Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Kateřina Nová
Název práce	Analysis and visualization of OCR output
Rok odevzdání	2022
Studijní program	Informatika
Specializace	Obecná informatika

Autor posudku	Jiří Mírovský	**Role**	Oponent
Pracoviště	ÚFAL		

Prosím vyplňte hodnocení křížkem u každého kritéria. Hodnocení *OK* označuje práci, která kritérium vhodným způsobem splňuje. Hodnocení *lepší* a *horší* označují splnění nad a pod rámec obvyklý pro bakalářskou práci, hodnocení *nevyhovuje* označuje práci, která by neměla být obhájena. Hodnocení v případě potřeby doplňte komentářem. Komentář prosím doplňte všude, kde je hodnocení jiné než *OK*.

K celé práci	lepší	OK	horší	nevyhovuje
Obtížnost zadání	☒	☐	☐	☐
Splnění zadání	☐	☐	☒	☐
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	☐	☒	☐	☐

Komentář Téma práce je na bakalářskou práci poměrně rozsáhlé, resp. téma je možné zpracovat do různé hloubky s různou náročností.
Ze čtyř bodů uvedených v zásadách pro vypracování autorka bezesporu splnila první dva, tedy " Shromáždit vícejazyčný vzorek zlatých dat naskenovaných textů a jejich ručních přepisů" a "Vybrat konkrétní systém OCR a analyzovat jeho kvalitu na zlatých datech" (autorka se zaměřila na tři open source systémy OCR).
Další dva body zásad pro vypracování ("Zapojit procedury počítačového zpracování přirozeného jazyka (NLP) do analýzy výstupu OCR" a "Uživatelsky přívětivě vizualizovat výstup OCR a jeho analýzu") autorka splnila spíše v menší míře.
Z procedur NLP jsou užity pouze tokenizace a POS tagging, přičemž potenciál analýzy výsledků OCR vzhledem k POS taggingu nebyl zcela vyčerpán.
Největší nedostatek práce je možno najít v posledním bodě týkajícím se vizualizace výstupu OCR - užitá vizualizace se týká jen výsledků statistické analýzy, nikoliv přímo výstupu OCR, což autorka sama stručně zmiňuje a ne příliš jasně vysvětluje v závěru. Nicméně vzhledem k celkové náročnosti zadání a dobré kvalitě jiných částí bych to nepovažoval za rozhodující v otázce možnosti práci obhájit.

Textová část práce	lepší	OK	horší	nevyhovuje
Formální úprava ... *jazyková úroveň, typografická úroveň, citace*	☐	☒	☐	☐
Struktura textu ... *kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	☐	☒	☐	☐
Analýza	☐	☒	☐	☐
Vývojová dokumentace	☐	☒	☐	☐
Uživatelská dokumentace	☐	☒	☐	☐

Komentář Po formální stránce je text práce dobře strukturován a napsán přehledně, dobrou angličtinou, byť s větším počtem přehlédnutých chyb, které ovšem nebrání srozumitelnosti.
Analýza problematiky Optical Character Recognition (OCR) je poměrně rozsáhlá, data a jejich zpracování detailně popsána, rovněž použitým nástrojům je věnován dostatečný (někdy možná ve výpisu parametrů zbytečně velký) prostor.
Pečlivě jsou popsány navržené a použité míry pro měření kvality výstupu OCR.
Text je doplněn řadou obrázků dobře ilustrujících probírané jevy. Závěr (Conclusion) mi přijde příliš stručný.
Vývojová/uživatelská dokumentace je přehledná a dostatečná.

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu ... *architektura, struktury a algoritmy, použité technologie*	☐	☒	☐	☐
Kvalita zpracování ... *jmenné konvence, formátování, komentáře, testování*	☐	☒	☐	☐
Stabilita implementace	☐	☒	☐	☐

Komentář Implementace dosahuje požadovaných kvalit s využitím standardních postupů, namátkou vybrané části kódu lze dle přiložené dokumentace bez potíží spustit.
Datová struktura je přehledná, korpus lze vytvořit pomocí Makefile.

Celkové hodnocení Dobře (spíše lepší)

Práci navrhuji na zvláštní ocenění Ne

Datum 12. srpna 2022

Podpis