Posudek oponenta závěrečné práce

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

Student: Michal Štefančík
Oponent práce: Mgr. Jan Starý, Ph.D.
Název práce: Automatické pojmenovávání skupin slov
Obor: Softwarové inženýrství

Datum vytvoření: 24. 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:
Jedná se o využití databáze WordNet se zdokumentovaným API skrze existující knihovny.

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:
Aplikaci nelze použít zamýšleným způsobem.

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:
Z číslovaných stran 1-36 jsou čtyři prázdné a čtyři tvoří seznam literatury.
Všechny části práce jsou pro práci podstatné, až na sekce 1.1.3 a 1.3.2 popisující teoretický aparát, který se v dalším nepoužije.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*40 (F)*

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:
Úvod nijak nezmiňuje existující řešení - rychlý google search odhalí články
"Text Cluster Labeling using WorrdNet" a podobné.
tf_{i,j} na straně 5 je zřejmě tf_{w_i,d_j} z předcházejícího vzorečku.
Na straně 6 se jako příklad uvádí dosud nezmíněná zkratka "SVD".
Podle str.9 provedením SVD dostaneme místo jedné matice tři, "čo je velkou výhodou" - proč?
Jedna z matic pak sestává ze "singulárných hodnot druhych mocnin eigenhodnot",
bez další zmínky o těchto pojmech. Také můžeme "zmenšit maticu sigma"
(jak, když její rozměry jsou předem dané?), a najít "najviac singifikantných konceptov",
bez dalšího vysvětlení pojmů.
Úvod kapitoly 1.4 Značkování shluků je poněkud matoucí:
jako tři možné vstupy se rozlišuje kolekce dokumentů (ze které generujeme shluky, což není náš úkol),
soubor shluků, nad kterými budeme provádět LSA a LDA (což také není náš úkol),
a konečně samotný shluk slov "s mírou relevance jednotlivých slov" (bez dalšího vysvětlení).
Zcela chybí základní input, který bych ze zadání očekával, totiž skupina slov jako "pes, kočka, liška, velryba, medvěd",
na které očekáváme odpověď "zvíře" (nebo ještě lépe: "savec"). Příkladem testovacího inputu, který nalezneme na CD,
je ale například exe/results/sensors obsahující:

alcohol;0.011594612441212345
drink;0.010585382310790977
beverage;0.009547317033786142
beer;0.006634967228855908
wine;0.006144769736936957
year;0.004616506968013172
united_states;0.004241650062428092
game;0.0041839797692611565
people;0.0037802877170926095
spirit;0.003751452570509142
country;0.003722617423925674
player;0.0035207713978414007
brand;0.003318925371757127
song;0.003261255078590192
name;0.003059409052505918
time;0.002972903612755515
term;0.0028575630264216444
number;0.0028287278798381768
vodka;0.0027710575866712415
part;0.0026845521469208385

Není jasné, jaký label pro tento shluk očekáváme ("sensors"??),
co znamenají čísla u jednotlivých slov (největší "váhu" mezi "sensors" má "alcohol"),
a proč nás takový input zajímá.

Podle str. 12 je ontologie WordNet často aktualizovaná (naposledy 2006), a aktuální verze je 3.0 (ve skutečnosti 3.1).
Na str. 13 funkce I(C) škáluje hodnoty do intervalu (-1, 1), nikoli [0,1] - jaký vliv to má na skóre kandidátů?
Podle str. 16 jsou jako kandidáti předem vyřazena všechna slova z WordNetu do úrovně 5,
což "dramaticky zvýšilo úspěšnost algoritmu". Jak budeme labelovat shluky, jejichž kandidáti
pocházejí z úrovně 5 a nižší, se dále nerozvádí, ani proč nezakázat třeba prvních 6 úrovní,
(a tím nezvýšit úspěšnost ještě dramatičtěji).

Kapitola 2.2 popisuje "generování testovacích dat".
Místo testování na předem daných, jednoduchých skupinách s jasnými labely
("kočka, pes, delfín, žížala: zvíře"; "koruna, frank, dolar, rubl, peso: měna" a podobně)
se budou generovat shluky slov s přiděleným "tématem" - podle dalšího se zřejmě myslí
slova z jednotlivých kategorií na wikipedii.

Většina takto vytvořených shluků obsahuje slova "time", "year", "version" a "part" -
zřejmě relikt z wiki stránek (last update?), což samo činí tyto shluky dosti pochybným testovacím materiálem.
Každý z nich je výsledkem čtení _abstraktů_ 1200 dokumentů - kterých přesně, kde je jejich seznam? Jakou část celé kategorie tvoří?
Na CD jsou uloženy v adresáři exe/results, je jich 14645 - to je jaká část struktury kategorií?
Příklad uvedený ve 2.3.1 (téma "The Beatles") neodpovídá skutečnosti - přiložený soubor ve skutečnosti obsahuje další slova, jako "version", "year", "time", "united_states" a "number", přičemž "number" má ve shluku "The_Beatles" větší "váhu" než "john_lennon").
Na těchto datech pak budeme testovat "úspěšnost" algoritmu (!).

Kapitola 2.3 popisuje Java třídy implementující samotný labeling
- tato sekce patří do kapitoly Implementace.

Kapitola 2.4 popisuje implementaci generátoru testovacích dat.
Na takto vygenerovaných datech předně nelze rozumně otestovat "úspěšnost" navrženého algoritmu (viz výše).
Dozvíme se nicméně podrobnosti o tom, jak SeekableByteChannel načítá řetězce pomocí ByteBuffer apod.
Samotná data se místo z (redukované) DB wikipedie čtou z obrovského 1.3GB textového souboru.

Odstavec 3.1 Volba ontologie patří do Návrhu, nikoli do Implementace.

Závěrečná kapitola 4 popisuje testování.
Korektnost samotné implementace se netestuje nijak, jedná se o testování "úspěšnosti" labelovacího algoritmu.
Postup tohoto "určení procentuelní úspěšnosti určení shluku" (tj. nalezení labelu) je dosti pochybný.
Předně, na použitých shlucích nelze kvalitu labelování rozumně testovat.
Dále, shluky, pro které jsme žádný label nenašli se vůbec neberou do úvahy (!).
Které to jsou, kolik z původních 14645 to bylo?
Za "úspěšné" se pak považuje takové pojmenování shluku, které se alespoň v jednom svém významu
shoduje s alespoň jedním významem slovního spojení z jeho nadpisu, převzatého z wikipedie,
jiný label nikdy "úspěšný" nebude. Příklad takového úspěšného či neúspěšného pojmenování uveden není,
ani elementární příklad nějakého přiřazeného labelu vůbec. Kde je seznam těch, které jsme prohlásilili za "úspěšné"?
Zároveň metodika testování dává false negatives: na shluku "pes, kočka," atd dává label "vertebrates" (obratlovci),
shluk ovšem vzešel z wiki jako Animals, takže toto pojmenování prohlásíme za neúspěch, přestože je specifičtější.

Na dvou řádcích je pak zmíněna úplně jiná sada testovacích dat, vzešlých ze stránky Enchanted Learning Wordlist
(tezaurus pro výuku angličtiny). Přesně taková data je třeba použít na skutečné testování: pod labelem "colors"
se uvádějí všechny možné barvy, pod "holidays" různé svátky, atd. Na rozdíl od shluků vzešlých z wiki abstractů
jsou to sémantick koherentní skupiny slov s očividnými označeními. Slova v těchto clusterech ovšem nemají žádné "váhy",
které algoritmus očekává na vstupu.

Nakonec se konstatuje, že algoritmus dosahuje 28% resp 21% "úspěšnosti", kterýžto údaj ale neznamená vůbec nic.
V závěru je pak algoritmus označen za nepoužitelný, a jako řešení se navrhuje použít jinou ontologii nebo vytvořit vlastní.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*60 (D)*

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:
Úroveň slovenského textu nemohu posoudit.
Anglický abstract vůbec neobsahuje členy.
Text obsahuje mnoho neslabičných předložek na konci řádku.
Na str. 7 začíná řádek uzavírající závorkou.
"mxn" apod. na str. 9 má zřejmě být $m \times n$.
Práce obsahuje prázdný seznam obrázků, rovněž obsah CD začíná prázdnou stránkou (a další stránku přetéká).
Tento obsah navíc neodpovídá skutečnosti, na DVD je 3.5GB zip file.
Použité příkazy jako wget a bzip2 nejsou v textu nijak typograficky odděleny.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
6. Práce se zdroji	*60 (D)*

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 odkazuje na 30 referencí, mezi jinými na tři přehledové články o WordNetu (9, 10, 11),
ale Fellbaum, Christiane: WordNet and wordnets (2005) referencované na "about" stránce wordnetu mezi nimi není.
Autoři jsou většinou uvedeni plnými jmény, někdy jen iniciálou. Rok vydání je někdy na konci, někdy ještě před názvem.
Názvy jsou někdy italikou, jindy ne. Poznámka "[online]" není mezerou odsazena od názvu.
V některých odkazech z neznámého důvodu zaujímá výsadní postavení země vydání,
jako např. v [4] California: Digital Government Society nebo [19] Netherlands: Computational Linguistics.

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:
Aplikace zadaná jako "automatické pojmenovávání skupin slov" by měla jako input
přijmout skupinu slov a jako output vypsat nějaké její pojmenování.

Předložená aplikace ovšem takový vstup nerozpoznává: vyžaduje adresář
předem naplněný soubory slov a jejich "váhami". Tyto "váhy" ale uživatel typicky nemá,
zajímá jej label nějaké skupiny slov (bez ohledu na jejich výskyt ve wiki abstractech nebo v jiných dokumentech).
Podle dokumentace lze tyto váhy vynechat (default = 1.0), ale aplikace pak prokazatelně dává prázdný výstup.

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:
Aplikace ve své současné podobě možná implementuje popsaný algoritmus,
ale vzhledem k metodice a úrovni "testování" nelze o jeho využitelnosti nic tvrdit.
Nicméně: pro 10703 z 14645 shluků je jedním z navržených labelů "causal_agent".
Pro 81 z druhé sady 115 shluků je jedním z navržených labelů "causal_agent",
včetně např. shluku "vegetables", který je seznamem různých druhů zeleniny.

Nejjednodušší možný interface, totiž čtení slov na standardním vstupu a psaní slov na standardní výstup,
aplikace nemá, zato očekává například specifikaci ";" jakožto oddělovače.
Bez přiřazení "vah" jednotlivým slovům vrací prázdný výstup.

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:
V čem je použitý labeling lepší než nejbližší společné hyperonymum?
Proč jako kandidáty neberete _průnik_ (tj. společná hyperonyma) místo sjednocení?
Kolik z původních 14645 testovacích shluků se "nebere v potaz" a které to jsou?
Ze kterých přesně dokumentů pocházejí shluky v results/? Jakou část celé wiki kategorie v jednotlivých případech tvoří?
Na CD jsou uloženy v adresáři exe/results, je jich 14645 - to je jaká část struktury kategorií na wikipedii?
Která pojmenování shluků v results/ a homogenresult/ byla "úspěšná", a kde je seznam "procentní úspěšnosti" vytvořených labelů?
Kolik je slov v prvních pěti úrovních hyperonymické struktury Wordnetu a která to jsou?
Kolik z 14645 testovacích shluků tak přislo o případného kandidáta na label?
Jaký label přiřadí algoritmus shluku slov, který zasahuje do těchto 5 "zakázaných" levelů?
Proč jste pro zpracování textu zvolil zrovna implementaci v Javě?
Proč na DVD distribuujete 3.5GB zip místo adresářové struktury (která rozbalená je téměř stejně velká)?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*40 (F)*

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:
Vytvořená aplikace a především doprovodný text myslím nesplňuje
elementární nároky na závěrečnou práci technického zaměření.

Podpis oponenta práce: