Hodnocení kvalifikační práce na ÚISK FF UK

Název práce: **Porovnání vybraných metod komprese textových dat**

Typ práce: bakalářská

Hodnocení práce: **velmi dobře**

Vedoucí práce: prof. RNDr. Jiří Ivánek, CSc.

Oponent/-ka práce: Dr. Jan Dvořák

Konzultant/-ka práce: —

Řešitel/-ka práce: Ondřej Malý

1. Hodnoticí kritéria práce a bodové ohodnocení

Obsahové posouzení práce (max. 70 bodů)

Aspekt práce	Vysvětlení	Body
Kvalita shrnutí současného stavu poznání, používání odborné terminologie, zohlednění různých pohledů na téma práce	Práce je v tomto ohledu vyhovující. Autor v teoretické části věnuje velkou pozornost terminologické čistotě. V praktické části přebírá terminologii poskytovatelů posuzovaných softwarových produktů, což se bohužel propisuje do jisté terminologické nejednoznačnosti závěru práce. Nejmodernější metody entropické komprese jsou zmíněny, avšak nikoli popsány.	8/10
Definice cílů, výzkumná otázka a její zasazení do současného stavu poznání	Stanovené výzkumné otázky jsou v pořádku. Autor ignoruje výpočetní náročnost komprese či dekomprese dat, což vede k velmi jednostrannému porovnání pouze na základě kompresního poměru. Analýza uživatelského rozhraní kompresních programů (jak grafického, tak volání z příkazové řádky) je přínosná. Autor také nijak nezmiňuje validní požadavek interoperability. Formát *zipx* (na kterém byl naměřen nejlepší kompresní poměr) je proprietární: nelze s ním pracovat jinak než opět v programu WinZip. Pro celou řadu použití tím je tento formát diskvalifikován.	6/10
Metodologické zpracování výzkumné otázky, vysvětlení volby využité metody s ohledem na výzkumnou otázku	Nastíněná metodologie je při již zmíněné v zásadě v pořádku, velmi však kulhá její provedení. Experimenty na jednom jediném souboru nemohou poskytnout žádný obraz chování různých algoritmů a softwarových produktů, poskytnou nanejvýš pár datových bodů v nesmírně komplexním prostoru. Volba použitého souboru není nijak odůvodněna.	8/20

Analýza, interpretace výsledků a formulace samostatných závěrů vč. zohlednění současného stavu poznání	Celkově v zásadě v pořádku, až na pár případů: V poddílu 4.4.1 („7-Zip“): Archivní soubory s příponou *tar* a *wim* žádnou kompresi nepoužívají, jak lze zjistit např. ze serveru https://file.org či z katalogu souborových formátů na webu americké Library of Congress. Striktně vzato tedy tyto formáty nespadají do tématu práce. Také je poněkud přehnané tvrzení z pododdílu 4.5.1 („Uživatelské rozhraní a příkazová řádka“), že vyvolání programu z příkazové řádky vyžaduje elementární znalosti programování. Také tvrzení ze závěru, že uživatel nemůže při používání komerčních softwarových produktů vybrat konkrétní komprimační algoritmus, sice platí v grafickém rozhraní, celkově však je příliš paušální. Při volání z příkazové řádky to možné je, jak autor sám ukazuje praktickým experimentem ve své práci. To, že některé programy používají kódy bez mnemotechnické souvislosti s označovaným algoritmem, je malá nepříjemnost vyžadující dohledání v dokumentaci, nikoli nepřekonatelná překážka.	10/20
Invence, inovativní a původní přístup	—	5/10
Počet bodů (obsah)		37/70

Formální posouzení práce (max. 30 bodů)

Aspekt práce	Vysvětlení	Body
Formální struktura práce a návaznost kapitol	Práce je dobře strukturovaná, oddíly na sebe navazují dobře.	10/10
Aktuálnost zdrojů, korektní citování, cizojazyčné zdroje	Jsou použity relevantní, přiměřeně aktuální zdroje, a to včetně zdrojů v anglickém jazyce. Citování je korektní.	10/10
Slohové zpracování a gramatická správnost	Slohově je práce velice zdařilá, čtivá. Gramaticky je až na jedinou výjimku (poslední položka ve slovníku zkratek na str. 41) v pořádku.	9/10
Počet bodů (forma)		29/30
CELKEM BODŮ		66/100

2. Komentář k hodnocení práce

Práce působí dojmem malých ambicí: porovnání algoritmů pouze podle jediného kritéria (kompresního poměru) umožňuje jen velmi jednostranné zhodnocení. Práce alespoň v posledním odstavci svého závěru uvádí, jak by komplexní analýza měla vypadat, aby byla užitečná. Teoretická část práce je však zdařilá, výklad je názorný.

Kontrola plagiátů systémem Turnitin nalezl shody pouze v seznamu literatury, kde jsou naprosto oprávněné. Systém Theses.cz dokonce nenalezl shodu vůbec žádnou.

3. Otázky či tematické okruhy k obhajobě

Archivní formáty jako *tar* a *wim* sice samy nepodporují kompresi vkládaných souborů, avšak celý výsledný archiv může být komprimován následně. Takto mohou vznikat např. soubory se složenou příponou *tar.gz*. Posuďte výhody a nevýhody takovéhoto řešení ve srovnání s formátem *zip* na konkrétním příkladu datové sady *ORCID Public Data File 2023* (DOI https://doi.org/10.23640/07243.24204912.v1). Je mezi těmito přístupy rozdíl, pokud chcete extrahovat všechny soubory nebo pouze jeden určitý? (Pro zodpovězení této otázky není nutné stahovat žádný soubor dané datové sady.)

4. Závěrečné zhodnocení

Práci doporučuji k obhajobě se známkou **velmi dobře**.

Bodový zisk za práci	Hodnocení
83–100 bodů	Výborně (1)
82–66 bodů	Velmi dobře (2)
65–50 bodů	Dobře (3)
≤49 bodů	Neprospěl/a, nedoporučeno k obhajobě

V …… dne ……. ……

jméno a příjmení zhotovitel/-ka posudku