Posudek práce

předložené na Přírodovědecké fakultě JU

☐ posudek vedoucího ☑ posudek oponenta
☑ bakalářské práce ☐ diplomové práce

Autor/ka:	**Patrik Dohnal**
Název práce:	**Detekce kategorie obsahu webové stránky prostřednictvím metod strojového učení**
Studijní program a obor:	Aplikovaná informatika
Rok odevzdání:	2021

Jméno a tituly vedoucího/oponenta:	Ing. Jiří Jelínek, CSc.
Pracoviště:	KI PřF JU
Kontaktní e-mail:	jjelinek@prf.jcu.cz

Odborná úroveň práce:
☐ vynikající ☑ velmi dobrá ☐ průměrná ☐ podprůměrná ☐ nevyhovující

Věcné chyby:
☐ téměř žádné ☑ vzhledem k rozsahu přiměřený počet ☐ méně podstatné četné ☐ závažné

Výsledky:
☐ originální ☑ původní i převzaté ☐ netriviální kompilace ☐ citované z literatury ☐ opsané

Rozsah práce:
☐ veliký ☑ standardní ☐ dostatečný ☐ nedostatečný

Grafická, jazyková a formální úroveň:
☐ vynikající ☑ velmi dobrá ☐ průměrná ☐ podprůměrná ☐ nevyhovující

Tiskové chyby:
☑ téměř žádné ☐ vzhledem k rozsahu a tématu přiměřený počet ☐ četné

Celková úroveň práce:
☐ vynikající ☑ velmi dobrá ☐ průměrná ☐ podprůměrná ☐ nevyhovující

Slovní vyjádření, komentáře a připomínky vedoucího/oponenta:

Cílem práce je využití metod strojového učení pro kategorizaci webových stránek dle obsahu. Autor se ve své práci zabýval klasickými metodami klasifikace textových souborů, přičemž pro experimentální ověřování použil čtyři pevně definované kategorie. V zadání práce je také zmiňována důležitost syntézy textové a vizuální informace, což v práci zohledněno není.

Práce je mimo příloh rozdělena do pěti kapitol. První uvádí čtenáře do problematiky analýzy textu z webových stránek, druhá se pak věnuje obecnému popisu struktury práce. Třetí kapitola se zaměřuje na teoretické základy klasifikace dat. Použité klasifikační metody jsou K nejbližších sousedů, naivní Bayesův klasifikátor, třetí metodou bylo využití modelu SVM. Ve vzorcích uvedených v této kapitole se občas vyskytují matematické nepřesnosti (například konec strany 7). Další části kapitoly jsou věnovány předzpracování textu a převodu textu na číselnou reprezentaci. Určité nepřesnosti lze najít v podkapitole nazvané Word2Vec (nejde o dopředné neuronové sítě). Podkapitola 3.4 se pak věnuje způsobům vyhodnocení úspěšnosti klasifikace.

Čtvrtá kapitola tvoří praktickou část práce. Autor nejprve popisuje možné cesty získání vstupních dat pro klasifikaci a zdůvodňuje, proč vybral cestu ručního sběru. Pro uložení výběru příslušných stránek byl použit formát HTML místo použití databáze. Nesystémově pak působí metoda specifikace (vymezení) kategorie „Ostatní“. Následuje popis tříd, které aplikace využívá, ten je velmi konkrétní, avšak ne zcela dobře čitelný a srozumitelný. Za poněkud neefektivní lze opět považovat práci s daty a ukládání korpusů do CSV souboru corpus_top_words. Nejasnosti jsou i kolem parametru N (K?) na stránce 27 a příkladu uvedenému na obrázku Zdrojový kód 3. Kapitola 4.3 se věnuje testování a výsledkům, autor zde popisuje i experimenty vedoucí k nastavení klíčových parametrů jednotlivých klasifikačních modelů. Uvedené postupy jsou závislé na vstupních datech, nicméně pro konkrétní datové sady jde o stanovení určité metodiky. Zvláštně působí graf 1, kdy po přidání obecné kategorie bylo dosaženo horších klasifikačních výsledků než bez její přítomnosti, což není dostatečně vysvětleno. U vyhodnocení experimentů chybí finální tabulka s jasně definovanými hodnotami dosažené přesnosti a dalších případných měr kvality klasifikace.

V rámci implementace autor využil řadu knihoven jazyka Python určených pro práci s textovými daty a pro klasifikační úlohy, rozsah vlastního kódu autora je tak omezený. V rámci poslední kapitoly v závěru práce si je autor vědom nedostatků jeho řešení, a to zejména z hlediska možných synonym a významově podobných slov. Právě v této oblasti by mohla výrazně pomoci metoda Word2Vec, kterou však autor označuje za nevhodnou. V práci bohužel není zahrnuta problematika využití neuronových sítí pro obdobnou klasifikační úlohu, což by vhodně doplnilo téma práce.

Práce je vybavena přílohami v odpovídajícím rozsahu, které obsahují jak vstupní data, tak příslušný zdrojový kód jejich zpracování i kompletní soubory s výsledky. K dispozici je rovněž jednoduchá instalační a uživatelská dokumentace, přičemž dodání funkční verze v podobě virtuálního stroje či kontejneru by jistě bylo přínosem.

Práci celkově považuji za dobře zpracovanou s drobnými formálními, gramatickými, obsahovými i věcnými nedostatky a navrhuji hodnocení 1 až 2 podle ústní obhajoby. Zadání bakalářské práce bylo naplněno.

Případné otázky při obhajobě a náměty do diskuze:

1. Souhrnně prezentujte dosažené výsledky, zejména z hlediska dosahovaných hodnot přesnosti.
2. Jakým způsobem byste řešil otázku synonym a významově příbuzných slov, bez které je navržená aplikace použitelná ve velmi omezené míře? Proč nebyla pro tuto oblast využita metoda Word2Vec.

3. Zvažoval jste využití cizojazyčné ontologické struktury, jako například Wordnet s jejím případným překladem do českého jazyka?
4. Proč nebyla v projektu využita databáze pro ukládání dat o webových stránkách?

Práci
☑ doporučuji
☐ nedoporučuji
uznat jako diplomovou/bakalářskou.

Navrhuji hodnocení stupněm:
☐ výborně ☑ velmi dobře ☐ dobře ☐ neprospěl/a

Místo, datum a podpis vedoucího/oponenta:

V Českých Budějovicích dne 17. května 2021

Ing. Jiří Jelínek, CSc.

Digitálně podepsal Ing. Jiří Jelínek, CSc.
Datum: 2021.05.17 04:32:23 +02'00'