Posudek bakalářské práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Ondřej Michálek
Název práce	Biblické parafrázování
Rok odevzdání	2020
Studijní program	Informatika
Studijní obor	Obecná informatika
Autor posudku	Mgr. Petra Barančíková 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	

Autor naimplementoval desktopovou aplikaci určenou k nahrazování slov na základě kosinové podobnosti word2vecových vektorů. V souladu se zadáním práce, aplikace nahrazuje jednotlivá slova ve větě. Zvolené řešení je velmi jednoduché - nahradit každé substantivum, adjektivum a sloveso slovem s nejbližším word2vecovým vektorem (případně stejného slovního druhu). Bohužel se toto řešení míjí s cílem bakalářské práce - parafrázování biblického textu do modernějšího jazyka.

Nicméně proti první verzi práce došlo k výraznému zlepšení - práce je přehlednější díky ubrání množství nekvalitních modelů a dodání objektivnějšího hodnocení výsledků. Algoritmus samotný považuji za hlavní nedostatek práce. Nejsilnější stránkou je teoretická část věnující se bibli a potom samotný kód, který je díky své stručnosti velmi přehledný.

Pochybnosti z minula, zda autor rozumí word2vecu a tomu jak funguje, u mne stále přetrvávají kvůli nepřesným tvrzením typu "*Word2vec s vektory umí provádět následující úlohy...*" (str.10), nebo tomu, že autor stále ignoruje parametr kontextového okna word2vecu. Tento poměrně zásadní parametr není uveden ani mezi nastavitelnými, ani mezi pevně stanovenými parametry. Jeho hodnotu se nedozvím ani z kódu, který přestože je zkopírován z Githubu, tak zrovna řádek s tímto parametrem autor vymazal a nechal defaultní (tedy 5).

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		
Vývojová dokumentace		X		
Uživatelská dokumentace		X		

V textové části došlo k výraznému zlepšení oproti první verzi bakalářské práce - ubylo gramatických chyb, přibyly citace, zdroje dat a objevila se i stručná rešerše relevantních literatury. Na výtku, že by některá tvrzení (např. "*CBOW lépe funguje pro frekventovaná slova*") měly být podpořeny relevatní citací, sice autor přidal citaci [15], ta ale chybí v seznamu literatury! Seznam použitelné literatury je stále nejednotný ve formátování - některé reference začínají příjmením, jiné jménem, jiné iniciálou.

Mrzí mne, že ačkoli autor již zmiňuje Leeuwenberg et al. (2016) a i to, že word2vec je nespolehlivý nástroj na hledání synonym, nepokusí se s tím nic dělat, např. vyzkoušet Leeuwenbergovu relativní kosinovu podobnost.

Naopak oceňuji, že přibyla alespoň nějaká analýza výsledků a "parafráze" nehodnotí jen samotný autor. Bylo by přehlednější, kdyby byly použito relativní hodnocení (0.75) místo absolutního (15 z 20), zvlášť u několika bodových škál. Také mi tam chybí počet nahrazení u jednotlivých modelů - tipla bych si a autor to také lehce naznačuje, že existuje nepřímá úměra mezi hodnocením modelu a množstvím substitucí.

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		

Algoritmus samotný je velmi jednoduchý, což by ničemu nevadilo, využívá dvou knihoven stažených z GitHubu, hlavní autorův přínos je vytvoření GUI pro jednoduché použití programu.

Bohužel použité knihovny pro word2vec nejsou šťastně vybrané - trénování modelů na větších datech a generování "parafrází" pomocí velkých modelů je příliš pomalé, než aby bylo reálně použitelné. Tento problém by byl jednoduše eliminovatelný použitím kvalitnější knihovny jako je fasttext (který je i v doporučené literatuře), nebo gensim, který má velmi rychlé vyhledávání díky kNN aproximaci.

Proti původní verzi programu došlo k několika malým, ale výrazným zlepšením - jedná se především o větší zapojení MorphoDiTy, která nahradila některé použité heuristiky a opravu překlepů v uživatelském rozhraní.

Značně zarážející je, že příklady uvedené v bakalářské práci neodpovídají výstupům programu - pro model TitulkyABible_100_5_5 (stáhnutý na základě autorova odkazu) mi program generuje zcela jiné věty. Např.:

Vstup: Na počátku bylo slovo, to slovo bylo u Boha, to slovo byl Bůh.

Výstup v BP: Na začátku zbylo sdělení, to sdělení zbylo u Boha, to sdělení byl Bůh.

Výstup programu s Morph.: Na začátku bylo jméno, to jméno bylo u Boha, to jméno byl Bůh.

Výstup programu bez Morph.: Na začátku nebylo jméno, to jméno nebylou bědovali, to jméno byl Bůh.

Celkové hodnocení Dobře

Práci navrhuji na zvláštní ocenění Ne

Datum

Podpis