FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Hodnocení vedoucího závěrečné práce

Student:	Vojtěch Skalák
Vedoucí: práce:	Ing. David Knap
Název práce:	Modulární anonymizér streamu dat
Obor:	Bezpečnost a informační technologie

Datum vytvoření: 10. 6. 2020

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 4:*
1. Splnění zadání	*1=zadání splněno,* *2=zadání splněno s menšími výhradami,* *3=zadání splněno s většími výhradami,* ***4=zadání nesplněno***

Popis kritéria:
Posuďte, zda předložená ZP dostatečně a v souladu se zadáním obsahově vymezuje cíle, správně je formuluje a v dostatečné kvalitě naplňuje. V komentáři uveďte body zadání, které nebyly splněny, posuďte závažnost, dopady a případně i příčiny jednotlivých nedostatků. Pokud zadání svou náročností vybočuje ze standardů pro daný typ práce nebo student případně vypracoval ZP nad rámec zadání, popište, jak se to projevilo na požadované kvalitě splnění zadání a jakým způsobem toto ovlivnilo výsledné hodnocení.

Komentář:
Zadání nebylo splněno. Konkrétněji z bodů zadání:

Cíl: Cílem práce je navrhnout modulární architekturu a vzorovou implementaci anonymizéru, který zajistí odstranění citlivých údajů ze vstupního streamu dat na základě zadaných vzorů.

Nesplněno: Výsledný software nijak nepracuje se vstupními streamy a neodstraňuje citlivé údaje na základě zadaných vzorů. Namísto toho dokáže zpracovat CSV soubory s ukázkou dat a odstranit z nich data v pevně zakódovaných sloupcích.

1. Porovnejte možné přístupy k anonymizérům založeným na vyhledávání dat podle vzorů.

Nesplněno: Autor se existujícím anonymizérům nebo možným přístupům v práci nijak nevěnuje. Částečně oblast pokrývá rešerše vyhledávacích algoritmů, ale autorem zvolené a doporučené algoritmy slouží k vyhledání konkrétních podřetězců v textu, nikoliv k vyhledání vzorů.

2. Prozkoumejte vzorky telekomunikačních dat dodané vedoucím práce a navrhněte vhodnou detekci obsažených citlivých údajů.

Splněno s většími výhradami: Autor práce v textu prohlašuje, že cit. „vzorky dat neobsahují (...) nebo jiné přímé identifikátory, které by šlo vyhledávat pomocí algoritmu pro hledání vzorů“. Přesto posléze nějaká, blíže neurčená data z vzorků vylučuje v rámci své implementace, která nepracuje na principu vzorů. Návrh vhodné detekce částečně pokrývá výše zmíněná rešerše vyhledávacích algoritmů.

3. Navrhněte architekturu vlastního anonymizéru s ohledem na rychlost zpracování a modularitu vstupního formátu, výstupního formátu a konfigurace předloh dat k anonymizaci.

Nesplněno: Součástí práce není žádný návrh či diskuse softwarové architektury ani v podobě jednoduchého diagramu, pouze stručný popis implementované funkcionality. Součástí textu není ani diskuse vhodných technologií, není uveden použitý programovací jazyk ani žádný úryvek kódu. Modularita je řešena částečně, ale i v centrální části programu je například omezení na pouze 58 sloupců vstupu podle ukázkových dat. Konfigurace předloh dat nebyla implementována, autorův software s předlohami dat a vzory nijak nepracuje.

4: Pro zvolenou architekturu připravte prototypovou implementaci s alespoň jedním modulem každého typu a anonymizér vyzkoušejte na dodaných datech.

Splněno s většími výhradami: Autor připravil implementaci, která zpracuje právě ty datové soubory se vzorovými daty, které obdržel, a žádné jiné.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*15 (F)*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části. Dále posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře. Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3. Posuďte, zda student využil a správně citoval relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami. Zhodnoťte, zda převzatý software a jiná autorská díla, byly v ZP použity v souladu s licenčními podmínkami.

Komentář:
Po formální stránce co do rozsahu má práce 45 obsahových stran textu, neobsahuje žádné obrázky, tabulky ani grafy; psána je spíše snadnějším jazykem („nelze se spolehnout, že Regex nespadne do exponenciální složitosti“) a obsahuje větší množství překlepů a gramatických chyb. Matematická vyjádření jsou uvedena v prostém textu bez použití matematické notace („2,8 x 10 na 18“). Citované literatury je dostatek, byť některé, jako například webové stránky Regionálního muzea v Žatci či články o starověkém Egyptě se nezdají být příliš tematicky relevantní. Převzaté části textu jsou zpravidla korektně označené a doplněné odkazem do bibliografie. Zejména v první půli chybí v práci křížové odkazy do ostatních kapitol, a čtenář tak zůstává napnut, zda bude použitá zkratka později vysvětlena a tvrzení podpořeno nějakým argumentem.

Po obsahové stránce práce z drtivé většiny sestává z rešeršní části, praktické realizaci jsou věnovány tři stránky na konci. U odborné závěrečné práce předpokládám formulaci nějakých vlastních myšlenek, hypotéz a tvrzení, které jsou následně podpořené či vyvrácené relevantními zdroji či vlastním výzkumem. Bohužel autor v této práci postupuje opačně: ze zhruba sedmdesáti kapitol jich asi 26 je ve formátu dlouhého, víceřádkového citátu či definice z cizího zdroje, následováno uvozením „to znamená, že“ a přeformulováním stejné věci autorovými slovy. To je v pořádku u části práce, která se zabývá legislativními požadavky, ale autor stejným způsobem přistupuje také k popisu procesů anonymizace, diskutovaným algoritmům či definicím různých typů dat. Značná část práce tak zcela postrádá autorův názor či pohled na věc, a stává se bohužel jen kompilátem citací z jiných zdrojů.

Následující hodnocení je poměrně obsáhlé, ale psané s cílem poukázat na konkrétní problémy, které se podílely na výsledném hodnocení.

V první části autor předkládá cíle práce. Za zmínku stojí, že zde autor vytyčuje za cíl „mazat ty části dat, které budou definovány na vstupu“. Tento vlastní cíl však autor později opouští.

Druhá kapitola obsahuje samotnou rešeršní práci a je stěžejní částí práce.

Prvně se zabývá rozborem, co je to anonymizace dat, jaké jsou možné úrovně anonymizace a jaké pro tyto úrovně plynou důsledky. V této části bych jako vedoucí práce podle litery zadání očekával průzkum existujících řešení anonymizérů, jaké úrovně anonymizace nabízejí a jaké výhody či nevýhody který z nich má. Bohužel místo toho je tato část vesměs zmíněným kompilátem citací a definicí s jen stručnou poznámkou, zda je technika vhodná pro vzorová data či ne. Na druhou stranu jsou tu také zmíněny některé zajímavé příklady z reálné praxe, kdy došlo vlivem špatně nastavených korporátních procesů k masivním únikům dat a jaké to mělo pro dotčené firmy důsledky.

Autor dále pokračuje přehledem algoritmů pro hledání vzorů; ovšem většina algoritmů, které pro porovnání zvolil jsou místo toho určené pro vyhledávání konkrétních podřetězců a pro účely práce jsou tedy nevhodné: autor volí jako nejlepší řešení „Aho-Corasick algoritmus s předem připraveným (...) slovníkem hodnot“ z důvodu jeho lineární složitosti. I přesto, že v pozdější části práce autor určuje jako citlivá data například jména fyzických osob, nijak se nezamýšlí nad tím, jak vytvořit konečný slovník všech ve světě možných jmen fyzických osob, ani jak náročná taková operace může být. Na druhou stranu koncept regulárních výrazů (v práci uveden jen jako knihovna Regex pro C++) autor pro jeho časovou složitost zavrhuje bez větší diskuse jako nevhodný.

Následuje detailní popis druhů dat z nejrůznějších pohledů dělení, opět bohužel spíše ve formátu dlouhých citací než jednoduchého a přehledného rozdělení; a dále historické okénko do mladšího paleolitu a do války mezi Aténami a Spartou, které se mi ale k tématu streamového anonymizéru nezdá příliš relevantní.

Další část se věnuje vzorovým souborům poskytnutým autorovi z průmyslové praxe zadavatele, vesměs se jedná o útržky telekomunikačních dat zachycených na síťových prvcích, vhodně transformovaných pro testovací účely. V této části autor bez větších okolků uvádí, že v datech žádná citlivé údaje pro vzorové vyhledávání nejsou, a proto to nebude dělat. „Vzorky dat neobsahují žádná jména a příjmení, e-mailové adresy nebo jiné přímé identifikátory, které by šlo vyhledávat pomocí algoritmu pro hledání vzorů. (...) Dodaná data proto nebudu analyzovat z pohledu hledání vzorů. (...) Pro zpracování citlivých údajů tak bude nejlepší zadat programu sloupce, které z něj má odstranit.“

Poskytnutá data nicméně údaje vhodné na hledání shod se vzory obsahují, na první pohled viditelné údaje zahrnují především telefonní čísla a interní a externí IP adresy, po drobné rešerši o telekomunikačních datech se dají snadno poznat také dobře definované identifikátory SIM karet (IMSI) či různé textové řetězce potenciálně identifikující polohu uživatele („Praha_Brezineves“) nebo typ jeho přístroje. Vzhledem k tomu, že komise pro státní závěrečnou zkoušku pravděpodobně nebude mít tato data snadno k dispozici a nikde v práci není uveden žádný jejich úryvek pro představu doplňuji, že například datový soubor data3 v příloze práce začíná hned v prvním řádku hodnotami „420604099135, 10.49.32.250, ...“, následované větším množstvím technických dat.

V této části tedy autor deklaruje, že vzhledem k údajné absenci dat vhodných pro strojové testování proti vzorům bude pouze odstraňovat ze vstupu zadané sloupce bez ohledu na jejich obsah. Jaké sloupce to budou a jakým způsobem k nim autor došel, není v práci popsáno.

