HODNOCENÍ DIPLOMOVÉ PRÁCE

Oponent DP

Jméno diplomanta: Václav Voldřich **Garantující katedra:** KKY

Název diplomové práce: Metody shlukové analýzy pro klasifikaci dokumentů pro účely následného využití v oblasti automatické adaptace jazykových modelů

	Předmět hodnocení	Nadprůměrné	Průměrné	Podprůměrné
1	Jazyková a grafická úprava	☐	☐	☒
2	Formální a obsahová stránka práce	☐	☐	☒
3	Vhodnost použitých metod	☐	☒	☐
4	Způsob zpracování a vyhodnocení	☐	☐	☒
5	Správnost získaných výsledků	☐	☐	☒
6	Vlastní přínos	☐	☐	☒
7		☐	☐	☐

Doplnění hodnocení, připomínky, dotazy:

Diplomová práce vychází z bakalářské práce obhájené diplomantem v roce 2011. Název této bakalářské práce "Metody shlukové analýzy pro klasifikaci dokumentů" naznačuje, že přidaná hodnota diplomové práce by měla spočívat v oblasti automatické adaptace jazykových modelů. Nicméně v celé diplomové práci není pojem jazykový model (natož pak pojem adaptace jazykového modelu) vysvětlen a dále využit. Navíc není vůbec zřejmé, jakým způsobem by navržené metody měly přispět k adaptaci jazykového modelu.

Považuji za nutné shrnout rozdíly diplomovou prací a zmíněnou bakalářskou prací:

- Diplomant popisuje rozšíření příznakového vektoru o n-gramové příznaky a popisuje metodu latentní sémantické analýzy (rozsah 2 strany).
- Dále popisuje metody PCA a ICA (rozsah 1 strana), nicméně tyto metody neimplementuje, používá existující knihovnu funkcí.
- Pro shlukovou analýzu používá různé variace metody K-means (stejně jako v bakalářské práci), navíc uvádí metody Mean-shift a Elbow method (rozsah 2 strany).
- Kapitoly věnované vyhodnocení výsledků.

Diplomant tedy po dvou letech práce rozšířil svoji bakalářskou práci o 5 stran teoretického popisu metod běžně používaných v úloze shlukové analýzy a znovu provedl vyhodnocení výsledků.

Klíčová kapitola číslo 4 (Výsledky) by měla shrnovat dosažené výsledky. Bohužel popis vstupních dat je neúplný, tabulka 4.1 postrádá vysvětlení, co které sloupce znamenají. Ve vyhodnocení je použita F-míra pro vyhodnocení přesnosti klasifikace. Tento způsob vyhodnocení však předpokládá, že ke každému obrazu v testovací množině je přiřazena informace o cílové třídě. Metody učení bez učitele, mezi které použité metody shlukové analýzy patří, však informaci o cílové třídě neposkytují, pouze umožňují zařadit daný obraz do jednoho ze shluků. Na straně 20 je pouze naznačeno, jakým způsobem byly jednotlivým shlukům přiřazeny cílové třídy. Díky tomuto zásadnímu nedostatku není možné určit, zda výsledky vyhodnocení pomocí F-míry jsou validní. Druhé kritérium pro vyhodnocení prezentovaných algoritmů – doba běhu algoritmu – je opět velice diskutabilní. Tabulka D.2 sice shrnuje průměrné doby běhu jednotlivých algoritmů, chybí však popis metodiky vyhodnocení. Zároveň není zřejmé, v jakých jednotkách je doba běhu uvedena (minuty, dny, apod.).

Práce jako celek nesplňuje zadání v těchto bodech:

1. – práce žádným způsobem nedává do souvislosti shlukování a klasifikaci dokumentů s automatickou adaptací jazykových modelů pro automatický překlad.

2.a) – práce nestanoví požadavky řešeného problému a nevybírá nejlepší metodu.

3. – práce uvádí obecně platné poznatky, faktické chyby v kapitole 4 (Výsledky) spolu s chybějícími požadavky na řešený problém brání jakékoli smysluplné analýze.

Práce samotná je psána velice nepečlivě. Chybí odkazy na tabulky a rovnice (str. 18 a 22). Odkazy na použitou literaturu jsou naprosto nevyhovující – seznam literatury obsahuje 5 položek, z čehož jedna je odkaz na Wikipedia.com, druhá pak odkazuje na webové stránky s příkladem použití metody ICA. Ze seznamu odborné literatury, který je součástí zadání práce není použita ani jedna položka. Diplomant uvádí popis použitých metody bez citace konkrétních zdrojů. Navíc neuvádí ani odkaz na svoji bakalářskou práci, přestože předkládaná diplomová práce z ní vychází. Diplomant používá množství zavádějících, nejasných nebo nesmyslných pojmů, např. "trojúhelníková podobnost" (str. 11) nebo věta "Tato matice je převážena TF-IDF." (str. 4). Práce obsahuje i řadu překlepů (shluokvaných, podnebným způsobem). Mnoho vět je nesmyslných, pro příklad zmíním druhou větu z kapitoly Závěr:

> Bohužel výzkum byl tak intenzivní, jak mohl být, ovšem ani základní algoritmy, které byly v této práci testované přinesly výsledky.

Z těchto důvodů považuji zadání za nesplněné, práci nedoporučuji k obhajobě a klasifikuji ji stupněm nevyhověl.

Splnění bodů zadání	☐ úplně	☒ částečně	☐ nesplněno	
Doporučení práce k obhajobě			☐ ano	☒ ne
Celkové hodnocení práce	☐ **výborně**	☐ **velmi dobře**	☐ **dobře**	☒ **nevyhověl**
Jméno, příjmení, titul oponenta: Jan Švec, Ing.				
Pracoviště oponenta: KKY				

5.9.2013
Datum

J. Švec
Podpis