Posudek oponenta bakalářské práce

Jméno a příjmení:	Jaroslav Malát
Název tématu:	Bezpečnostní analýza v oblasti elektronické výměny informací o pacientech
Vedoucí práce:	Ing. Petr Včelák (NTIS)

Obsah práce

Logická struktura

Práci chybí logické členění do očekávaných částí jako seznámení se s problematikou, analýza, návrh, implementace, ověření a diskuze. Tyto části se v práci různě prolínají. Z názvů kapitol práce je zřejmé, že autor postrádá jasný cíl své práce i přehled popisované problematiky.

Již v kapitole Úvod si autor suverénně *mění cíle práce* a stanovuje si body, které jsou v rozporu s vytyčenými zásadami pro vypracování.

Obsah

Ve druhé kapitole lze nalézt náznaky analýzy, ale texty jsou velmi obecné a povrchní. Následuje kapitola *Právní předpisy*, kde jsou uvedeny platné vyhlášky a zákony. Na 9. straně autor nepravdivě tvrdí *„V zákoně č. 227/200 Sb., o elektronickém podpisu s novelizací ... lze nalézt důležité informace týkající se zabezpečeného přenosu dat"*. Žádná zmínka o zabezpečeném přenosu dat se v těchto zákonech nevyskytuje. Ostatně elektronický podpis, jako obdoba klasického podpisu, není určen pro zabezpečení dat z pohledu jak autor zamýšlí. Pro práci na téma bezpečnostní analýzy je to závažným nedostatkem. Autorem jasně vymezený kontext chybí.

U některých dalších zákonů nebo vyhlášek lze pochybovat o jejich souvislosti s řešeným tématem. Zejména v kapitole 3.6 (str. 11 12) je uvedena Vyhláška č. 62/2015 Sb.[1] což je popis *„náležitostí dokumentace u používaných zdravotnických prostředků"*. Mezi zdravotnické prostředky vyhláška řadí např. v § 6 inhalátory, respirační zdravotnické prostředky nebo kontaktní čočky. Student dle mého text vyhlášky ani nečetl.

Zákon č. 181/2014 Sb. o kybernetické bezpečnosti, se vzdáleně s tématem práce sice shoduje, ale v uvedeném textu postrádám uvedení kontextu, proč autor zákon vůbec zmiňuje v souvislosti s výměnou informací o pacientech. Konkrétní příklady opatření uvedeny nejsou a jedná se jen o obecná konstatování nebo popis zákona a nařízení nebo vyhlášek. Podobně je tomu s vyhláškou č. 317/2014 Sb. o významných informačních systémech a jejich určujících kritériích.

Ve čtvrté kapitole, tj. na listech 19 31 student povrchně popisuje DICOM formát, ale uvádí informace, které nemají s tématem práce přímou souvislost. Ze 13 listů práce o formátu DICOM se vyskytuje náznak technického řešení bezpečnosti na listech 27 a 28 v doslova pár bodech při popisu historie zavedení DICOM standardu v ČR (kapitola 4.3), které však nejsou autorovi původní a nerespektuje pravidla pro citování textů.

V závěru práce student uvádí (druhý odstavec) zavádějící tvrzení *„Kromě očekávaných údajů ve formátu DASTA a DICOM jsou identifikovány a odstraněny citlivé údaje z obrazových souborů."* V praktické části student řešil pouze přímo obrazové soubory (PNG) a uváděné formáty neměl student k dispozici.

Rozsah

V zadání doporučený rozsah 30 stran původního textu nepovažuji za jednoznačně splněný. Dle formalizovaného postupu pro zjištění počtu stran kvalifikační práce na KIV vychází hodnota cca 36 listů. Dle výše popsaného obsahu práce jsou kapitoly 3.6 (zdravotnické prostředky; cca 1 list) a 4 (cca 12 listů ze 13) zcela mimo téma práce. Lze je označit pouze jako výplňový text.

[1] O provedení některých ustanovení zákona o zdravotnických prostředcích

Práce obsahuje jednu přílohu se snímkem po provedené anonymizaci textu v obrazu. Součástí je CD-ROM s textem dokumentu a autorem vytvořený program se zdrojovým kódem a potřebnou knihovnou *JSoup*.

Kvalita řešení a dosažených výsledků

Do teoretické části měla spadat (1) analýza legislativních bezpečnostních požadavků a (2) analýza dostupných technických prostředků. V dokumentu analýza uvedena není ani v kapitolách pojmenovaných nebo evokujících u čtenáře analýzu. Shodně chybí i jasně ohraničený třetí bod zásad vypracování spočívající ve stanovení možných bezpečnostních rizik a hodnocení nebo analýza dopadu.

V 5. kapitole autor rovnou přistupuje přímo k praktické části práce, kde řeší anonymizaci obrazových dat pro opakované využití při výzkumu. Kapitola již nějaké použitelné informace obsahuje, ale bohužel jsou logické části návrhu a implementace spojeny a to i přes existenci 6. kapitoly *Implementace*, kde je spíše programová dokumentace a popis testování a výsledku řešení.

Zdrojové kódy spočívají v jediném nekomentovaném Java souboru s 284 řádky. Ve své podstatě kód pouze volá externí aplikaci *convert*, pro zvětšení snímku, a OCR nástroj *tesseract*, který provede nalezení textu ve snímku. Kvalita zdrojového kódu je trestuhodná. Seznam slov, jež mají být při anonymizaci ignorovány, je uveden jako jednotlivé podmínky. Také řada dalších řetězců je přímo ve zdrojovém kódu a nikoliv v externím souboru.

Student na 32. stránce práce uvádí použití Hammingovy vzdálenosti slovy: *„je pro můj program zásadní, jelikož bez jejího použití by se procentuální úspěšnost anonymizace znaků značně snížila“*. Zarážející je, že ve zdrojových kódech použití Hammingovy vzdálenosti není ani náznakem.

S regulárními výrazy (5.1.5) autor pracuje stejně jako s Hammingovou vzdálenosti, tedy nijak. V textu regulární výrazy uvádí, ale ve zdrojových kódech testuje konkrétní znak získaný metodou *String.charAt()* na shodu.

Student neřešil žádnou optimalizaci OCR procesu nebo testování různých parametrů a spokojil se s výchozím nastavením nástroje a bez trénování, které *tesseract* umožňuje. Jediným přínosem práce je tvrzení, že v testovaných případech po zvětšení snímku/obrázku dává OCR proces lepší výsledky, ale není jasné jaké je zlepšení.

Chybí fáze testování nebo ověření s uvedením úspěšnosti anonymizace se vzorkem dat, jež měl student k dispozici. Autor pouze uvádí 80% úspěšnost rozeznávání znaků OCR nástrojem Tessearact. Není však jasné, zda k této hodnotě autor dospěl sám nebo jen (opět) neuvedl citaci. Za své výsledky prezentuje jediný výsledek a to je anonymizovaný snímek v příloze. V textu není zmíněna konkrétní verze používaných externích nástrojů *convert* a *tesseract*. Bakalářská práce neobsahuje kapitolu diskuze nebo porovnání s jinými autory.

Formální úroveň

Po formální stránce patří práce mezi slabší. Obsahuje řadu překlepů, neúplných vět nebo chybných konstrukcí věty. Autor nedodržuje typografická pravidla. Čitelnosti textu nepomáhají dlouhé výčty (str. 18). Názvy kapitol nezačínají vždy na nové stránce. Číslování kapitol je chybné viz chybějící podkapitoly 5.1.2 a 5.1.4. Seznamy vložených objektů (obrázky, tabulky), ale i zkratek v práci chybí.

Práce s literaturou

Práce s literaturou je zcela nedostatečná a některé části textu lze označit za plagiát. Seznam literatury obsahuje pouhých 9 položek. Ve všech případech se jedná o elektronické zdroje. Text práce nepovažuji v celém rozsahu za autorův původní.

- Minimálně na 3 4. straně své práce autor použil cizí text (doslovně) o rozsahu cca 3/4 strany a bez provedení přímé citace textu. Odkaz na první položku seznamu literatury autor v jednom místě sice uvádí, ale stylem, že uvedený text působí jako vlastní.

- Kapitola 4.3 (str. 26 31) i přes uvedení zdroje (č. 6) na úplném počátku, je „velmi podobná" strukturou i textem a rozhodně neodpovídá zvyklostem pro použití cizích zdrojů a jejich citování. Ani obrázek 4.1 (str. 31) nemá cizí zdroj uveden.

Splnění zadání

Zadání *nebylo* splněno.

1. Autor uvádí seznámení s legislativou, ale nikoliv analýzu vyplývajících požadavků. Chybí jednoznačná kapitola nebo souhrn odpovídající vlastní analýze popsané legislativy.
2. Analýza dostupných technických prostředků není uvedena. V rámci popisu nasazení DICOM standardu v ČR jsou některé technické prostředky zmíněny, ale o analýzu autora se nejedná.
3. Bezpečnostní rizika student nehledal ani nestanovil. Stejně tak chybí analýza jejich dopadu.
4. Vyhodnocení uvedených analýz z různých pohledů student vůbec neřešil.

Doplňující informace k práci

Autor si v úvodu práce (strana 1) zvolil body práce odlišné od oficiálního zadání. Až na první bod, který se původnímu podobá, jsou zbývající tři body odlišné. Popis formátu DICOM nemá zásadní opodstatnění pro toto téma a vlastně ani pro následně prováděnou anonymizaci dat, neboť student při anonymizaci používal jen obrazová data uložená v PNG souborech.

Z odevzdané práce není jednoznačně jasné co mělo být výsledkem. Pokud předpokládám, že jeho hlavním přínosem (bez ohledu na dosažený výsledek a kvalitu) měla být anonymizace dat pro opakované využití při výzkumu, pak není nejmenší důvod proč by nemohl body oficiálního zadání student splnit právě s ohledem na využití daného typu dat o pacientech pro další výzkum.

Dotazy k práci

1. V textu zmiňujete Hammingovu vzdálenost jako klíčovou metodu. Vysvětlete nebo vhodně ilustrujte algoritmus jak, ve vašem případě, Hammingovu vzdálenost počítat? Zejména jak získáte druhý vektor pro výpočet Hammingovy vzdálenosti? Jak zjištěná vzdálenost vektorů pomůže při rozhodnutí o nalezení neznámého textu získaného z OCR?
2. Vysvětlete význam zákona č. 181/2014 Sb. (o kybernetické bezpečnosti) a vyhlášky č. 317/2014 Sb. (o významných informačních systémech a jejich určujících kritériích) v kontextu vašeho tématu zaměřeného na elektronickou výměnu dat o pacientech.

Závěr

Zadání a zásady pro vypracování bakalářské práce student *nesplnil.* Komisi navrhuji zvážit změnu téma nebo požadovat po studentovi významné přepracování textu při dodržení zásad obvyklých pro psaní kvalifikačních prací a citování pramenů.

Celkově navrhuji hodnocení známkou *nevyhověl* a práci *nedoporučuji k obhajobě.*

V Plzni 5. 8. 2015

Ing. Petr Včelák
NTIS, ZČU