Po tomto prohlášení autor pokračuje rešerší legislativních požadavků, kde mimo jiné hned o čtyři strany později v kapitole

2.6.1.2 říká „Protože se (...) počítá mezi osobní údaje i síťový identifikátor, je nutné považovat za osobní údaj IP adresu fyzické osoby“, resp. „Mezi přímé identifikátory patří (...) telefonní číslo nebo číselný identifikátor“. Bohužel se po této rešerši už autor nevrací zpět, aby revidoval svůj předchozí názor.

Zbytek rešeršní části je tvořen podrobnou analýzou relevantních zákonů, která sice nebyla požadována a opět je tvořena převážně citacemi, ale vpravdě popisuje vše důležité, co je k tématu potřeba vědět.

Následuje už jen třístránkový popis realizovaného řešení s velmi zevrubným popisem kroků programu, bez technických detailů, ilustrací či návrhových vzorů.

Autor se tak zcela odklonil od textu zadání a ve výsledku se tak věnuje úplně jinému scénáři použití, než jaký byl předepsán.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*0 (F)*

Popis kritéria:
Dle charakteru práce se případně vyjádřete k nepísemné části ZP. Například: SW dílo – kvalita vytvořeného programu a vhodnost a přiměřenost technologií, které byly využité od vývoje až po nasazení. HW – funkční vzorek – použité technologie a nástroje, Výzkumná a experimentální práce
– opakovatelnost experimentů

Komentář:
Po obsahové stránce je vytvořená aplikace k tématu irelevantní, zadaný úkol neřeší. Aplikace nedetekuje citlivá data na základě konfigurovatelných vzorů, místo toho zpracuje vzorky dat vymazáním celých sloupců v konkrétních třech souborech a žádné jiné zpracovat neumí. Ani zmiňovaná modularita není ideální, bez zásahu do univerzálního dispečeru není například možné načítat soubory s více než 58 sloupci.

Po stránce formy je kód čitelný a vhodně doplněný komentáři pro generování technické dokumentace. Rozsah kódu je pro závěrečnou práci hrubě podprůměrný, ale odpovídá funkcionalitě navržené autorem. Pro tuto funkcionalitu je však C++ zbytečně komplexní jazyk, odstranění zadaných sloupců ze souboru by bylo možné například implementovat jedním příkazem v jazyce AWK.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Hodnocení výsledků, jejich využitelnost	*0 (F)*

Popis kritéria:
Dle charakteru práce zhodnoťte možnosti nasazení výsledků práce v praxi nebo uveďte, zda výsledky ZP rozšiřují již publikované známé výsledky nebo přinášející zcela nové poznatky.

Komentář:
Výsledek práce není pro pro praxi použitelný, zadání nebylo splněno.

Hodnotící kritérium:	*Způsob hodnocení – následující škálou 1 až 5:*
5. Aktivita a samostatnost studenta	5a: *1=výborná aktivita,* *2=velmi dobrá aktivita,* ***3=průměrná aktivita,*** *4=slabší, ale ještě dostatečná aktivita,* *5=nedostatečná aktivita* 5b: *1=výborná samostatnost,* *2=velmi dobrá samostatnost,* ***3=průměrná samostatnost,*** *4=slabší, ale ještě dostatečná samostatnost,* *5=nedostatečná samostatnost*

Popis kritéria:
V souvislosti s průběhem a výsledkem práce posuďte, zda byl student během řešení aktivní, zda dodržoval dohodnuté termíny, jestli své řešení průběžně konzultoval a zda byl na konzultace dostatečně připraven (5a). Posuďte schopnost studenta samostatné tvůrčí práce (5b).

Komentář:
Autor zpracovával téma samostatně, na mé e-maily a výzvy ke konzultaci nereagoval, první verze se ke mě jako vedoucímu dostala až tři dny před termínem odevzdání. To vidím jako pravděpodobný důvod nesplnění zadání – při včasných a častých konzultacích by snad byla možnost vysvětlit jakékoliv nejasnosti a nedorozumění, ať šlo o zadání nebo dodaná vzorová data.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Celkové hodnocení	*15 (F)*

Popis kritéria:
Shrňte stránky ZP, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení nemusí být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích. Obecně platí, že bezvadně splněné zadání je hodnoceno klasifikačním stupněm A.

Text hodnocení:
Nedostatečná rešerše a absence konzultací s vedoucím práce vedly zřejmě k nedorozumění či nepochopení zadání, kvůli čemuž se autor v průběhu práce zcela odklonil od vytyčeného tématu a zadání ani rámcově nesplnil. Vzhledem k tomu, že praktická část řeší zcela jinou úlohu a ani rešeršní část se hlavním částem zadání příliš nevěnuje, doporučuji práci ke kompletnímu přepracování. Práci musím hodnotit stupněm F – Nedostatečně, a k obhajobě ji tak nedoporučuji.

Podpis vedoucího práce: