UNIVERZITA PARDUBICE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

POSUDEK OPONENTA DIPOLOMOVÉ PRÁCE

Jméno studenta: **Klečanský Pavel**

Název práce: **Deduplikace dat a jejich využití**

Autor posudku: **Ing. Martin Pozdílek Ph.D.**

Zadání odborného problému a použití metod řešení v rámci diplomové práci

Cílem práce je podrobný popis algoritmů a problematiky deduplikace a spojování záznamů. Praktickým výstupem práce bude vytvoření knihovny, která umožní provádět spojování záznamů mezi dvěma zdroji dat a deduplikaci jednoho zdroje dat. V knihovně budou použity vybrané algoritmy popsané v teoretické části. V teoretické části bude popsán celý workflow deduplikace dat a spojování záznamů, a to od čištění dat až po klasifikaci. Práce také popíše a ukáže možnosti využití například Jaro-Winklerovy vzdálenosti, Levenshteinovy vzdálenosti a Damerau-Levenshteinovy vzdálenosti. Součástí práce bude i Jaccardův index a Q-gram podobnost.

Konkrétní výsledky diplomové práce

Teoretická část práce správně popisuje problematiku spojování a deduplikace záznamů mezi datovými zdroji. V práci jsou podrobně popsány všechny potřebné fáze od předzpracování dat přes různé metody blokování a porovnání záznamů až po slučování nalezených duplicitních záznamů. Pro každou fázi jsou popsány různé metody jejich princip, výhody a nevýhody. Teoretická část práce je zpracovaná kvalitně.

V praktické části autor vytvořil v jazyce Java svoji knihovnu, která implementuje jednotlivé fáze spojování dat. V knihovně jsou přítomny různé metody a algoritmy popisované v teoretické části práce.

Dílčí připomínky a náměty

Práce by mohla obsahovat uživatelský manuál pro jednodušší používání ostatními uživateli, ukázky kódu, popsání parametrů apod. Například u krátkého popisu jednotlivých hodnot parametrů by mohl být uveden dopad na výkon a přesnost.

Celkové posouzení práce a zdůvodnění výsledné známky:

Cílem práce bylo vytvořit knihovnu pro úlohu deduplikace a spojování záznamů. Toto student splnil a cíle práce byly naplněny. Z práce je patrné, že student rozumí aspektům práce, a to jak po teoretické stránce, tak i po praktické stránce. Fungující finální knihovna ukazuje, že navržené postupy řešení problémů byly správné.

Práce obsahuje minimum chyb a překlepů. Po formální stránce práce nemá žádné závažné nedostatky.

Vyhodnocení kontroly textu práce pomocí systému pro odhalování plagiátu:

Diplomová práce prošla kontrolou plagiátorství a výsledek je, že práce není plagiátem.

UNIVERZITA PARDUBICE
FAKULTA ELEKTROTECHNIKY A INFORMATIKY

Otázky k obhajobě

V práci nebyly uvedeny jednotlivé rychlosti klasifikačních algoritmů. Mohl byste prosím upřesnit, jaké jsou rychlosti provádění jednotlivých klasifikačních algoritmů, které jste implementoval?

Jaké jsou výhody knihovny proti ostatním konkurenčním knihovnám?

Jakými funkcionalitami byste případně rozšířil knihovnu v budoucnu?

Práci doporučuji k obhajobě.

Navržená výsledná známka: A

V Pardubicích**, dne 24. května 2024**

_____

podpis