FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Ing. Tomáš Kalvoda, Ph.D.
Student:	Michal Melko
Název práce:	Automatická analýza změn v datech s využitím knihovny great expectations
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	10. června 2024

Hodnotící kritéria

1. Splnění zadání

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

Zadání práce je psáno obšírnějším stylem, ale pokud extrahuji jeho hlavní body, tak se dostávám k následujícímu konstatování:

* Požadavek implementace profileru nesplněn (důvodem je jeho existence v knihovně Great Expectations (GX) až v době po vytvoření zadání, to se může stát; očekával bych v takovém případě alespoň výstižný popis použití a možností profileru z knihovny GX).
* Implementace sady rozšíření knihovny GX pro detekci data driftu: implementováno jedno rozšíření.
* Požadavek ukázky použití na alespoň dvou datasetech: práce uvádí výsledky experimentů na uměle vygenerovaných datasetech, ale ukázka použití samotného rozšíření v rámci knihovny GX chybí.

2. Písemná část práce 45 /100 (F)

Hlavní text práce zabírá prakticky 26 stránek (včetně obrázků) a je na mnoha místech velmi stručný.
Rozsahem z mého pohledu naráží na dolní hranici požadavků kladených na bakalářskou závěrečnou práci.
Konkrétněji mám následující připomínky k jednotlivým částem:

* První odstavec kapitoly Introduction je spojení prvních tří odstavců zadání! To představuje obsahově podstatnou část této kapitoly. Sekce "Thesis overview" je v podstatě stručná verze obsahu bez jakéhokoliv komentáře.
* Druhá kapitola, popisující rešerši existujících řešení zabírá pouhé dvě stránky. Určitě by

šlo přidat i informace například o rozšířenosti/oblíbenosti jednotlivých nástrojů, jakým způsobem jsou distribuovány, pod jakou licencí, možná se podívat i mimo Python ekosystém (Julia).
* Čtvrtá kapitola popisující různé experimenty a metody mi přijde nejpovedenější. Trochu mi ovšem chybí rozbor (nebo alespoň odkaz do přílohy) zdrojového kódu. Jaké nástroje student použil, jak může čtenář experimenty reprodukovat? V jisté míře na tyto otázky odpovídá Readme.md v příloze, ale to není dostatečné.
* Pátá kapitola, mající se zabývat implementací, je prakticky prázdná! Najdeme zde pouze konstatování, že knihovna GX již obsahuje data profiler a proto tento nebudeme implementovat (chápu důvod, funkcionalita byla v knihovně implementována až po vytvoření zadání). O samotném rozšíření knihovny GX se zde ale nic konkrétního nedočteme. Nabízí se řada otázek: co tedy student implementoval; jak je možné jeho implementaci využít; kde je kód k dispozici; vhodná by byla ukázka použití,... Tato kapitola má jednu jedinou stránku! Bez zkoumání přílohy na tyto otázky nelze zodpovědět.

K práci mám dále několik věcných poznámek. Vysvětlení pojmu "data drift" pomocí pravděpodobnosti je mlhavé a bez podrobnějšího zavedení notace a popisu kontextu pro čtenáře těžko pochopitelné. Podobně tomu je u následného popisu termínu "concept drift". Tato kapitola, která by měla být výchozím bodem pro následující úvahy, se odvolává jenom na jeden článek. Vzhledem k očividné důležitosti této problematiky mi přijde nepravděpodobné, že by teorie potřebná k uchopení problému byla tak nerozpracovaná.

V práci se vyskytuje několik dalších nepřesností:

* V druhé rovnici na str. 14 má být zřejmě proměnná x__2, místo x__1.
* Seznam výpisů kódů ("List of code listings" na str. iv) neobsahuje žádné výpisy kódu.
* Pod rovnicí v sekci 3.1.1 na str. 11 (nebo i 3.1.2, 3.1.3 a jinde) nemá začínat další odstavec (odražení).
* Odkazy na rovnice jsou pravděpodobně sázeny pomocí makra \ref (bez uzávorkování) místo správného \eqref (s uzávorkováním).
* Výraz pro maximum v matematické formuli je vhodné sázet pomocí makra \max k tomu určenému, ne jako "max", k čemuž se LaTeX chová jako k součinu tří proměnných. Autor občas zapomíná matematická prostředí (m a n v sekci 3.1.1).

Práce je doplněna skromnějším seznamem literatury.
V rámci textu se student na původní zdroje odkazuje, ale najdou se i mezery (např. zdroj pro sekci 3.1.2 není uveden).
Nemohu se ubránit pocitu, že autorova rešerše existující literatury nebyla dostatečně důkladná.

Práce je psána anglicky na poměrně dobré úrovni. Občas se autor nevyvaruje chybějící interpunkci, některé pasáže by šlo formulovat lépe, ale obecně čtenář nemá problém pochopit popisovanou myšlenku.

3. Nepísemná část, přílohy 75 /100 (C)

Nepísemná část práce, která bohužel v textové části práce není vůbec rozebrána/ zmíněna, obsahuje dva hlavní výsledky:

* Rozšíření Python knihovny Great Expectations (124 řádek Python kódu) v souboru začínajícím komentářem "This is a template for creating custom BatchExpectations".
* Skripty pro generování experimentálních dat a jejich vyhodnocení (popsáno v práci v Kapitole 3; 207 řádků Pythonu a Jupyter notebook).

Kód je funkční, Jupyter notebook je poměrně strohý, bez doplňujícího komentáře.

4. Hodnocení výsledků, jejich využitelnost 65 /100 (D)

Hlavní výsledek, rozšíření knihovny Great Expectations (GX), zcela jistě využitelné je. Z textu práce není jasné, jestli je výsledné rozšíření snadno použitelné pro aktuální uživatele GX.

Celkové hodnocení 45 /100 (F)

Mé celkové hodnocení výrazně sráží textová část práce, z které je cítit velká nedokončenost (viz 2. sekce tohoto posudku). Hned několik kapitol je obsahově velmi řídkých. Věřím, že pokud bude student moci věnovat dodatečný čas textové části práce (a případně ještě i té netextové části), tak by nakonec mohlo jít o pěknou bakalářskou práci. Komisi navrhuji v tento okamžik hodnotit práci známkou F.

Otázky k obhajobě

Testoval jste vaše rozšíření i na nějakých reálných větších datech, než jenom na menších synteticky připravených datasetech?

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.

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.