Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK OPONENTA BAKALÁŘSKÉ PRÁCE

Jméno studenta: Vlasta Matějková

Název práce: Nástroje pro přípravu českého textu na strojové zpracování

Autor posudku: Ing. Martina Husáková, Ph.D.

Cíl práce: Vytvoření přehledu dostupných nástrojů pro zpracování českého textu spolu s jejich otestováním a porovnáním na reálných textech. Nalezení specifik mezi českým a anglickým textem vzhledem k vybraným metodám, které se používají pro zpracování textů.

Povinná kritéria hodnocení práce	**Stupeň hodnocení** (známka)					
	A	**B**	**C**	**D**	**E**	**F**
Práce svým zaměřením odpovídá studovanému oboru	X	☐	☐	☐	☐	☐
Vymezení cíle a jeho naplnění	☐	☐	☐	☐	X	☐
Zpracování teoretických aspektů tématu	☐	X	☐	☐	☐	☐
Zpracování praktických aspektů tématu	☐	☐	☐	X	☐	☐
Adekvátnost použitých metod, způsob jejich použití	☐	☐	☐	☐	X	☐
Hloubka a správnost provedené analýzy	☐	☐	☐	☐	X	☐
Práce s literaturou	☐	☐	☐	☐	X	☐
Logická stavba a členění práce	X	☐	☐	☐	☐	☐
Jazyková a terminologická úroveň	☐	☐	☐	☐	X	☐
Formální úprava a náležitosti práce	☐	☐	X	☐	☐	☐
Vlastní přínos studenta	☐	☐	X	☐	☐	☐
Využitelnost výsledků práce v teorii (v praxi)	☐	☐	☐	X	☐	☐

Vyjádření k výsledku anti-plagiátorské kontroly

Antiplagiátorská kontrola vykazuje 0% podobnosti s jinými pracemi v systému Odevzdej.cz.

Dílčí připomínky a náměty:

- Str. 9, str. 16: apod (apod.)
- Str. 10: *„…nejspíš jsou to rodiči zmíněných dětí…“*
- Str. 14: *„…pokud slovo může být více slovních druhů…“*, *„tři rody (střední, ženský, mužský) jsou rozšířen…“*
- Str. 17: *„Jedním z nejznámějším LLM je Chat GPT, který…“*
- Str. 17: *„Některé se LLM přímo zaměřují na úlohy spojené s přípravou jazyka…“*
- Str. 21: *„…většina testování probíhalo v Pythonu.“*

- Doporučuji volit jiný způsob uvádění zdrojů, tj. nikoliv ve formě horního indexu. Číslo zdroje je pak velmi malé.
- Doporučuji volit jiné formátování odstavců, resp. odsadit první řádek odstavce nebo uvést prázdný řádek mezi odstavci. Vizuálně to bude lépe působit.
- V textu je často používáno dlouhých vět, přitom mohou být lehce rozděleny na sadu jednoduchých vět, viz např. str. 11: „*Obě techniky slouží k normalizaci textu, obě mají své výhody a nevýhody, liší se hlavně způsob, jakým fungují a jejich výsledkem*[8].“, „*Odstranění číslovek je zvláštní případ, dá se říct, že význam věty „Karel a Eva vychovali tři děti.“ je sdělit, že dva lidé mají děti a mimo kontext nezáleží, jestli jich mají deset nebo jedno; v širším kontextu záleží o jaký počet se jedná a z věty se může získat více informací, například že pár má zkušenost s výchovou dětí, nejspíš jsou to rodiči zmíněných dětí apod.*“, str. 14: „*Taggery POS se učí na základě dat, existuje několik druhů: pravidlový, statistický (stochastický) a hybridní – každý přístup má své výhody a nevýhody* [11].“
- Místy jsou věty hůře pochopitelné, viz např. 13: „V češtině může být odhalit význam slova, slovního spojení nebo i věty na základě kontextu výzva i pro člověka natož pro stroj.“
- Slovo chatbot se zpravidla používá dohromady kombinací slov chat a bot, spíše než oddělením těchto slov (chat bot).
- Pokud pro citování zdrojů v samotném textu používáte styl [cislo], pak v seznamu literatury by měla být tato forma [cislo] taktéž použita.
- Seznam použité literatury: zdroj 3 (informace ke zdroji nejsou úplné)

Celkové posouzení práce a zdůvodnění výsledné známky:

Bakalantka si zvolila zajímavé téma, u kterého lze jistě produkovat užitečné výstupy, které by napomohly zájemci o zpracovávání textů s výběrem metody i vhodného nástroje. Ač je z textu patrná jistá snaha o předložení detailní analýzy nástrojů (před)zpracovávající nestrukturované texty s využitím vybraných metod, nelze konstatovat, že bylo tohoto cíle plně dosaženo. Níže jsou uvedeny argumenty, které podporují toto tvrzení:

1. *Strukturování textu*: Jistým neduhem práce je způsob strukturování vět. Často jsou vytvořeny dlouhé věty, které by bylo velmi jednoduché zkrátit, resp. čárkami je často pospojováno několik vět hlavních, přitom žádná není vedlejší, viz připomínky výše.
2. *Faktické nedostatky*: Pozor na odlišení LLM a aplikace, která ho používá, resp. ChatGPT není LLM, viz str. 17. V práci není explicitně uvedena metodika výběru nástrojů, resp. dle jakých kritérií byla zvolena daná paleta nástrojů. V podkapitole 4.2 (Používané nástroje) je zmíněno: „*Veřejně dostupné nástroje jsou převážně k dispozici jako open source (otevřené) knihovny, většina testování probíhalo v Pythonu. Mezi tyto nástroje patří…*“ Z této věty se nedá usuzovat na to, které nástroje budou porovnávány, testovány. Lze se tedy jen domnívat, že kritériem byla volná dostupnost a open-source charakter.
3. *Citování a práce se zdroji*: Na str. 18 je zmíněn zákon č. 429/2022 Sb. o právu autorském spolu s následujícím textem: *„(1) Do práva autorského nezasahuje ten, kdo …“* Pokud to má být přímá citace, domnívám se, že ano, pak není nijak označena a odlišena od okolního textu. Není uveden zdroj, odkud citace přímo pochází, resp. v textu je zmíněn zákon, který není uveden v seznamu s použitou

literaturou. Zároveň pokud jsou pro analýzu nástrojů použity vybrané knihy (viz Saturnin a Povídání o pejskovi a kočičce), je nutné je citovat a uvést v seznamu s literaturou. Kapitola 4.3 se zmiňuje o používaných nástrojích. Pro většinu z nich nejsou uvedeny odkazy (zdroje, kde je lze získat). U některých zdrojů, uvedených v seznamu literatury, chybí informace, tj. u internetových zdrojů chybí informace o datu citování zdroje. U jiných zdrojů, zmíněných v seznamu literatury, je uvedeno „Dostupné z“, u jiných zdrojů toto chybí.

4. *Analýza nástrojů*: Analýza nástrojů, která začíná od kap. 5., není provedena úplně s přihlédnutím ke všem nástrojům. V podkapitole 5.1 se analyzují stopslova. Jak jsou na tom s identifikací stopslov ostatní nástroje, tj. Stanza, CorPy, Simplemma, Polygot? Mají vlastní seznam stopslov nebo nikoliv? Bylo by vhodné se i něm vyjádřit, nejen k NLTK. Tabulka 7 uvádí výsledky tokenizace napříč zkoumanými texty a nástroji. Jak tokenizace dopadla u nástroje Polygot? Je zde důvod, proč není zmíněn? Pokud tokenizace není schopen, nutné se k tomuto vyjádřit. V podkapitole 5.3 (Lemmatizace) jsou porovnávány nástroje CorPy, Stanza a Simplemma. Není uvedena informace o nástorji Polygot a NLTK, přitom např. NLTK nabízí lemmatizaci s podporou WordNet slovníku. V podkapitole 5.4 (POS tagování) jsou porovnány nástroje CorPy a Stanza. Ostatní nástroje nejsou do analýzy zapojeny, přitom např. nástroj NLTK POS tagování podporuje. Obdobně je tomu v podkapitole 5.5, ve které jsou porovnány nástroje CorPy, Polygot a Stanza. Ostatní nástroje nejsou do analýzy zapojeny. Přitom opět, např. nástroj NLTK NER tagování podporuje.

   Na str. 22 je zmíněna tvorba vlastního frekvenčního algoritmu. Tento algoritmus, nebo odkaz na něj na jiné místo v práci, není zmíněn. Tabulka 5 uvádí výsledky odstranění stopslov z knihy Saturnin. Jakým nástrojem byla provedena lemmatizace? Také pomocí CorPy? Z tabulky není patrné. Obdobně u Tabulky 6.

5. *Ostatní*: Alespoň ve formě příloh by bylo vhodné doplnit práci zdrojovými kódy, resp. jak byly jednotlivé techniky (před)zpracovávající text realizovány jednotlivými nástroji. Pokud uvažujeme samotný text od kapitoly 1 (Úvod) po kapitolu 7 (Závěry a doporučení) práce čítá pouhých 27 stran.

Předkládaná bakalářská práce má dobrou teoretickou úroveň. Jednotlivé metody (před)zpracovávající text jsou dobře vyloženy. Bohužel je z textu patrná určitá nesystematičnost a nedůslednost při zpracovávání jinak velmi zajímavého tématu.

Otázky k obhajobě:

1. Na str. 18, v souvislosti s nástrojem Simplemma, používáte slovní spojení „*povrchové vyhledávání*“. Vysvětlete tento pojem blíže.
2. Na základě jakých kritérií jste vybrala danou paletu nástrojů?
3. Jaký citační styl je využit u zdrojů uvedených v seznamu literatury?
4. Existuje důvod, proč byly z analýzy nástrojů některé vyčleněny?

Práci doporučuji k obhajobě.

Navržená výsledná známka: E

V Hradci Králové**, dne 15. května 2024**

podpis