Posudek oponenta závěrečné práce

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

Student: Bc. Jan Navara
Oponent práce: Ing. Petr Procházka, Ph.D.
Název práce: Compression of natural Czech text
Obor: Systémové programování

Datum vytvoření: 8. 1. 2017

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:
1. Jedná se o implementační práci, která zahrnuje: důkladné studium dostupné literatury, analýzu a návrh kompresního algoritmu, implementaci kompresního algoritmu a podrobné experimentální vyhodnocení včetně diskuse. Použitý kompresní algoritmus je netriviální, dále jsou nutné základní znalosti lingvistiky. Na druhou stranu k danému tématu existuje velké množství dostupné literatury i programových knihoven.

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 bylo beze zbytku naplněno.

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 práce je více než dostatečný v rámci definovaného zadání. Jednotlivé kapitoly podávají všechny potřebné informace. Naopak žádné přebytečné informace ("výplň") jsem nezaznamenal.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*95 (A)*

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 je logicky strukturována do kapitol: Úvod, Analýza, Návrh, Implementace, Experimenty a Závěr. Úvod obsahuje přesný popis použitých kompresních metod včetně správných citací. Velmi zajímavý je popis/přehled dostupných lingvistických nástroj. Velmi oceňuji věcnou správnost zejména u kapitoly Analýza, která obsahuje přesný popis navrhovaných algoritmů a experimentů. Méně zajímavá, ale logicky správná je následující kapitola Návrh popisující jednotlivé třídy programu, jejich odpovědnosti a propojení. Velmi kvalitně a korektně jsou zpracovány experimenty a to včetně diskuse nad dosaženými výsledky. Objevené závěry jsou poměrně zajímavé (zejména tabulky v apendixu B).

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

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:
Velmi oceňuji jazykovou úroveň celé práce. Autor dokázal zřetelně a jasně formulovat podstatné informace, aniž by obtěžoval čtenáře nadbytečným textem. Pozitivně lze hodnotit rozdělení do jednotlivých kapitol, správnou volbu obrázků, tabulek a odkazování na ně. Taktéž typograficky lze hodnotit práci velmi vysoko. Text je logicky členěn do odstavců, nové termíny, případně programová volání jsou zvýrazněna použitím odlišného fontu. Velmi přehledně jsou zpracovány tabulky s dosaženými výsledky. Čtenář se tak snadno orientuje a dokáže prezentované výsledky snadno a rychle vyhodnotit.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*85 (B)*

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:
Práce se zdroji je standardní. Počet odkazovaných zdrojů považuji za dostatečný, pouze v kapitole 1.8, která se věnuje kompresi s ohledem na lingvistiku, bych jich očekával více. Citace v textu by měly být vždy s pevnou mezerou a před interpunkčními znaky. V této práci je to často až za nimi.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*90 (A)*

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:
Dosažené výsledky jsou uspokojivé. Požadovaný program by implementován a poskytl zajímavé experimentální srovnání kompresních metod.

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:
Práce poskytuje výsledky, které dokazují vliv formální stránky jazyka na možnost jeho komprese. Jako zajímavé se jeví výsledky uvedené v tabulce B.3. Tím to bohužel končí. Dosažené výsledky potvrzují očekávání (a výsledky předchozích prací) a sice, že využitím lingvistických znalostí lze dosáhnout jistých zlepšení (v kompresním poměru), ale tato zlepšení jsou zanedbatelná a navíc (jako v tomto případě) na úkor dalších parametrů algoritmu (časové složitosti). Některé dílčí detaily dané lingvistikou jsou již běžně využívány (ať už v kompresi dat nebo v Information Retrieval). Např. case folding (na což autor sám narazil v práci), stemming, stopping. Další potenciál lingvistiky může v budoucnu být např.:
- Ve využití výchozích jazykových modelů (což autor sám zmiňuje v práci).
- Ve ztrátové kompresi přirozeného jazyka se zachování sémantické roviny textu.

Na druhou stranu pozitivně je nutné s ohledem na toto kritérium hodnotit implementaci PPM začleněnou do knihovny ExCom.

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:
Kapitola 2.8.1 Basic word-based compression:
- Vysvětlete, proč používáte samostatný model separátorů (white tokens) pro každý znak, který se objevil jakožto poslední z předcházejícího slova (black tokens)?

Kapitola 2.8.3 Experiment with part-of-speech tags (1)
- Popište, jak pracuje model M_{BT_X}, tj. model určující black token (alfanumerické slovo) na základě k předchozích alfanumerických slov stejného slovního druhu (POS) X? Nebylo by vhodnější omezit v kontextu daném k předchozími BT (alfanumerickými slovy) výsledná slova podle slovního druhu (POS)?

Kapitola 3.6 Language compression.
- Zdůvodněte zvolené hodnoty délky kontextu k pro modely M_{BT} a M_{WC}. Podle dostupné literatury by vyšší hodnoty než zvolené (k_{BT} = 1 a k_{WC} = 0) měly vést k efektivnější kompresi.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*95 (A)*

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:
Celkové hodnocení je výrazně kladné. Byly splněny všechny zadané úkoly. Logická i formální úroveň práce je nadstandardní. Taktéž zpracování experimentální části práce je velmi vyzrálé. Dosažené výsledky navržených kompresních metod jsou zajímavé, byť zůstaly za očekáváním.

Podpis oponenta práce: