Posudek bakalářské práce

Název práce: Software na klasifikaci textů
Diplomantka: Veronika Vlasáková

Cílem předložené bakalářské práce bylo vytvořit software (framework v jazyce Python) umožňující trénovat strojově učící se metody na uživatelem definovaných skupinách textů a získané modely následně evaluovat. Vytvořený software je doplněn plnohodnotnou textovou prací, která software popisuje a uvádí řadu ilustrativních aplikací spolu s jejich výsledky a interpretací.

Pomocí představeného softwaru lze pro libovolné skupiny textů nechat počítač zjistit, zda řadou metod (např. SVM, LDA, KNN, Naive Bayes nebo MDS a tSNE) je možné tyto skupiny na základě uživatelem zadaných vlastností odlišit a tedy např. dále zjišťovat to, zda tyto vlastnosti nesou o těchto skupinách informace. Software je napsaný pečlivě, přehledně a umožňuje další rozšíření. Jeho uživatelskou předností je plně definovatelná tzv. *pipeline* na zpracování textů od tokenizace, redukce, lemmatizace až po odstranění stop-slov. Software dále nabízí několik možností, jak texty pro účely strojového učení vektorizovat, a to za pomoci libovolné kombinace modelu *bag-of-words* a kvantitativně lingvistických indexů, včetně volitelných normalizací. Je tak například možné definovat, že vstupní vlastnosti k rozpoznávání autorství mají obsahovat informace o způsobu užití předložek a vybrané stylometrické indexy.

Při náhledu do zdrojového kódu se nicméně objevuje několik málo nedostatků. Nejzávažnější chybou je například individuální aplikace normalizace na trénovací a testovací dataset – správné řešení obnáší *přenesení* popisných statistik nebo extrémů z trénovacího datasetu do testovacího. Tuto chybu je možné jednoduše odstranit, ale může mít dramatické důsledky na evaluaci metod. Dále, nastavení SETTING_METHODS uvádějící seznam metod k trénování by mohl být svou podstatou jen pouhý seznam.

Text tištěné práce následně obsahuje popis, jak software používat a jaké úlohy v něm lze řešit, včetně způsobu zadání a předzpracování textů a následné evaluace výsledků na trénovacích a testovacích datech. Ilustrační aplikace jsou doplněné diskuzí a výsledky jsou doložené tabulkami, shrnujícími grafy a interpretacemi. Text by však mohl obsahovat větší lingvistické vhledy do problematiky metod a příkladů a širší interpretace získaných výsledků. Otázkou relevance jsou pak některé přeplněné grafy metod MDS a tSNE aplikované na rozsáhlé datasety (např. graf 42, str. 62). Text práce je relativně strohý, nicméně srozumitelný a věcný. Je však zřetelné, že samotný text práce byl napsán ve spěchu a neprošel řádnou korekturou. V práci tak můžeme najít překlepy a další problémy spojené např. s větnou strukturou i se sazbou; jinak je text formálně v pořádku.

Jak diplomantka sama uvádí v závěru, aplikaci by přiléhal uživatelsky přívětivý interface. Pozitivním rysem celého softwaru je, že takové rozšíření např. do podoby webové aplikace by neměl být problém.

Práce sice obsahuje některé nedodělky uvnitř softwaru a v psané práci, ty lze ale snadno odstranit. Vzhledem k rozsahu a kvalitě práce lze práci doporučit k obhajobě se stupněm A.

Otázka:
Při reprezentaci textů pomocí *bag-of-words* a následném trénování a evaluaci modelů je třeba, aby byla obsažena pouze slova obsažená v trénovací části datasetu (tj. slova, která jsou dopředu známá). Texty obsažené v testovacím datasetu jsou pak také reprezentovány právě a pouze těmito slovy, tj. slovy z trénovacího datasetu. V jiném případě hrozí zkreslení evaluace obecnosti modelu. Je tato praxe v kódu programu reflektována? Pokud ne, jak by bylo třeba kód upravit?

Mgr. Vladimír Matlach, Ph.D.
V Olomouci dne 25. srpna 2020.