

Hodnocení vedoucího diplomové práce

Autor/autorka práce: Bc. Karel Zíbar

Název práce: Automatická extrakce klíčových slov pomocí metod trénovaných bez učitele

Obsah práce:

Práce se zabývá extrakcí klíčových slov z textů a to za pomoci metod, které nevyžadují supervizi (nevyžadují anotovaná data). Diplomant testoval několik metod (TF-IDF, LDA, GloVe), které poté různými způsoby kombinoval pro dosažení optimálních výsledků.

Kvalita řešení a dosažené výsledky:

Autor na podnět vedoucího práce použil state-of-the-art metody pro danou úlohu. Experimenty byly prováděny na datech ze SemEval 2010 (International Workshop on Semantic Evaluation), konkrétně na úloze #5. Výsledky metod z této práce jsou porovnány s výsledky účastníků na SemEval 2010. Diplomantova nejlepší metoda dosáhla F-skóre 25,95%, což by eventuálně stačilo na druhé místo v soutěži.

Spolupráce s vedoucím a aktivita studenta:

Diplomant zpracovával práci během dvou let. V prvním roce nebyla práce dokončena. V druhém roce práci odevzdal v řádném termínu. Diplomant k řešení přistupoval aktivně až posledních několik měsíců. Po tuto dobu svojí práci pravidelně konzultoval a dbal podnětů vedoucího. Kdyby tímto způsobem pracoval celou dobu, mohla být práce podstatně lepší a dokončená o rok dříve.

Formální úroveň práce:

Diplomová práce se skládá z 56 stran vlastního textu (81 stran včetně úvodních stránek, referencí a příloh). Práce je vysázena v LaTeXu. Struktura textu je rozumně navržena. Autor věnuje dostatečnou část práce úvodu do problematiky a popisu metod pro extrakci klíčových slov. Matematický popis je srozumitelný a až na malé nepřesnosti 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. Zde uvádím několik formálních nedostatků:

- U několika referencí chybí tečka za iniciály autora. U reference č. 5 chybí "š" ve jméně autora.
- Na několika místech autor zaměňuje pojem úspěšnost s pojmem F-skóre.
- Princip kompozicionality nesouvisí s morfologií jazyka. Tato část by měla být spíše součástí sekce Sémantika textu.

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.

Splnění zadání:

Práce splňuje zadání. Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 1. 6. 2017

Ing. Tomáš Brychcín, Ph.D.

**SOUHLASÍ
S ORIGINÁLEM**

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

①