Posudek oponenta diplomové práce

Autor/autorka práce: **Veronika Kutková**

Název práce: **SharePoint Add-In pro vytěžování dat z dokumentů**

Obsah práce

Práce se zabývá částečnou automatizací procesu digitalizace faktur. Autorka v teoretické části stručně uvádí čtenáře do problematiky strojového rozpoznávání (OCR) a zpracování textu a představuje platformu SharePoint, se kterou má být řešení spjato. V praktické části experimentálně srovnává 15 existujících OCR nástrojů a 8 konfigurací algoritmů pro určení míry podobnosti textů. Dále navrhuje vlastní vytěžování údajů převážně na základě regulárních výrazů. V kapitole 7 popisuje implementaci add-inu, který je komplexním modulárním řešením problému. V příloze se nachází detailní 27 stránková instalační a uživatelská příručka.

Kapitola *3.2.1 Předplatné Office 365* popisující cenový model různých variant služby dle mého názoru nesouvisí s prací, jelikož autorka poznatky neuplatňuje v žádné analýze.

Čistý text práce odpovídá 60 normostranám a vyhovuje tak požadavkům na kvalifikační práci.

Kvalita řešení a dosažených výsledků

Autorka v textu dobře využívá UML diagramy a výpisy HTTP komunikace, které usnadňují pochopení implementace. Zdrojové kódy jsou vhodně komentované.

Experiment ke srovnání OCR nástrojů je velmi dobře provedený.

K experimentu srovnávajícímu metody určení dodavatele na základě přečteného textu a adresáře dodavatelů mám výhrady. Autorka využívá při experimentu výstupy získané z OCR software, které v sobě obsahují různé množství chyb. Je tedy nemožné ve výsledku experimentu rozlišit, jaká část chyby je důsledkem nekvalitního vstupu a jaká část je způsobená nevhodným algoritmem. Autorka se sice pokouší v diskuzi toto rozlišit, nicméně preciznější by bylo používat v experimentu bezvadný vstupní text, případně provést 2 experimenty (s čistými a reálnými vstupními daty) a výsledky analyticky porovnat. Dále je nevhodné, že u 2 z porovnávaných konfigurací je jiná množina vstupních dat. Tabulka 6.2 s výsledky experimentu obsahuje i absolutní hodnoty, které jsou však kvůli tomu nesrovnatelné.

U experimentů by bylo vhodné rozhodnout i o statistické významnosti výsledků.

Při popisování kvality navržených metod na vytěžování čísel není opět možné rozlišit, do jaké míry se jedná o chybu vlastního algoritmu. Na první pohled jsou zarážející hodnoty recall 27% pro čárové kódy a 13% pro čísla objednávky. Jak však autorka dále komentuje, problémem je, že OCR špatně zpracoval čárové kódy u 73% faktur a většina [sic] faktur má čísla objednávek v jiném tvaru než bylo uvedeno ve specifikaci.

Ve zhodnocení výsledků postrádám informaci, zda byly naplněny požadavky z kapitoly 7.1 – obzvláště zrychlení rutinní činnosti účetních (7.1.1 bod 1) a uživatelů SharePoint (7.1.2 bod 3). Dále bych zde čekal důslednější popis limitací výsledného řešení. Například možnost zpracovat pouze 25 stran textu denně ve vybraném OCR (je zmíněno pouze v poznámce v kapitole 7.3.4).

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Formální úroveň

Práce má velmi dobrou formální úroveň a množství překlepů je zanedbatelné vzhledem k rozsahu práce.

Práce s literaturou

Autorka pracuje se 78 relevantními zdroji. Jedná se převážně o online zdroje (dokumentace nástrojů a algoritmů, popis technologií, zpravodaje), což je pochopitelné vzhledem k řešené problematice. Knižní zdroje jsou z oblasti účetnictví, OCR a porovnávání textů.

Splnění zadání

Zadání bylo splněno bez výhrad.

Dotazy k práci

Jakým způsobem očekáváte, že bude probíhat validace vytěžených dat?

Odhadněte množství ušetřeného času zaměstnanců ve zvolené společnosti v důsledku nasazení dokončeného řešení. Srovnejte s náklady, které jsou spojeny s provozem řešení.

Navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

V Plzni 8. 6. 2017

Ing. Martin Kryl

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky