



Posudek oponenta závěrečné práce

Student: Bc. Jakub Hynek
Oponent práce: prof. Ing. RNDr. Martin Holeňa, CSc.
Název práce: Zpracování toků krátkých textů v reálném čase
Obor: Teoretická informatika

Datum vytvoření: 21. 1. 2021

<i>Hodnotící kritérium:</i>	<i>Způsob hodnocení – následující škálou 1 až 4:</i>
1. Splnění zadání	<i>1=zadání splněno, 2=zadání splněno s menšími výhradami, 3=zadání splněno s většími výhradami, 4=zadání nesplněno</i>
<i>Popis kritéria:</i> 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.	
<i>Komentář:</i> Zadání bylo splněno s výjimkou posledního bodu, v němž analýza kvality shluků byla provedena zcela nedostatečně. Diplomant se omezil na výpočet rozptylu shluků, z běžně používaných měř kvality shluků, jako je siluetová, mezerová, Calinského-Harbaszova nebo Davies-Bouldinova, nevyhodnotil ani jedinou.	
<i>Hodnotící kritérium:</i>	<i>Způsob hodnocení – bodové hodnocení 0 až 100 bodů (známka A až F):</i>
2. Písemná část práce	<i>65 (D)</i>
<i>Popis kritéria:</i> 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 odlišený 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:

S písemnou částí práce nejsem spokojený. Má řadu nedostatků jak obsahových, tak formálních.

Po obsahové stránce úroveň práce snižuje zejména:

1. Analýza výsledku shlukování byla provedena zcela nedostatečně (viz 1. Splnění zadání).
2. Charakterizace přístupu t-SNE používá pojem „vzájemné pravděpodobnosti“. Nikde ale není vysvětleno, co je tímto pojmem myšleno.
3. Tokenizace se odvolává na ekvivalentnost informací, není ale vysvětleno, v čem jsou ekvivalentní.
4. algoritmu ClusTree je prostě sdělení skutečnosti, že používá 1. a 2. empirický moment, nahrazeno velmi složitým vysvětlením v délce celého odstavce.
5. Obsah podsekcce 2.2.10 zcela chybí.
6. Podsekcce 2.2.1 by podle názvu měla obsahovat vysvětlení všech základních pojmů, je v ní však vysvětlen jen jediný pojem.
7. Teoretické části není u klíčové metody Word2Vec citován článek, ve kterém byla navržena. Místo toho u ní diplomant cituje Tensorflow, ve kterém je implementována.

Po formální stránce mně na práci vadí zejména:

8. Mnoho odborných termínů ponechal diplomant v angličtině, přestože k nim existují české ekvivalenty (např. machine learning; stochastic neighbor embedding; stop words; cluster; self-organizing-maps; outlier; buffer; kód je open source; set zpráv).
9. Přestože práce je psána česky, tedy v mateřštině diplomanta, občas se v ní vyskytují formulace, které gramaticky nedávají smysl a jejich význam si čtenář tudíž musí domýšlet (např. není čitelný běžným člověkem, natož s ním interagovat; bez vědomostí těchto parametrů).
10. Nadto se v ní poměrně často používají hovorové obraty, nehodící se do odborného textu (např. leda odbornou; masivní publikum; kus informace; příbytek datových bodů; když má času nazbyt).
11. Text v obr. 2.1 není česky.
12. Práce obsahuje značné množství překlepů.

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:

Ačkoliv vytvořený prototyp není použitelný v souladu s původním záměrem práce, kterým byla analýza sentimentu ve zprávách, oceňuji úsilí, které diplomant jeho vytváření věnoval, zvláště v kontextu toho, že se musel potýkat s velkým množstvím problémů z oblasti zpracování přirozeného jazyka, k čemuž měl na FITu mnohem menší podporu než mají magisterští studenti na UFAL MFF UK.

Hodnotící kritérium:

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

4. Hodnocení výsledků, jejich využitelnost

75 (C)

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:

Vytvořený prototyp bohužel není využitelný v souladu s původním záměrem práce, kterým byla analýza sentimentu ve zprávách, viz komentář k 3. Nepísemná část, přílohy. Věřím však, že pomohla odbornému růstu diplomanta, zvláště pokud se někdy bude zabývat zpracováním přirozeného jazyka.

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:

1. Proč si myslíte, že závěry týkající se nastavení parametrů získané při statickém shlukování mohou být přeneseny na shlukování proudových dat?
2. Klíčovou roli ve shlukování hraje použitá metrika. Jak v práci správně píšete, zda daná funkce 2 proměnných je metrikou, závisí na tom, zda splňuje axiomy metriky. Ty axiomy jste ale v práci nepřipomněl. Můžete je připomenout aspoň dodatečně?

Hodnotící kritérium:

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

6. Celkové hodnocení

83 (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:

S písemnou částí práce nejsem spokojen, ale diplomant na mě při schůzce kvůli nejasnostem, které jsem k ní měl, udělal velmi dobrý dojem. Domnívám se, že nízká úroveň výsledku jeho práce je do značné míry důsledkem toho, že se týkala oblastí zpracování přirozeného jazyka a analýza sentimentu, které jsou vzdálené náplni jeho studia na fakultě. Pokud však v něm vzbudila zájem o tyto rychle se rozvíjející oblasti, pokládám hodnocenou diplomovou práci za přínosnou.

Podpis oponenta práce: