

## POSUDEK DIPLOMOVÉ PRÁCE

**Typ posudku:** *oponent práce*

**Autor/ka práce:** Václav Čepelák

**Název práce:** **Analýza biografických vyprávění pamětníků s užitím počítačové textové analýzy**

**Vedoucí práce:** Martin Hájek

**Oponent/tka:** Petr Soukup

**Navržené hodnocení:** *výborně*

**Zdůvodnění Vašeho hodnocení práce (zejména výhrad a kritických připomínek) rozepište podle níže uvedených kritérií.**

1. Je cíl práce (výzkumná otázka) jasně formulován a odpovídají mu závěry? Je práce vhodně strukturována?

*Cíl práce je stanoven jasně. Je možné se přít, zda není příliš obtížný, ale po mém soudu jej autor naplnil. Struktura práce je zcela přehledná a čtení usnadňují i velice dobrá dílčí shrnutí.*

2. Opírá se autor/ka o dostatek literatury relevantní tématu práce? Využívá i cizojazyčné texty?

*Autor využívá nadstandardní rozsah relevantní literatury z mnoha oborů (kvalitativní a kvantitativní výzkum, statistika, lingvistika, historie. Z textu práce je zřejmé, že nejde o umělý seznam, ale s každým citovaným zdrojem autor poctivě pracuje.*

3. Jaká je kvalita použitých dat nebo zdrojů (včetně jejich výběru, sběru a popisu) a metod jejich analýzy?

*Za data považuji použité rozhovory funkcionářů, disidentů apod. Jde o velice kvalitní zdroj pro analýzu a je jisté dobře, že se jej autor pokusil vytěžit. Nadto autor de facto v práci posuzuje kvalitu datového zdroje a konstatuje, že jde o kvalitní data.*

4. Jaká je kvalita argumentů, o něž autor/ka opírá závěry, k nimž dospěl/a?

*Vynikající jsou první tři části práce, kde autor poukazuje na různé přístupy textové analýzy a možnosti zpracování textů. V části věnované vícerozměrnému škálování jsou sice dílčí chyby (viz dále), ale ty nijak nesnižují celkovou kvalitu textu.*

*V části čtvrté mám pochybnosti o interpretaci grafů, hned první bych četl zcela jinak, protože některé body (slova) v naznačených elipsách k sobě mají velice daleko a pochybuji o jejich častém spoluvýskytu. Obdobně při analýze a interpretaci bych čekal, že bude posouzeno řešení ve více než dvou rozměrech a srovnáno s dvourozměrným (hodnoty charakteristika Stress jsou pro klasické statistiky neúměrně vysoké).*

5. Jsou v práci autorova/autorčina tvrzení a zjištění jasně odlišena od tvrzení a zjištění převzatých? Ano, zcela bez problémů.

6. Jaká je úroveň odkazového aparátu, jazyka a dalších formálních náležitostí?

*Diplomant působí jako zkušený autor odborných textů, není co vytknout.*

7. Jiné přednosti a/nebo nedostatky, které neodpovídají výše uvedeným kritériím (jsou-li jaké).

Předností je jasná struktura i úvahy přesahující téma práce.

Nedostatky jsem (s ohledem na své oborové zaměření) shledal v partii věnované vícerozměrnému škálování. Cituji dále s uvedením stran v elektronickém pdf dokumentu.

*K vizualizaci je pak využito mnohorozměrné škálování, tedy statistická metoda, která umožňuje vícerozměrný prostor zredukovat do dvou dimenzí. Str. 31*

Věta není pravdivá, redukuje se na méně dimenzí než má původní úloha (počet různých slov), ale automatická redukce na dvě dimenze není správná ani vhodná.

*Metodu lze klasifikovat do několika typů. Lze rozlišit škálování jednoduché (na základě jedné datové matice) a replikované (více datových matic). V rámci eukleidovského modelu lze stanovit váhu jednotlivých dimenzí (vážený a nevážený eukleidovský model). Pracovat lze se symetrickou maticí (vzdálenost z bodu A do bodu B je též jako vzdálenost z bodu B do bodu A) či asymetrickou (obě vzdálenosti se liší). Data mohou do analýzy vstupovat jako matice vzdáleností či jako matice proměnných. Str. 41*

Vážený model má jinou logiku než že by autor analýzy stanovoval váhu dimenzí. V analýze jde o vzdálenosti případů či proměnných a dle toho adekvátní matice, vyjádření není přesné.

*generovaných dat pro různé počty proměnných uvádí následující vzorec pro výpočet stresu pro m-dimenzionální zobrazení matice n prvků. Hodnota stresu pro konkrétní zobrazení by měla být významně nižší než hodnota náhodně generovaného stresu. Na základě tohoto vzorce však lze jen těžko usuzovat na statistickou významnost rozdílu naměřené hodnoty stresu od této simulované hodnoty. Str. 43*

Vzhledem k ryze explorační podobě techniky pochybuji o možnosti testování statistické významnosti.

*Tabulka 1: Modelová kontingenční tabulka pro výpočet koeficientů vzdálenosti*

*Tabulka 2: Vzorce výpočtu jednotlivých koeficientů vzdálenosti*

Vždy jde o míry podobnosti ne koeficienty vzdálenosti, vyšší hodnota znamená vyšší shodu a ne větší vzdálenost.

Drobnosti: čtvercový korelační koeficient, iterativní majorizace

Jde o druhou mocninu korelačního koeficientu, tzv. koeficient determinace a iterativní nejspíš bude maximalizace.

8. Náměty na diskuzi při obhajobě práce.

Autor by se mohl vyjádřit k možnosti interpretace os výsledných řešení a případném užití řešení s více než dvěma dimenzemi.

Celkové hodnocení práce:

*Práce je velice kvalitní a velmi rád ji doporučuji ji k obhajobě a navrhuji hodnotit jako **výbornou**. V případě kvalitní obhajoby navrhuji zvážit pochvalu děkana. Osobně autora vyzývám k publikaci dílčích částí (může vzniknout článek o provedené analýze nutnými úpravami, téměř bez úprav může vzniknout text o možnostech analýzy textů).*

Datum: 11.6.2012

Podpis: