Hodnocení vedoucího bakalářské práce

Autor práce: Michal Tušl

Název práce: Explicitní sémantická analýza

Obsah práce:

Práce je zaměřena na sémantickou analýzu textů, která tvoří jádro v oblasti automatického zpracování přirozeného jazyka. Práce studuje metody Distribuční sémantiky pro reprezentaci významu slov založené na strojovém učení bez supervize. Od studenta bylo vyžadováno pochopení a využití poměrně náročných algoritmů v teoreticky náročné oblasti strojového učení a zpracování přirozeného jazyka. Experimenty byly provedeny na standardních korpusech v českém a anglickém jazyce. Student měl možnost vyzkoušet si práci s obrovskými textovými korpusy. Pro strojové učení byla použita Wikipedie obsahující téměř 2 miliardy slov pro anglický jazyk a téměř 100 milionu slov pro češtinu.

Kvalita řešení a dosažené výsledky:

Student naimplementoval metody Explicitní sémantická analýza a Latentní sémantická analýza. Obě metody jsou založené na Distribuční hypotéze a využívají Bag-of-Words reprezentaci kontextu. Student následně navrhnul jejich rozšíření a vytvořil metodu, která jako kontexty využívá hierarchii kategorií na Wikipedii. Na výsledek aplikoval singulární rozklad matic, který redukuje paměťové nároky (snižuje množství trénovacích parametrů) a zároveň vylepšuje výsledky metody.
Všechny metody otestoval na standardních korpusech pro měření sémantické podobnosti slov. Implementované metody dosahují na češtině i angličtině dobrých výsledků.

Spolupráce s vedoucím a aktivita studenta:

Student přistupoval k řešení velmi svědomitě a aktivně. Pravidelně konzultoval svojí práci s vedoucím a dodržoval stanovené termíny. V tomto směru byla spolupráce naprosto bezproblémová.

Formální úroveň práce:

Bakalářská práce se skládá z 36 stran vlastního textu (45 stran včetně úvodních stránek a referencí). Práce je vysázena v LaTeXu. Struktura textu je dobře navržena. Autor věnuje dostatečnou část práce úvodu do problematiky a popisu metod sémantické analýzy. Matematický popis je srozumitelný a správný. Z experimentů je patrné, co a jak bylo uděláno. Vyjadřování v českém jazyce je odpovídající. Student používá vektorovou grafiku pro všechny obrázky a grafy.

Úroveň kódu:

Program je napsán v jazyce Java. Odevzdaný kód je plně funkční a dle přiloženého návodu je možné všechny implementované metody jednoduše spustit a natrénovat. Autor dodržuje konvence pro psaní kódu v jazyce Java. Kód je srozumitelný a dostatečně komentovaný. Autor používá Maven jako buildovací nástroj. Jediná velmi nepatrná výtka je, že v odevzdaném programu je špatně nastavená závislost na knihovnu strojového učení Brainy a na algoritmus pro normalizaci slov HPS. Po malé úpravě vše pracuje tak jak má.

Splnění zadání:

Práce zcela splňuje zadání, a proto navrhuji hodnocení známkou **výborně** a práci doporučuji k obhajobě.

V Plzni 28. 4. 2017

Ing. Tomáš Brychcín, Ph.D.