Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Adam Osuský
Název práce	Predicting Word Importance Using Pre-Trained Language Models
Rok odevzdání	2024
Studijní program	Informatika
Specializace	Umělá inteligence

Autor posudku	Mgr. Dominik Macháček, PhD.	Oponent
Pracoviště	Ústav formální a aplikované lingvistiky	

K celé práci

	lepší	OK	horší	nevyhovuje
Obtížnost zadání	X			
Splnění zadání	X			
Rozsah práce *... textová i implementační část, zohlednění náročnosti*	X			

Tématem práce je detekce důležitých slov v textu. Autor popisuje motivaci, definuje úlohu a vytváří webovou aplikaci pro ruční detekci důležitých slov v textech anotátory. Dále s pomocí dobrovolníků – anotátorů vytváří evaluační dataset. Následně navrhuje, trénuje a vyhodnocuje automatický model pro detekci důležitých slov.

Obtížnost zadání a rozsah práce hodnotím velmi kladně. Autor vhodně použil znalosti NLP, včetně orientace v odborné literatuře, hluboké učení, matematické definice, statistické metody a softwarové inženýrství. Zadání je splněno bez výhrad.

Textová část práce

	lepší	OK	horší	nevyhovuje
Formální úprava *... jazyková úroveň, typografická úroveň, citace*	X			
Struktura textu *... kontext, cíle, analýza, návrh, vyhodnocení, úroveň detailu*	X			
Analýza	X	X		
Vývojová dokumentace	X			
Uživatelská dokumentace	X	X		

Formální úprava Práce je psaná anglicky, jazyková úroveň je velice dobrá. Autor používá poměrně mnoho matematických pojmů a symbolů v kap. 1 a 4. V kap. 1 na str. 10 chybí vysvětlení symbolů $\mathcal{P}$ a *, což ztěžuje porozumění. Velmi kladně hodnotím, že v celé práci, hlavně v sekci Related work na str. 7, autor správně používá velmi mnoho odkazů na jiné zdroje, především na nedávné odborné články z oblasti NLP. Zanedbatelnou výhradou je, že mnohé reference, např. [36], odkazují na předběžnou verzi článku na Arxiv.org místo finální, recenzované a publikované verze např. v ACL Anthology. K typografické úrovni nemám výhrady, pouze doporučení: 1) v obr. 4.2 a podobných prezentovat pořadí slov více čitelným způsobem než odstíny modré, 2) v tab. 4.6 apod. zvýraznit nejvyšší hodnoty a graficky odlišit baseline, vlastní a převzaté modely a zvážit zachovat pořadí metod ve všech sekcích tabulky, 3) obrázky a tabulky blíže místu, odkud jsou poprvé odkazovány (např. tab. 4.6 odkázána na str. 35, uvedena na str. 39).

Struktura textu Ke struktuře textu nemám výhrady. Úroveň detailu je vhodná, doporučil bych ale následující: 1) uvést požadavky na HW a výpočetní čas pro trénování modelu; 2) v kap. 1 bych doporučil uvést ukázky textů ze všech domén a také více informací o použitém literárním textu (str. 16), např. autor, dílo, období. 3) Chybí zdůvodnění výběru modelu `bert-base-uncased` (str. 23). 4) Doporučuji zveřejnit kolekci anotací (sekce 2.2.2) pro použití v dalších pracích.

Analýza Analytická část práce je dostatečná. Kladně hodnotím zejména přehled Related Work, a sekci 1.2 – definici důležitosti slov a z toho vyplývající instrukce pro anotátory.

Nejdůležitější a nejrozsáhlejší část analytické části práce je kap. 4 – evaluace automatických metod pro řazení slov v textu podle důležitosti. Její malou slabinou je to, že autor navrhuje použití Pearsonova korelačního koeficientu bez ověření jeho předpokladů. Místo něj bych doporučil tzv. rank correlation koeficient, např. Kendalův či Spearmanův. Kromě Pearsona ale autor navrhuje a vyhodnocuje ještě dvě vhodné metody, k-inter a Overlap, tudíž prezentované výsledky jsou správné a relevantní i tak.

Dokumentace Dokumentace obou částí implementací je velmi vhodná. Chybí popis minimálních či doporučených HW požadavků na trénování – velikost GPU RAM. Příprava trénovacích dat podle instrukcí z dokumentace nepoužívá paralelizaci a trvá velmi dlouho. Malým nedostatkem je chybějící upozornění na tuto časovou náročnost a instrukce, jak zapnout paralelizaci, případně návod, jak přípravu dat přeskočit s pomocí autorem předzpracovaných dat.

Implementační část práce	lepší	OK	horší	nevyhovuje
Kvalita návrhu *... architektura, struktury a algoritmy, použité technologie*	X			
Kvalita zpracování *... jmenné konvence, formátování, komentáře, testování*	X			
Stabilita implementace	X			

Implementační část práce hodnotím jako velmi kvalitní a plně funkční. Je složena ze dvou částí, 1) webová aplikace na anotaci dat, a 2) automatický ranker důležitých slov – příprava trénovacích dat, dotrénování jazykového modelu, evaluace.

U webové aplikace (část 1) hodnotím kladně zejména jednoduchost a přehlednost aplikace z pohledu anotátorů a správnost a robustnost z pohledu administrátora. To je zajištěno správným použitím frameworku Django. Silnou stránkou je rovněž leaderboard anotátorů, pro něj bylo potřeba navrhnout netriviální algoritmus.

U části 2) hodnotím kladně zejména správné, vhodné a pokročilé použití knihoven HuggingFace Transformers a Datasets. Kladně hodnotím také to, že všechna data a předtrénovaný model, které program používá, se při prvním spuštění stáhnou z veřejných či autorových repozitářů z Internetu. Kód je přehledný a dobře zdokumentovaný, až na instrukce o paralelizaci. Ty se ale dají vyčíst z kódu a z dokumentace použité knihovny. Kladně hodnotím i to, že na zapnutí paralelizace není potřeba měnit kód, ale jen uživatelskou konfiguraci.

Celkové hodnocení	Výborně
Práci navrhuji na zvláštní ocenění	Ano

Datum: 22. 8. 2024 Podpis