

Posudek diplomové práce

Diplomantka: Veronika Straková

Předložená diplomová práce se zabývá otázkou, zda lze na základě kvantifikovaných vlastností sociálních sítí postav extrahovaných z beletrie se statistickou významností určovat autorství. K tomuto účelu je využita paleta nástrojů zahrnující zpracování přirozeného jazyka (NLP), teorii grafů, statistiku a programování v jazycích Python a R. Předložená práce je poměrně komplexní a její vypracování zřetelně obnášelo rozsáhlejší proces, než je ten, který je prezentován ve finálním textu. Komplexita celé práce tkví především v několika netriviálních vrstvách problémů, které diplomantka musela postupně řešit, aby se vůbec dostala k cílové otázce celé práce.

Text práce obsahuje (1) praktické základy teorie grafů, včetně rešerše literatury odpovídající na to, jak jsou extrahovány propojení postav z textů knih a jak z nich produkovat grafy či sítě. (2) Společně s rešeršemi vysvětluje úskalí procesu extrakce jmen postav z textu. V tomto (pro práci kritickém) kroku diplomantka navíc vylepšuje detekční a extrakční schopnosti použité knihovny (v průmyslovém standardu) vlastní metodou (s navýšením původní úspěšnosti z 93,4 na 98,6 % F1). (3) Pro získané sítě postav měří kvantitativní vlastnosti, které (4) analyzuje vybranými statistickými metodami za účelem jejich redukce a následně s cílem určování autorství textů v jazyce R. (5) Práce poskytuje zajímavý závěr.

Práce je ve své podstatě velmi rozsáhlá, její text je ale psán stručně s vědomím konkrétního cíle. Jednotlivé popsané vrstvy jsou diplomantkou zvládnuty velmi pěkně, velká péče je především věnována extrakci postav z textu. Diplomantka je dále explicitní ve všech potenciálních problémech, které mohou při užití konkrétních metod nastat a diskutuje jejich případná řešení i další případné nedostatky. Práce je doplněna kvalitními grafy, ukázkami sociálních sítí postav a řadou příkladů, které pomáhají porozumět rozmanitým popisovaným problémům. Práce je z formálního hlediska kvalitní. Obsažena je i skutečně obsáhlá příloha, ve které nalezneme veškerá potřebná data i programové kódy. Využité texty pro analýzu dle jasných kritérií nasbírala diplomantka samostatně z elektronické knihovny Gutenberg.

V práci nalezneme i jisté drobné potíže. Jedná se především o potenciálně nejasné či chybějící formulace, jako např. - str. 49: "... *for each pair of variables* ...", když je porovnávána jedna proměnná délka vůči všem vlastnostem (tj. nikoliv každý pár každé vlastnosti); - str. 52 chybí vysvětlení důsledků matice na obrázku 17; - str. 54, není jisté, zda je EFA aplikována na původní matici obsahující data z nezkrácených textů nebo na matici výsledků po zkrácení (ačkoliv z logiky následujícího textu je patrné, která z verzí to je). V některých fázích textu lze pozorovat, že chce mít autorka již práci za sebou a některé zajímavé kvalitativní důsledky vyplývající z finálních výsledků práce nerozvíjí.

Při hodnocení této práce je dále nutné přihlédnout k tomu, že samotné téma vyžaduje řešení celé řady vrstev, které by často mohly být i jediným tématem diplomové práce. Práce také přináší velmi zajímavé výsledky. Finálně lze vyzdvihnout samostatnost diplomantky jak při práci se zdroji, psaní, tak i v programování detekce, extrakce a měření vlastností sítí, nutno podotknout, že po snad jediném odchozeném semestru programování. Práci lze doporučit k obhajobě se stupněm A.

Otázky:

- 1) Na str. 45 jsou ukázky grafů, co znamenají velikosti a barvy uzlů?
- 2) Proč bylo výhodné z matice bodových grafů a korelací přejít na faktorovou analýzu?
- 3) Jak dlouho zabralo vylepšení NER SpaCy?

Mgr. Vladimír Matlach, Ph.D.
V Olomouci dne 22. 8. 2023.