Posudek oponenta bakalářské práce

Autor/autorka práce: Ondřej Matura

Název práce: **Vícejazyčné vyhledávání v textových dokumentech**

Obsah práce

Cílem práce bylo prozkoumat možnosti vylepšení historického portálu Porta fontium (PF) v oblasti dvojjazyčného vyhledávání v textu a na základě analýzy navrhnout a implementovat softwarový modul, který bude možno využít pro lepší vyhledávání.

Kvalita řešení a dosažených výsledků

Student identifikoval stávající nedostatky vyhledávacího modulu v portálu PF. Logicky navrhl, že stávající modul by bylo možné vylepšit doplněním vhodného před a po-zpracování a dále pak vlastním přetrénováním modelů fastText. Analýzou před a po-zpracování se zabývá obr. 3.2, který by si ale zasloužil odkaz z textu a lepší popis a rozbor.

Pro implementaci použil student model ColBERT-X, což je vícejazyčné rozšíření transformer modelu BERT. Volbu považuji za vhodnou a výsledné řešení je funkční s omezením na velikost souboru indexu.

Student se též věnuje popisu dostupných datových sad a evaluačních metrik. V textu jsou popsány tři datové sady. Dále práce obsahuje souhrnnou tabulku 2.3, kde jedna použitá sada chybí, ale na druhou stranu jsou zde dvě navíc. Bylo by vhodné toto sjednotit. V popisu existujících prací jsou uvedeny výsledky, proto by bylo vhodné uvést popis metrik dříve.

Výsledky práce bude možné využít pro vylepšení stávajícího vyhledávacího systému PF.

Formální úroveň

Průvodní dokument je vytvořen v systému LaTeX. Má logickou strukturu, je na výborné jazykové úrovni, neobsahuje pravopisné chyby, jen několik překlepů a formálních prohřešků (např. neslabičná slova na konci řádků). U odkazů na obrázky a sekce je potřeba uvést, zda se jedná o obrázek nebo sekci, jinak může dojít k neurčitosti.

Práce s literaturou

Student uvedl v práci celkem 10 odborných publikací, které byly většinou v anglickém jazyce. Tento počet považuji za lehce nižší, než bych u bakalářské práce očekával. Nicméně z textu práce plyne, že musel autor prostudovat i další články, které v seznamu uvedeny nejsou (např. systém word2vec nebo FastText).

Splnění zadání

Zadání bylo splněno v plném rozsahu.

Dotazy k práci

1) Jakou byste doporučil optimální konfiguraci před a po-zpracování (viz obr. 3.2)?

2) Uvádíte, že jste použil lemmatizaci a stemming. Nikde v práci jsem ale nenašel, o jaké metody/knihovny šlo. Vysvětlete prosím.

3) V práci uvádíte problém s velikostí indexu u systému ColBERT-X. Jakým způsobem byste tento problém řešil?

4) Stávající výsledky jsou vyhodnocenuy na úrovni celých odstavců (dokumentů). Proto není na 1. pohled vidět úspěšnost dotazu. Proveďte prosím vyhodnocení na úrovni vět.

5) Stávající systém PF pracuje s historickým jazykem, ale Vy metody testujete na současnou češtinu / němčinu. Zdůvodněte prosím.

Vzhledem k připomínkám uvedeným výše navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 30.5.2023 doc. Ing. Pavel Král, Ph.D.