Hodnocení vedoucího bakalářské práce

Autor/autorka práce: **Ondřej Matura**

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

Cílem práce bylo seznámit se se stávajícím systémem pro mezijazyčné vyhledávání (v češtině a v němčině) aktuálně použitým v portálu Porta fontium. Dále pak na základě rešerše dostupných metod pro mezijazyčné vyhledávání navrhnout vylepšení systému a implementovat softwarový modul, který by mohl být do Porta fontium integrován.

Protože není možné provést porovnání přímo na datech z Porta fontium, bylo nutné nejprve zvolit vhodný dataset, který by porovnání na česko-německých datech umožnil. Pro tento účel byl vybrán dataset CLIR matrix.

V Porta fontium je aktuálně použita metoda vyhledávání pomocí dvojjazyčného rozšíření dotazu – pro hledané slovo je pomocí slovních vektorů (Fasttext) nalezena množina synonym v obou jazycích a dle ní se vyhledává. Tato metoda byla nejprve otestována na datasetu CLIR matrix a byla určena její úspěšnost.

Jako jedna cesta k vylepšení stávajícího sytému byla provedena analýza možností předzpracování a následného zpracování hledaných slov. V experimentu bylo ukázáno, že použití lemmatizace a vhodnějšího nastavení umožní významné vylepšení metriky nDCG.

Dle zadání pak byly prozkoumány modernější metody založené na kontextových vektorech. Na základě dostupných publikací pak byla vybrána metoda ColbertX, která dosáhla zajímavých výsledků a hodnoty 0,36 pro metriku nDCG. Metoda ColbertX pak byla použita i pro návrh a implementaci softwarového modulu.

Použitelnost navržené metody může být omezena velkou velikostí indexu, který aplikace vytváří. Toto omezení by bylo nutné při praktickém nasazení vyřešit.

Text bakalářské práce je přehledně členěný. Použité zdroje jsou odpovídajícím způsobem citovány.

Přístup k řešení byl aktivní a postup byl pravidelně konzultován s vedoucím.

Zadání bylo splněno bez výhrad

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

V Plzni 30. 5. 2023 Ing. Ladislav Lenc, Ph.D.