Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Martin Mareš
Název práce	Pseudonymizace textových datových kolekcí pro strojové učení
Rok odevzdání	2021
Studijní program	Informatika **Studijní obor** Umělá inteligence

Autor posudku	Doc. RNDr. Ondřej Bojar, Ph.D. **Role** vedoucí
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

Diplomová práce Martina Mareše reaguje na významnou potřebu oboru strojového učení, a sice hlad po trénovacích datech. V dnešním elektronizovaném světě existuje u soukromých subjektů i institucí celá řada kolekcí dat, na nichž by bylo možné trénovat a testovat užitečné metody umělé inteligence. Z důvodu ochrany osobních údajů však data typicky nelze předat jinému subjektu (např. výzkumné instituci), natož pak je zveřejnit, aby se vývoje nových metod pro danou úlohu mohla účastnit široká vědecká komunita.

Martin Mareš vytvořil nástroj určený pro původní vlastníky a správce dat, který podstatným způsobem urychlí odstraňování citlivých informací z textů v kolekci, tak, aby kolekce dat mohla být zveřejněna. Nástroj přirozeně nedává žádné záruky, konečná odpovědnost je na vlastníkovi dat a jeho dohodě s lidmi, kteří do kolekce přispěli, ale snadný proces po technické stránce je základním předpokladem, aby vlastníci dat vůbec o zveřejnění začali uvažovat.

Práce je přehledně strukturována. Po úvodu a motivaci následuje velmi cenný rozbor problematiky ochrany osobních údajů i citlivých údajů, které sice z pohledu zákona není nutné speciálně chránit, ale jejichž neuvážené zveřejnění by mohlo způsobit vážné škody. Součástí analýzy je návrh aplikace jako webového nástroje, který bude snadno spustitelný přímo u správce dat, ale podle potřeby umožní i vzdálený přístup anotátorů, kteří budou pseudonymizaci s pomocí stroje provádět.

Citlivé a zejména osobní údaje se v textu nejčastěji vyskytují v podobě osobních jmen nebo v jejich okolí. Po technologické stránce jejich vyhledávání se proto práce soustředí na tzv. rozpoznávače pojmenovaných entit. Martin Mareš empiricky porovnává tři rozpoznávače a pečlivě je vyhodnocuje právě pro účel identifikace osobních a jiných citlivých údajů. Po relativně jednoduchém ale velmi účinném zlepšení postupu nejlepší z rozpoznávačů dosahuje citlivosti 85 %, tj. jen na 15 % citlivých údajů v testovací kolekci Martinův postup automaticky neupozornil.

Zbylé dvě rozsáhlé kapitoly pak přinášejí podrobnou vývojovou a uživatelskou dokumentaci

vyvinuté aplikace. Konfiguraci a instrukce pro spuštění jsou uvedeny v přílohách.

Diplomová práce je po všech stránkách zpracována velmi kvalitně, je dobře strukturovaná, čtivá, dostatečně podrobná a obsahuje jen zcela zanedbatelné chyby. Rád bych vyzdvihl i spolupráci s Martinem Marešem a jeho samostatný a dlouhodobě velmi organizovaný a důsledný přístup k diplomové práci. Na našich pravidelných schůzkách jsme vždy jen vytyčili směr a další rámcové úkoly. O dva nebo tři týdny později Martin představil hotové výsledky a spíše výjimečně další otázky k vyjasnění. Martin se nebál zabrousit i mimo svůj záběr znalostí, zejména pak do problematiky právních aspektů ochrany osobních údajů.

S Martinovou prací jsem celkově velmi spokojen a doporučuji, aby byla přijata.

Práci doporučuji k obhajobě.

Práci navrhuji na zvláštní ocenění.

Vzhledem k aktuálnosti práce a užitečnosti vytvořeného nástroje (práce má šanci podstatně rozšířit repertoár dostupných datových kolekcí, pokud se k jejich zveřejnění po pseudonymizaci jejich vlastníci rozhodnou) doporučuji zvážit zvláštní ocenění a tím práci získat větší publicitu. Konkrétní návrh na typ ceny však nemám.

V Praze dne 23. 8. 2021

Podpis: