FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Hodnocení vedoucího závěrečné práce

Vedoucí práce:	Mgr. Lukáš Bajer
Student:	Olena Marchenko
Název práce:	Kontextuálni pasivní DNS prediktor
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	24. srpna 2021

Hodnotící kritéria

1. Splnění zadání

[1] zadání splněno
▸ **[2] zadání splněno s menšími výhradami**
[3] zadání splněno s většími výhradami
[4] zadání nesplněno

Předložená práce "Contextual Passive DNS Resolution" se zabývá problémem odhadu použitého doménového jména (DNS) v situacích, kdy v síťové telemetrii není toto jméno přítomno a je k dispozici pouze informace o IP adrese a další standardní telemetrické informace. Práce se drží navrženého zadání a navrhuje kokrétní systém využití těchto dodatečných informací pro predikci doménového jména, a to jak pro případ predikce jmen druhého řádu (SLD), tak celého doménového jména (hostname). V závěrečné části je pak stručně splněno i rozšíření úlohy o predikci rizika.

Oproti zadání práce ani její příloha neobsahuje odkaz na použitelný veřejný dataset (použitelný dataset nebyl nalezen studetem ani vedoucím) a nebyl dodán ani vygenerovaný zjednodušený syntetický dataset, který by umožnil jednoduché ověření dodaných kódů. Toto odchýlení od zadání bylo se mnou konzultováno v průběhu a považuji za opodstatněné z důvodu množství energie strávené se zpracováními poměrně velkých dat dodaných vedoucím práce.

2. Písemná část práce 70 /100 (C)

Odevzdaná práce po formální i obsahové stránce splňuje základní požadavky na diplomovou práci. Práce je smysluplně rozdělena do čtyř kapitol s vhodnou typografickou úpravou, cituje použité zdroje, kterými jsou většinou vědecké články nebo monografie. Práce je na vhodných místech doplněna tabulkami nebo diagramy, až na dvě výjimky (str. 36 a 38) neobsahuje nedostatky ve formálních nebo semi-formálních zápisech a definicích.

Práce je z velké většiny napsána správným anglickým jazykem (nakolik mohu posoudit), s pochopitelným množstvím standardních překlepů. Nicméně obsahuje také věty se špatným slovosledem nebo chybějícími či přebývajícími větnými členy; nejčastější chybou je pak zřejmě zdvojení členu "the the most/best".

Teoretická část práce je poměrně dobře srozumitelná. Některé části by zřejmě bylo možné zkrátit a odkázat čtenáře na relevantní literaturu (např. 1.3.2 o útocích na DNS, nebo část 2.5.3 o MLP), respektuji ale volbu autorky. Po slohové stránce by bylo možné zlepšit propojení mezi teoretickou částí a praktickým problémem/daty v kapitole 3, popis navrženého řešení v kapitole 3.3 by také dle mého názoru zlepšil obecnější úvod.

Coby vedoucí práce jsem spokojen s rozsahem experimentů (parameter search, stratified subsampling), s typem navrženého řešení, extrakcí příznaků (feature extraction) i výsledného klasifikátoru. Výsledky tzv. baseline modelu (viz str. 40) odpovídají nezávislým výsledkům naměřených na jiných datech, a výsledky navrženého přístupu přinášejí zajímavé zlepšení.

Jako slabší článek práce hodnotím popis experimentů a zpracování jejich výsledků. Studentka nevhodně použila třídy 0 a 1 na str. 46 opačně, než je ve strojovém učení obvyklé. To má patrně za následek zavádějícím způsobem nízké hodnoty recall/precision v následujících hodnoceních, např. tabulce 3.13 nebo odst. 3.5.1.2. Uvítal bych také přehled výsledků výběru parametrů a porovnání s finálním výběrem modelu. V některých obratech by bylo lepší přesnější vyjádření o skupinách dat, na kterých byla prováděna která operace (parameter tunning, škálování hodnot, viz např. předposl. odstavec str. 32).

Zpracování výsledků postrádá diskuzi či pokus o jejich interpretaci a navázání na použítá data a metody -- zpravidla nejhodnotnější části vědeckých publikací. Delší komentář by si např. zasloužil dramatický subsampling v sekci 3.5.2. nebo velmi optimistické PR křivky v obrázku 3.13. Diskuze by pak možná odhalila velký rozdíl mezi binární accuracy a precision/recall výsledky v obr. 3.10.

3. Nepísemná část, přílohy

75 /100 (C)

Práce obsahuje zdrojové kódy pro extrakci příznaků (PySpark) a klasifikační modely. Vlastní trénování modelů a parameter tunning kolekce zdrojových kódů neobsahuje, opakovatelnost experimentů je tedy do jisté míry omezena. Některé části nejsou přiloženy mj. i z důvodu spolupráce s komerční firmou a možnými odkazy na jejich zdrojové kódy, obecnější kód by asi přiložit bylo možné.

Studentka prokázala schopnost experimentální práce v jazyku Python, stejně jako zpracování výsledků. Jako možnosti zlepšení do budoucna vidím zejména v systematičnosti experimentální práce, správě a systematickém ukládání výsledků z experimentů, verzování experimentů, měření všech podstatných veličin nákladných výpočtů apod. Tuto část nepísemná část práce také nepokrývá, hodnotím tedy spíše z průběhu práce během roku.

4. Hodnocení výsledků, jejich využitelnost

90 /100 (A)

Nové klasifikační modely, které byly v rámci práce navrženy a zpracovány, přinášejí znatelné zlepšení oproti současnému stavu (viz "baseline model" popisovaný v práci). Z definice úlohy a dodaných testovacích/trénovacích dat sice neumožňují přímé nasazení,

to však nebylo předmětem práce. Výsledky ukazují nové koncepty a jejich slibné výsledky nabízí další rozpracovaní pro reálné nasazení.

5. Aktivita studenta

[1] výborná aktivita
▸ **[2] velmi dobrá aktivita**
[3] průměrná aktivita
[4] slabší, ale ještě dostatečná aktivita
[5] nedostatečná aktivita

Studentka pracovala podle dohodnutého schématu a na konzultace přicházela připravená. Do její budoucí praxe by bylo přínosem, kdyby zlepšíla schopnost prezentovat své vlastní řešení a rozmýšlela řešení o několik kroků dopředu oproti zcela následujícím krokům.

6. Samostatnost studenta

[1] výborná samostatnost
[2] velmi dobrá samostatnost
▸ **[3] průměrná samostatnost**
[4] slabší, ale ještě dostatečná samostatnost
[5] nedostatečná samostatnost

Studentka je schopná samostatné práce na předem dohodnutých krocích a samostatně implementovala veškerá řešení. Do budoucna by pomohlo, pokud by dokázala potřebné kroky více sama navrhovat a přicházet s různými variantami řešení.

Celkové hodnocení 75 /100 (C)

Diplomová práce přináší zajímavý výsledek a jedno konkrétní řešení daného problému, který může být po úpravách převeden do praxe. Práce je zpracována dobře, studentka odvedla netriviální množství práce a prokázala schopnost zpracovat dané téma, včetně seznámení a použití nových technologií (v tomto případě Apach Spark) a práce s daty z praktické úlohy. Práci doporučuji k obhajobě, kvůli některým nedostatkům popsaným výše navrhuji hodnotit známkou 2.

Instrukce

Splnění zadání

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.

Písemná část práce

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. 52/2021, 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.

Nepísemná část, přílohy

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.

Hodnocení výsledků, jejich využitelnost

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.

Aktivita studenta

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.

Samostatnost studenta

V souvislosti s průběhem a výsledkem práce posuďte schopnost studenta samostatné tvůrčí práce.

Celkové hodnocení

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.