# Posudek vedoucího diplomové práce

DOBIÁŠOVSKÝ, Jan. *Přibližná shoda znakových řetězců a její aplikace na ztotožňování metadat vědeckých publikací = Approximate equality of character strings and its application to record linkage in metadata of scientific publications.* Praha: Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví, 2020. Vedoucí diplomové práce Dr. Mgr. Jan Dvořák.

Práce popisuje metody určování vzdálenosti znakových řetězců a jejich použití pro ztotožňování metadatových záznamů vědeckých publikací. Práce přináší data a poznatky, která všem provozovatelům informačních systémů o aktuálním výzkumu umožní optimalizovat rozpoznávání shody záznamů o publikacích v databázích Web of Science a Scopus.

## K obsahu práce

Práce začíná kapitolou, která podává obecný úvod do problematiky ztotožňování záznamů. Je představen klasický model dle Fellegiho a Suntera a proces ztotožňování záznamů s několika dílčími modifikacemi. Jsou diskutovány některé překážky, které se v reálné praxi vyskytují. Tato kapitola je dobře zpracovaná a obsahuje všechny důležité informace pro zbytek práce.

Dále následuje představení celkem pěti definic vzdálenosti znakových řetězců a jim odpovídajících metod výpočtu této vzdálenosti. Tato kapitola je také zpracovaná dobře. Velice názorně je vyložen algoritmus výpočtu Levenštejnovy vzdálenosti; další metody jsou zavedeny, avšak jejich vlastnosti ani postup jejich výpočtu již nejsou ilustrovány tak podrobně.

Následující kapitola diskutuje speciálně metadata vědeckých publikací. Probírá jednotlivé položky metadatového záznamu a diskutuje jejich vlastnosti. Zvláštní pozornost věnuje identifikátorům, které umožňují jednoznačně rozlišit dokumenty či zdroje, v případě DOI pak i jednotlivé publikace, a ztotožnit záznamy, u kterých se vyskytují rozdíly v textových položkách. V oddíle 4.6 jsou posléze nastíněna kritéria pro výběr údajů pro ztotožňování metadatových záznamů.

Poté je v krátkosti představen informační systém V3S, který ČVUT v Praze používá pro evidenci vědecké činnosti. Autor prezentuje informace z jeho publikovaných popisů a z nápovědy systému. Tato kapitola obsahuje všechny informace potřebné k tomu, aby si čtenář práce učinil představu o systému a o praktické úloze, která pro něj byla řešena.

Praktická část práce popisuje úlohu ztotožňování záznamů pocházejících z citačních databází Web of Science a Scopus, kterou je třeba řešit v systému V3S. Je představena zvolená metoda a datové množiny, na které je aplikována. Postup zpracování dat je popsán dostatečně podrobně. Autor se soustředí na názvy publikací jakožto primární údaj, podle kterého posuzuje shodu metadatových záznamů. Formou grafů a v některých případech i tabulek jsou prezentovány výsledky značně rozsáhlých výpočtů pro celkem sedm variant výše zmíněných pěti metod určování vzdálenosti znakových řetězců (Levenštejnova vzdálenost, Jarova podobnost, Jaro-Winklerova podobnost, kosinová vzdálenost 3- a 4-gramů, Jaccardův koeficient pro 3- a 4-gramy), které autor používá na názvy publikací. Na

publikacích s DOI je provedena kalibrace modelů pro optimální míru F, $F_2$ nebo $F_3$. Následně autor popisuje ještě ruční kontrolu vzorku záznamů bez DOI, kde jsou identifikovány typy chyb, které se vyskytly, a jsou i kvantifikovány.

Závěr práce vhodně shrnuje nejdůležitější poznatky.

## Ke zpracování práce

Práce je psána odborným stylem. Autor volí vcelku výstižné formulace. Občas se vyskytují drobné gramatické chyby. Při uvádění desetinných čísel autor používá desetinnou tečku, je v tom však konzistentní, takže srozumitelnost práce tím netrpí.

Práce je doprovozena významným dodatečným materiálem: datovou sadou a programovými skripty použitými pro zpracování dat. Oba tyto artefakty jsou umístěny na renomovaných úložištích (Zenodo v případě datové sady, GitHub v případě programů), které jsou současnými faktickými standardy pro nastupující trend Open Science. Obojí je v dostatečné míře zdokumentováno pomocí souborů README. Datová sada dostala přidělený perzistentní identifikátor DOI. Je jasná vazba mezi těmito artefakty a diplomovou prací, a to oběma směry.

## Ke spolupráci s autorem práce

Diplomant přistupoval ke své práci velmi zodpovědně. Sjednaných konzultací využíval k řešení v dané době nejpalčivějších otázek a byl na ně vždy dobře připraven. Naprostou většinu sdělených připomínek zapracoval. Diplomant překonal i obtíže se zpracováním dat relativně velkého rozsahu (teoretických kombinací pro vyhodnocení podobností řetězců je více než 800 milionů). Celkově hodnotím spolupráci s ním jako výbornou.

## Ke kontrole práce proti plagiátorství

### Kontrola systémem Turnitin

Systém Turnitin nalezl jen útržkovité shody s texty ve svých zdrojích, z nichž žádná není relevantní jako pokus o plagiátorství.

### Kontrola systémem Theses.cz

Systém Theses.cz nalezl jedinou shodu, která byla v seznamu literatury a byla zcela irelevantní. Systém tedy nenalezl žádné pokusy o plagiátorství.

## Celkové zhodnocení

| Aspekt kvalifikační práce | Vysvětlení | Možné hodnocení | Hodnocení oponentem práce |
|---|---|---|---|
| metodologie a věcné zpracování tématu | Práce je dobře zpracovaná, výklad postupuje logicky. | 0-40 bodů | 36 bodů |
| přínos a novost práce | Práce řeší konkrétní zadání, které se velmi často vyskytuje v institucionálních systémech o aktuálním výzkumu (CRIS). | 0-20 bodů | 18 bodů |
| citování, korektnost citování, využití inf. zdrojů | Citování je korektní, práce využívá relevantní zdroje. | 0-20 bodů | 20 bodů |
| slohové zpracování | Práce je psána odborným stylem. Některé formulace jsou složitější, nežli by bylo nezbytně nutné. | 0-15 bodů | 9 bodů |
| gramatika textu | Občasné drobné gramatické chyby. | 0-5 bodů | 3 body |
| CELKEM | | 0-100 bodů | 86 bodů |

Práci považuji za celkově zdařilou. Navrhuji známku výborně (1).

Jan Dvořák
ÚISK FF UK