Posudek oponenta závěrečné práce

České vysoké učení technické v Praze **Fakulta informačních technologií**

Student: Ivo Sklenář
Oponent práce: Ing. Karel Klouda, Ph.D.
Název práce: Automatizované zkracování textu
Obor: Teoretická informatika (bakalářský)

Datum vytvoření: 12. 6. 2015

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 5:*
1. Náročnost a další komentář k zadání	*1=mimořádně náročné zadání,* *2=náročnější zadání,* ***3=průměrně náročné zadání,*** *4=lehčí, ale ještě dostatečně náročné zadání,* *5=nedostatečně náročné zadání*

Popis kritéria:
Podrobněji charakterizujte diplomovou (bakalářskou) práci a její případné návaznosti na předchozí nebo běžící projekty. Dále posuďte, čím je zadání této ZP náročné. (U obtížnější ZP lze dále tolerovat některé nedostatky, které by u ZP standardní obtížnosti tolerovány nebyly; a naopak u jednoduché ZP mohou být zjištěné nedostatky hodnoceny přísněji.)

Komentář:
Obtížnost zadání byla dána zejména tím, jak k němu student přistoupil: při velkorysejším přístupu by se dalo hodnotit i jako náročnější. Po přečtení práce však hodnotím zadání jako průměrně náročné.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
2. Splnění zadání	*1=zadání splněno,* ***2=zadání splněno s menšími výhradami,*** *3=zadání splněno s většími výhradami,* *4=zadání nesplněno*

Popis kritéria:
Posuďte, zda předložená ZP splňuje zadání. V komentáři uveďte body zadání, které nebyly zcela splněny, případně rozšíření ZP oproti původnímu zadání. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Komentář:
Zadání neobsahovalo jasně definované úkoly (což nemyslím nutně jako kritiku), takže je těžké soudit, zda bylo splněno. Jedním z úkolů bylo zamyšlení se nad specifiky metod s ohledem na jazyk dokumentu. Jelikož dva uvažované jazyky (čeština a angličtina) se velmi liší, skutečně velmi ovlivňují kvalitu výsledků. Zamyšlení nad touto skutečností se ovšem v textu omezilo na celkem obecná (a i mylná, vizte níže) konstatování.

Hodnotící kritérium:	*Způsob hodnocení - následující škálou 1 až 4:*
3. Rozsah písemné zprávy	*1=splňuje požadavky,* ***2=splňuje požadavky s menšími výhradami,*** *3=splňuje požadavky s většími výhradami,* *4=nesplňuje požadavky*

Popis kritéria:
Zhodnoťte přiměřenost rozsahu předložené ZP vzhledem k obsahu, tj. zda všechny části ZP jsou informačně bohaté a ZP neobsahuje zbytečné části.

Komentář:
Rozsah textu je přiměřený bakalářské práci. Kapitola 2 s názvem „Návrh“ obsahuje ale spíše než návrh popisy implementovaných algoritmů, následující kapitola „Realizace“ je pak okomentovaným seznamem tříd, který by zpřehlednil a částečně nahradil vhodně zvolený diagram.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*50 (E)*

Popis kritéria:
Posuďte, zda předložená ZP je po věcné stránce v pořádku, případně vyskytují-li se v práci věcné chyby nebo nepřesnosti. Zhodnoťte dále logickou strukturu ZP, návaznosti jednotlivých kapitol a pochopitelnost textu pro čtenáře.

Komentář:
Práce nepřináší zcela nové poznatky, cílem bylo hlavně prozkoumat a porovnat existující základní algoritmy pro zkracování textu. Proto považuji způsob, jak byly tyto algoritmy popsány a jejich vlastnosti diskutovány, za jeden z nejdůležitějších aspektů práce. Musím bohužel konstatovat, že práce obsahuje mnoho nešikovných formulací, ale i mnoho věcných chyb, nepřesně vysvětlených pojmů, stejně jako nepřesnosti ve vzorcích a definicích. Pro ilustraci uvedu několik příkladů:

V části 2.2.1 jsou uvedeny atributy, které byly použity pro klasifikaci vět textu (je jich 23). První z nich se jmenuje „Zkrácená množina“, jeho popis pak zní následovně: „První atribut je třída, neboli atribut, který se snažíme určit pomocí klasifikace.“ Z tohoto popisu ani poměrně poučený čtenář nepochopí, o co jde.

Popis atributů pokračuje následovně: „Ostatní atributy jsou interpretovány jako diskrétní spojité náhodné veličiny s normálním rozdělením. Spojité atributy mohou dosahovat hodnot <0, nekonečno).“ V této větě je zřejmě chybně uvedeno slovo diskrétní, je taky jasné, že normální rozdělení vylučuje omezení na nezáporné hodnoty. Vrcholem pak je, že se takto mluví o opravdu diskrétních atributech, jako je např. počet čárek, počet slov začínajících velkým písmenem apod. Takové náhodné veličiny rozhodně nelze interpretovat jako normálně rozdělené.

Dále se mluví o normalizaci hodnot parametrů (na hodnoty mezi 0 a 1). O jiné normalizaci se ale mluví i v části 2.2.2.1 u popisu „naivního Bayese“. V části 4.4.5 „Výsledky naivního Bayese“ se pak uvádí výsledky bez normalizace a s normalizací. Z textu není jasné, která normalizace byla myšlena.

Část 1.3.1 začíná větou, ve které také dochází k zjevnému zmatení pojmů, tentokrát z oblasti jazykovědné: „V mnoha jazycích jsou slova skloňována, např. kvůli časování nebo životným rodům.“

Jako ilustraci nepřesností ve značení uvedu popis klasifikátoru na straně 12: vektor atributů má nejdříve n složek, následně se ale v sumě sčítají složky od 1 do k. Na straně 28 (a jinde) je dokazováno na metriku ROUGE-N s N = 1, v definici ROUGE-N se ale žádné N nevyskytuje. Celkově by bylo lepší v definici v části 1.4.1 popsat bez použití zbytečně matoucího pojmu n-gram (jinde v textu se nevyskytne) a možná se i omezit na jediný případ použitý v práci, tedy n = 1, což vlastně z n-gramu udělá slovo ve větě.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*50 (E)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 12/2014, článek 3.

Komentář:
Práce neobsahuje mnoho překlepů. Častější již jsou chyby v interpunkci. Za nejzávažnější nedostatek však považuji formulace vět: často jsou nesrozumitelné, chybí jim podmět, slova jsou ve špatném pádě či jsou podivně řazena apod. Uvedu pár příkladů: „To je nejspíše způsobeno dvěma důvody“, „Při měření algoritmu TextRank byla relaxovaná podmínka počtu iterací konvergence skóre z původních 30 na 3000.“ (obojí str. 30)

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*70 (C)*

Popis kritéria:
Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení ZP. Charakterizujte výběr studijních pramenů. Posuďte, zda student využil všechny relevantní zdroje nebo zda se pokoušel řešit již vyřešené problémy. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Komentář:
Citace pramenů jsou vesměs uváděny na místech, kde je to vhodné. Některé reference mají zvláštní tvar (např. [15]) a některá tvrzení by si citaci zasloužila. Např. chybné tvrzení, že angličtina má menší slovní zásobu než čeština (str. 30). Při hledání zdroje této informace by student přišel na to, že angličtina je jazyk s největší slovní zásobou ze všech jazyků světa.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*50 (E)*

Popis kritéria:
Vyjádřete se k úrovni dosažených hlavních výsledků ZP, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, apod. Případně také zhodnoťte, zda software nebo zdrojové texty, které nevytvořil sám student, byly v ZP použity v souladu s licenčními podmínkami a autorským právem. Popište případnou publikační činnost a získaná ocenění související s řešením této ZP.

Komentář:
Za hlavní výstup práce lze považovat srovnání jednotlivých algoritmů na dostupných testovacích datech v kapitole 4. Algoritmy jsou spouštěny s různým nastavením: s normalizací či bez, s vynecháním stop words či po provedení stemmingu, v případě KNN se pak používají různé metriky. Výsledky jsou jen minimálně interpretovány. Není mi také jasné, proč se používá stemming a nikoli lemmatizátor. Pro výpočet některých atributů bylo potřeba určit slovní druhy slov, pro češtinu však byla použita pouze malá databáze slov (na straně 31 se mluví o 1,5 kB), přitom jsou dostupné mnohem robustnější nástroje (např. od ÚFAL MFF UK). Nikde není vysvětleno, jak se tedy atributy počítají, když se u mnoha slov slovní druh neurčí. Také se vůbec neřeší vhodnost zvolených atributů, přitom by bylo jistě možné vymyslet mnoho dalších a bylo by zajímavé se zamýšlet nad tím, jestli nějaký atribut není vhodnější pro češtinu a jiný pro angličtinu. Taková diskuze ovšem také naprosto chybí.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*

8. Komentář o využitelnosti výsledků

Popis kritéria:
Uveďte, zda hlavní výsledky ZP rozšiřují již publikované známé výsledky a/nebo přinášející zcela nové poznatky. Uveďte možnosti využití výsledků ZP v praxi.

Komentář:
Vzhledem k výše uvedeným výtkám nelze výsledky analýzy považovat za spolehlivé.

Hodnotící kritérium:	*Způsob hodnocení - nehodnotí se*

9. Otázky k obhajobě

Popis kritéria:
Uveďte případné dotazy, které by měl student zodpovědět při obhajobě ZP před komisí (body oddělte odrážkami).

Otázky:
1. Byl ve Vaší práci použit lemmatizátor? Jakým způsobem by mohl stemming pomoci při klasifikaci vět v českém jazyce, nebyl by lemmatizátor vhodnější?
2. Na straně 9 píšete, že SVD lze chápat jako organizaci textu na témata. Jak to přesně myslíte?
3. Na straně 13 píšete, že se pravděpodobnosti hodnot atributů počítají pomocí hustoty normálního rozdělení. Jak se přesně spočítá pravděpodobnost, že atribut má např. hodnotu 3?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*50 (E)*

Popis kritéria:
Shrňte stránky ZP studenta, které nejvíce ovlivnily Vaše celkové hodnocení. Celkové hodnocení **nemusí** být aritmetickým průměrem či jinou hodnotou vypočtenou z hodnocení v předchozích jednotlivých kritériích 1 až 9.

Text hodnocení:
I když měl student s implementací a studiem algoritmů jistě práci, nepovažuji výsledek za příliš povedený. Text je velmi odbytý, nesrozumitelný a plný chyb a nepřesností. Samotná analýza se omezuje na prosté spuštění naimplementovaných algoritmů a dosazení do vzorců zvolených metrik. Student neprojevil žádnou vlastní iniciativu či invenci, což by zejména v případě zkracování českého textu bylo záhodno.

Podpis oponenta práce: