FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Oponent práce:	Mgr. Petr Novák, Ph.D.
Student:	Milan Vu
Název práce:	Extrakce strukturovaných dat z českých faktur
Obor / specializace:	Znalostní inženýrství
Vytvořeno dne:	23. 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

2. Písemná část práce 70 / 100 (C)

Práce je psaná srozumitelným jazykem a je přehledně členěná. Část věnovaná rešerši stávajících metod je relativně rozsáhlá a někdy zabíhá do přílišných detailů (např. u preprocessingu a u klasifikačních metod), což se odráží i v zbytečně detailním členění sekcí (např. 2.4.1.1). Není pak zcela vysvětlené, které z představených metod byly vybrány pro samotnou implementaci a proč (např. proč byl zvolen Tesseract pro OCR). Analýza výsledků by zasloužila detailnější rozbor. Chybí aspoň hrubé srovnání s dostupnými metodami, také interpretace zvolených optimálních hyperparametrů v části 6.2.2 a skóre důležitosti v části 6.2.3.

3. Nepísemná část, přílohy 70 / 100 (C)

Součástí práce je vlastní implementace prototypové aplikace pro práci s fakturami, která umožňuje jednak labeling faktur a jednak automatickou detekci dat na základě natrénovaných modelů. Aplikace je funkční a přehledná. Jak je v práci diskutováno, detekce je spíše pomalejší. Dále autor provedl analýzu porovnávající vybrané algoritmy strojového učení použité pro detekci, jejíž výstupy jsou shrnuty v písemné zprávě. Příslušné Jupyter Notebooky nebyly úplné, takže nebylo možné analýzu zreplikovat. Část experimentů byla prováděna jen na vzorku šedesáti ručně labelovaných faktur, což je může být potřeby strojového učení málo. Podle nezanedbatelného skóre důležitosti feature "martin" (tabulka 6.5) lze pojmout podezření, že použité faktury nejsou dostatečně různorodé, což může zkreslovat výsledky.

4. Hodnocení výsledků, jejich využitelnost 80 /100 (B)

Aplikace pro labelování a detekci dat z faktur je funkční, detekce je sice pomalejší, ale po dopracování uživatelského rozhraní (uživatelské účty, lokalizace, export dat, třídění faktur do složek) si lze představit použití aplikace v praxi. Analýza přesnosti detekce je stručnější, ale může sloužit jako základ pro další zkoumání problematiky.

Celkové hodnocení 75 /100 (C)

Předložená práce se zabývá metodami detekce a extrakce dat z českých faktur. Autor provedl rešerši stávajících metod, implementoval vlastní řešení a prozkoumal přesnost různých použitých metod strojového učení. Práce vykazuje výše zmíněné nedostatky, nicméně se domnívám, že splňuje požadavky kladené na bakalářskou práci a doporučuji ji za ni uznat.

Otázky k obhajobě

- Existují metody, kterými by bylo možné urychlit detekci pomocí Craft+Tesseract?
- Jak pracovat s případy, kdy faktura obsahuje IČO dodavatele i odběratele?
- Čím lze vysvětlit rozdíly v přesnosti mezi výsledky detekce pomocí Tesseractu oproti kombinaci Craft+Tesseract (tabulky 6.1 a 6.2), kdy některé položky jsou podstatně přesněji detekovány jednou metodou a jiné druhou?
- Jak lze interpretovat zápornou přesnost naivního Bayesova klasifikátoru v tabulce 6.4?

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.