOM souborů.

Text práce nepovažuji v celém rozsahu za autorův původní. Minimálně níže uvedené části jsou dle mého plagiátem:

- text na 5 7. straně o NZIS pochází z webu `http://ferovanemocnice.cz/casto-kladene-otazky.html` v plném rozsahu.
- poslední odstavec na 18. straně až 2. odstavec na straně následující `https://www.systemonline.cz/it-pro-verejny-sektor-a-zdravotnictvi/systemy-pacs-z-hlediska-databazovych-systemu.htm`.

Doplňující informace

Po studentovi jsem požadoval zaslání výsledků nad testovací sadou souborů a hodnocení úspěšnosti. Výsledky a hodnocení mi poskytl až po měsíci od mého požadavku s tvrzením 83% úspěšnosti (114 ze 137 řetězců). Jen při pohledu na poskytnuté výsledky to skutečnosti neodpovídá. Ve snímcích jsem napočítal celkově 256 řetězců jméno a příjmení (počítám jako jeden jako uvedl autor v emailu), datum, čas, rodné číslo nebo číslo vyšetření, název zařízení nebo oddělení (počítám jen jako jeden řetězec), které mají být začerněny. V anonymizovaných datech jsem napočítal těchto řetězců 176, tj. celkem bylo anonymizováno pouze 31 % případů. Ze všech souborů, kde byl nějaký text k odstranění jich bylo po provedené anonymizaci pouze 6 zcela čistých souborů, tj. s autorem výše uvedených 45 souborech s citlivými texty se dostaneme na 13% úspěšnost.

Splnění zadání

Zadání *nebylo* splněno.

1. Autor uvádí pouze legislativní požadavky v rámci ČR. Legislativu ze zahraničí vůbec neuvažuje. Popis formátu DICOM je velmi obecný, na úrovni popularizačních článků. Stejně tak řada informací o DICOMu se netýká přímo řešeného problému.
2. Analýza dostupných metod anonymizace spočívá v popisu jediného nástroje zaměřeného na anonymizaci dokumentů, ale bez informace o výsledku otestování s obrazovým vyšetřením. Uvedený text nepopisuje návrh algoritmu, který by vysloveně zlepšoval proces anonymizace obrazových medicínských dat. Navrhovaný a implementovaný nástroj pouze nějakou formou anonymizaci provede.
3. Otestování a ověření implementovaného způsobu anonymizace student neřešil.
4. Vyhodnocení výsledků nebo porovnání s jinými autory student vůbec neřešil.

Dotazy k práci

Nemám.

Závěr

Zadání a zásady pro vypracování bakalářské práce student *nesplnil.* Autor nerespektuje pravidla pro psaní kvalifikačních prací, práci s literaturou a citováním zdrojů, dopouští se plagiátorství a zároveň ani nesplňuje zásady pro vypracování.
Celkově navrhuji hodnocení známkou *nevyhověl* a práci *nedoporučuji k obhajobě.*

V Plzni 12. 8. 2016

Ing. Petr Včelák
NTIS, ZČU