Univerzita Hradec Králové
Fakulta informatiky a managementu

POSUDEK OPONENTKY DIPLOMOVÉ PRÁCE

Jméno studenta: Bc. Jiří Kánský

Název práce: Metody modelování témat

Autor posudku: doc. RNDr. Kamila Štekerová, Ph.D., MSc.

Cíl práce: Cílem práce je implementovat různé metody modelování témat v rámci extrakce informací z textu. Práce bude obsahovat praktická použití několika různých algoritmů pro modelování témat, které závěrem porovná a vyhodnotí, zda jsou algoritmy vhodné pro vybraná data.

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

Celkové posouzení práce a zdůvodnění výsledné známky:

Téma práce zapadá do oboru Datová věda. Diplomant si zvolil vcelku jednoduchý, mechanický cíl, a sice implementovat různé algoritmy extrakce témat za použití knihoven jazyka Python a porovnat jejich vhodnost pro jeden volně dostupný dataset. Je škoda, že se diplomant nezaměřil na řešení konkrétního problému z praxe.

Text práce je členěn do 7 číslovaných kapitol včetně úvodu a závěru.

Teoretické kapitoly 2, 3 a 4 jsou věnovány výkladu problematiky strojového zpracování přirozeného jazyka, extrakci informací z dokumentů a algoritmům modelování témat. Při překladu z angličtiny se do textu práce dostaly četné nepřesnosti. Kapitoly 2.1 - 3.3, 4.1 jsou koncipovány jako slovníček pojmů, nejedná se o souvislý odborný text. Formulace typu *„počítač sám o sobě nerozumí psanému textu, Tudíž největším problém je, jak docílit, aby počítač textu porozuměl...“* (s.41) nepatří do diplomové práce. Některé citované zdroje jsou nevhodné (Youtube video, popularizační webové stránky). Odkazy

na zdroje autor mechanicky vkládá na konce odstavců. Seznam použité literatury není upraven podle žádného z doporučených standardů, u mnohých položek chybí jména autorů a další údaje.

Praktická zkušenost s modelováním témat je popsána v kapitolách 5 a 6, které jsou věnovány implementaci pěti algoritmů a představení výsledků extrakce témat z volně dostupného datasetu *20 new group*. Tyto implementace jsou běžně dostupné, např. na webové stránce https://www.kaggle.com/ lze nalézt implementaci algoritmů LSA a LDA na témže datasetu. Shrnutí výhod a nevýhod jednotlivých metod (s. 71) je vágní a málo objevné, když za výhody metod (sloužících ke zpracování velkých objemů dat), jsou označeny mj. „*lehká implementace*“ či skutečnost, že metoda „*automaticky najde množství témat*“, nevýhodou pak má být „*potřeba velkého množství dat*“.

Cíl práce byl rámcově splněn, předložená práce může sloužit jako jednoduchý, nepříliš pečlivě zpracovaný úvod do problematiky extrakce témat. Vlastní přínos diplomanta hodnotím jako malý.

Otázky k obhajobě:

1. Výsledky (s. 66-70) jsou shrnuty do tabulek, v nichž extrahovaná témata jsou uvedena česky a klíčová slova anglicky. Proč?

2. Je možné k modelování témat použít velké jazykové modely, např. ChatGPT?

Práci doporučuji k obhajobě.

Navržená výsledná známka: D

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

podpis