Oponentní posudek diplomové práce

Autor práce: Bc. Michal Koktan
Název práce: Automatické rozpoznávání (analýza) sentimentu
Autor posudku: ing. Jan Kodera

Zadání práce odráží potřebu automaticky klasifikovat dokumenty podle postoje autora či mluvčího k danému tématu. Tato potřeba se vyskytuje v celé řadě oblastí, v případě ČTK jde o doplňkovou informaci využitelnou zejména při zpracování cizích textů.

Z pohledu ČTK jako zadavatele je škoda, že autor vyřadil texty dalších médií a pracoval výhradně s agenturním zpravodajstvím ČTK, ve kterém je rozptyl sentimentu z principu výrazně nižší.

Autor na začátku práce stručně a nepříliš výstižně popisuje cíl práce a její praktickou využitelnost, popisuje historii automatického určování sentimentu v textech a vymezuje úlohu jako speciální případ obecnější problematiky automatické klasifikace dokumentů. Na to navazuje popis tří vybraných statistických klasifikačních metod, přičemž jejich výběr zdůvodňuje pouze odkazem na literaturu (přesněji na práci založenou na klasifikaci filmových recenzí).

V následující části se pak autor zabývá paralelním zpracováním klasifikačních úloh a možností nasazení paralelního zpracování zvoleného algoritmu Expectation-Maximization. Volba algoritmu není zdůvodněna a nejsou zde zmíněny ani jiné algoritmy využitelné pro řešení daného problému.

Ve čtvrté kapitole autor popisuje využití korpusu v úlohách automatické klasifikace a zmiňuje problémy při jeho pořízení, resp. vytvoření pro účely diplomové práce. Podrobněji se pak tvorbou korpusu zabývá v kapitole sedmé, věnované programu, který pro tento účel vytvořil. Z popisu v této kapitole je také patrné, jaké texty byly pro korpus využity. Na tomto místě je třeba zdůraznit, že při zadání práce se nepočítalo s tím, že nebude možné získat hotový korpus a bude třeba ho pracně vytvářet. Tato část jde tedy nad rámec zadání a předpokládaného rozsahu práce.

Po stručném popisu dat, s nimiž autor pracoval, a jejich formátu následuje analýza existujících nástrojů, které je možné nasadit pro řešení úlohy s využitím zvolených algoritmů. Zvolený volně dostupný nástroj MinorThird i jeho úpravy provedeném autorem jsou zde popsány spolu s jeho parametry a používanými formáty. Představena je i metoda křížové validace, která je navržena pro statistické vyhodnocení úspěšnosti dosažených výsledků. Další vytvořený program, který slouží pro parametrizaci a převod textů do formátu vhodného pro vstup do MinorThird, je popsán už v části věnované tvorbě korpusu. Autor pro parametrizaci bohužel použil pouze jednu metodu (dokumentovou frekvenci slov původním tvaru, bez filtrování a lemmatizace), jejíž výběr zdůvodnil pouze odkazem na její jednoduchost.

Těžiště práce leží v kapitolách 8 a 9, které se zabývají postupem klasifikace včetně využití paralelismu, způsobem nalezení optimálních vah přiřazovaných jednotlivým metodám a vyhodnocováním získaných výsledků. Pro paralelní klasifikaci autor využil dvě různé metody určení výsledného zařazení do tříd, a to většinové hlasování a lineární kombinaci klasifikátorů. Autor zde podává i přehled výsledků a porovnává výsledky křížové validace po jednotlivých metodách. Výsledky jsou přehledně prezentovány ve formě tabulek a grafů. Pozitivem je rovněž srovnání výsledků všech metod v závislosti na zvolené velikosti trénovací a testovací sady. Autor použil čtyři velikosti sady, ve všech případech je nejvyšší účinnosti dosaženo u sady velikosti 350 dokumentů a u jediné větší sady (750 dokumentů) úspěšnost mírně klesá. Pro vysvětlení tohoto jevu bohužel není vyslovena žádná hypotéza. Na závěr jsou uvedeny i časy potřebné pro klasifikaci pro všechny čtyři velikosti sad a při sériovém a paralelním zpracování. Z výsledků je patrné, že při využití paralelního zpracování dojde jen k velmi mírnému zrychlení.

Jazyková úroveň práce je dobrá, vyskytují se jen ojedinělé překlepy a chyby. Slabší je však bohužel přehlednost práce, formulace vět v některých místech brání srozumitelnosti a pochopení textu. Srozumitelnosti by rovněž prospěla větší péče věnovaná návaznosti témat a jednotlivých kapitol a vysvětlení souvislostí. V textu se poměrně často vyskytují nepřesnosti, např. je v grafech a tabulkách uváděn znak „%“, přestože hodnoty se pohybují pouze v intervalu <0;1> a chybí tedy přepočet na procenta. Tyto nedostatky připisuji zejména nedostatku času při závěrečné redakci práce, způsobenému hlavně zpožděním vzniklým tvorbou vlastního korpusu.

Struktura dat dodaných na přiloženém CD je dobrá, autor přiložil i pomocné tabulky a výpočty ve formátu tabulkového procesoru, které umožňují hlubší zkoumání výsledků.

Dodaný zdrojový kód je dokumentovaný, třídy a veřejné metody jsou komentovány ve formátu javadoc. Uvítal bych doplnění dalších komentářů přímo v kódu, které by napomohly jeho snadnějšímu pochopení a případným budoucím úpravám. Aplikace jsou dobře strukturované, s přehlednými objektovými modely. Doporučil bych ovšem věnovat větší pozornost ošetření chybových stavů, vznikajících například při nesprávném zadání parametrů. Aplikace v takových případech obvykle končí neošetřenou výjimkou. Drobné chyby se vyskytují i jinde, např. při inicializaci uživatelského rozhraní aplikace pro tvorbu korpusu.

Vytvořený software je pro praktické využití orientován příliš na prezentaci a srovnávání dosažených výsledků a bylo by třeba ho upravit a doplnit o metody a vhodná rozhraní pro napojení na systémy, které by ho mohly využívat.

Celkově diplomant prokázal, že je schopen porozumět zadání, předložený problém analyzovat, nastudovat související literaturu a vyhledat nebo vytvořit a nasadit nástroje vhodné pro řešení. Prokázal rovněž dostatečnou znalost programovacích technik i jazyka Java.

Diplomovou práci doporučuji k obhajobě a hodnotím ji stupněm **dobře**.

Otázky:

- V tabulce 19 uvádíte velmi nízkou úspěšnost klasifikace (Acc) v případě paralelního zpracování oproti výsledkům jednotlivých metod samostatných. Můžete tato čísla vysvětlit?
- Jakým způsobem by podle Vás ovlivnilo úspěšnost klasifikace, pokud by subjektem hodnocení byly celé texty a nikoli pouze odstavce?

V Praze, 9. srpna 2012

ing. Jan Kodera
oponent DP