FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student:	Bc. Marek Mařík
Oponent práce:	Ing. Karel Hynek
Název práce:	Identifikace webového obsahu v šifrovaném provozu
Obor:	Znalostní inženýrství

Datum vytvoření: 5. 1. 2021

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 bylo splněno v plném rozsahu.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
2. Písemná část práce	*75 (C)*

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:
Text práce je napsaný poměrně srozumitelně. Během jeho čtení jsem si nevšiml žádných překlepů ani závažných typografických chyb. V některých větách pouze chyběly, nebo naopak přebývaly čárky. Rovněž chválím vkládání referencí přímo doprostřed věty, takže je jasné, k čemu se vztahují. S členěním textu příliš spokojený nejsem a výrazně bych jej pozměnil. Části, které patří do analýzy (např. představení použitého softwaru), jsou uvedeny v kapitole zabývající se realizací. Naopak v kapitole s rešerší se dozvídáme důležitá návrhová rozhodnutí. Místy jsou používána dlouhá souvětí, která se nečtou moc dobře a je jednoduché se v nich ztratit. Rovněž jsem si všiml neobratných výrazů, které nejsou v odborných textech obvyklé (neuronové sítě zažívají v posledních letech boom). Celkově text hodnotím jako průměrný.

Hodnotící kritérium:	*Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):*
3. Nepísemná část, přílohy	*85 (B)*

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:
Autor vytvořil dva hlavní nepísemné výstupy diplomové práce. 1) Systém pro automatickou tvorbu datové sady je pěkně navržený a dle mých testů i naprosto funkční. Mám pouze výhrady k jeho dokumentaci, která je nedostatečná. V rámci výstupů diplomové práce by měla být rozumně vyplněná dokumentace v doxygenu samozřejmostí. 2) Analyzátor šifrované webové komunikace je implementovaný v Jupyter notebooku, což považuji za dostatečné, protože se jedná o prototyp. I zde by neuškodilo přidání komentářů a více vysvětlujících prvků. Vytvořený notebook s experimenty je velice dlouhý a je tedy těžké se v něm orientovat. Uvítal bych tedy i častější používání Markdown polí.

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

4. Hodnocení výsledků, jejich využitelnost	*95 (A)*

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:

Generátor datové sady je perfektní a bude jistě užitečný v dalším výzkumu Laboratoře monitorování síťového provozu. Rovněž experimenty prováděné na vytvořené datové sadě jsou velice zajímavé. Odborná literatura se zatím příliš nevěnuje rozpoznávání jednotlivých stránek v rámci jednoho webu, a proto je přínos práce značný. Pouze mě mrzí, že se autor omezil na základní algoritmy strojového učení, které mu nabízí knihovna sci-kit learn a nepoužil složitější ensemble metody. Zejména kombinace několika různých algoritmů se v problematice rozpoznávání webových stránek často používá. Nicméně i tak jsem přesvědčen, že výsledky diplomové práce mohou sloužit jako základ budoucí odborné publikace.

Hodnotící kritérium:	*Způsob hodnocení – nehodnotí se*

5. Otázky k obhajobě

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:

Lze z používaných charakteristik určit i typ prohlížeče případně operační systém, ze kterého byla webová stránka načtena?

Zkoušel jste při odhadování optimálního počtu charakteristik (RFECV) použít i jiný klasifikátor/regresor než Random Forest? Pokud ano, pozoroval jste nějaké významné rozdíly?

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

6. Celkové hodnocení	*85 (B)*

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:

Celkové hodnocení práce je kladné. Autor si dokázal samostatně dostudovat problematiku monitorování síťového provozu a dokázal vytvořit systém pro automatické generování datové sady. Dále vytvořil datovou sadu, na které prováděl experimenty a trénoval modely strojového učení. Vytvořené experimenty dokazují, že je možné s poměrně vysokou přesností rozpoznat jakou stránku (přesný název článku na wikipedii) si uživatel prohlíží bez nutnosti dešifrování samotné komunikace. Rovněž ukázal, že je možné s rozumnou přesností zjistit i informace o navštívené stránce, jako je například počet obrázků. Bohužel samotný text silně kazí celkový dojem z práce, a proto ji hodnotím známkou B.

Podpis oponenta práce: