FAKULTA INFORMAČNÍCH TECHNOLOGIÍ ČVUT V PRAZE

Posudek oponenta závěrečné práce

Student: Peter Kanoš
Oponent práce: Mgr. Jan Starý, Ph.D.
Název práce: Sledování a analýza článků v médiích
Obor: Znalostní inženýrství

Datum vytvoření: 31. 5. 2018

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 nasazení pythonovské knihovny GenSim na texty článků.	

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: Práce splňuje zadání.	

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 odpovídající.	

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
4. Věcná a logická úroveň práce	*69 (D)*
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:
Kapitola 1 je úvod do měření sémantické podobnosti, s drobnými nepřesnostmi.

Řídkost matice Term-Document (str. 7) je způsobena tím, že dané slovo se vyskytuje jen v některých
z obrovskéo množství dokumentů, a daný dokument obsahuje jen některá ze všech slov;
nikoli tedy "mnohonásobně větším počtem slov než dokumentů".

Komentář následující po vzorečcích na str. 8 dole vysvětluje symboly d_i, d_j, a_x, b_x, které se ve vzorcích nevyskytují.
Podobně a_x, b_x na konci str. 9 jmá zřejmě být a_i, b_i z předchozího vzorečku.

Kapitola 2 popsuje existující nástroje pro textovou a sémantickou analýzu.
O NLTK se tvrdí, že má vestavěné "velké množství korpusů" (kolik? jakých?).
SEMILAR pak obsahuje "701 párů anotovaných výrazů" - proč párů? A není 701 dost málo?
Nakonec je zvolen GENSIM v Pythonu. Jeho model Doc2Vec je údajně jediný,
který umožňuje porovnávat celé texty na základě jejich obsahu - to se zdá těžko
uvěřitelné, takovou možnost mají i jiné nástroje sémantické analýzy.

Podle 2.5.1 se paragraph vector učí "spojitým distribuovaným vektorovým reprezentacím",
aniž by se zmínilo, jaké vektorové reprezentace jsou "distribuované" či "spojité".
Algoritmus, který je údajně hlavní myšlenkou Doc2Vec, se odbude jednou větou:
převést texty různé délky do vektorové reprezentace s konstantní délkou.
Jak se to děje? Jak se z nadpisu o deseti slovech stane vektor obsahující sto čísel?

U převzatého obrázku 2.1 není vysvětlen žádný z pojmů "average/concatenate" a "classifier"
a "učení vektorové reprezentace" nijak nepřibližuje. Zbytek strany 14 je změtí nikdy nevysvětlených pojmů:
"predikčná úloha ... sa vykonáva prostrednictvom viactriedneho klasifikátora",
"vektory ... sů zpriemerované na predikciu následujucého slova v kontexte",
"vektor sa správa aka pamäť, v ktorej je uložené chýbajúce slovo do daného kontextu",
"kontexty ... sú vzorkované z posuvného okna nad odsekom",
"pri každom kroku stochastickej aproximácie ... vypočítame gradient chýb zo siete".
Použitá metoda není popsána nijak blíže než takto (!).

Kapitola 3 popisuje stahování článků z iDnes.cz a z Novinky.cz a jejich zpracování.
Podezřele působí tvrzení z 3.5, že měnící se číselné údaje u sportovních zpráv (typicky výsledky zápasů)
"způsobovaly zkreslenost měřených podobností", takže je ignorujeme.
Nemáme právě takové změny v podobnosti sledovat?

Jako trénovací korpus je zvolen dump české wikipedie, s konstantní délkou reprezentujícího vektoru 100.
Jak takové vektory vzniknou, je popsáno následovně: "texty ... sa pomocou metody infer_vector prevedú
do vektorovej reprezentácie určenej daným modelom". Víc o natrénovaných datech nevíme.

Následuje ukázka článku, jehož titulek a perex mají podobnost 0.31873 (kteréžto číslo ale dosud nic neznamená).
Tato "relatvně malá podobnost" (relativně malá vůči čemu?) je vysvětlena mimo jiné tím,
že "reaguje Moskva" je v přítomném čase, kdežto "Moskva zareaguje" v budoucím,
"čo vyvoláva pocit vzájomného so protirečenia". Od té chvíle musí být čtenář ve střehu.
Dalším vlivem může být "náhodná inicilizace vektorů" při infer_vector (?).
(Kolik ze 100 čísel vektoru odpovídajícího krátkému nadpisu je tedy ve skutečnosti _náhodných_?)

Krátkou poznámou se přejde i to, že "model ve svém slovíku neobsahuje slova, která jsou obsažena v samotném textu".
Kolik typicky/nejméně/nejvíce slov v článku takto neznáme, kolik článků se to týká, jakou část celku tvoří?
Kolik takových slov je, která to jsou, jakou část všech slov tvoří, kde je jejich seznam?
A proč tedy nepoužíváme jiný model? Nic z toho se nedozvíme.

Sekce 3.7 je statistickým přehledem výsledných hodnot sémantické podobnosti,
tak jak je GENSIM přiřadil částem jednotlivých článků (titulek, perex, text).
Začíná třemi nejpodobnějšími dvojicemi titulek-perex, přičemž hodnota pro nejpodobnější dvojici
(totiž podobnost 0.741521) "přesně odpovídá skutečnosti" (?). (Neodpovídalo by skutečnosti spíše 0.736417?)

Následuje ukázka dvojic titulek-perex s malou podobností, kdy perex nerozvádí titulek,
ale podává jiné, další informace. Tabulky 3.5 a 3.6 zachycují statistiku podobnosti
mezi titulky, perexy a texty článků z Novinek resp. iDnes. Střední hodnota u obou je
v rozmezí od 0.22 do 0.33, se standardní odchylkou okolo 0.11 až 0.14.
Tomu těžko přiřadit jiný statistický význam než zjištění,
že některé titulky/perexy/texty jsou si podobné více a některé méně.

Rozdílná podobnost jednotlivých částí je "očividně způsobena jejich rozdílnými délkami".
Jednak je jistě způsobena především tím, že se jedná o různé texty,
a jednak těžko předpokládat, že porovnávané texty budou stejně dlouhé,
na tom přeci nemůže sémantická metoda záviset.

Sekce 3.8 pak analogicky srovnává podobnost různých verzí článků, a sekce 3.9 dva články na totéž téma z obou novin.
Míra jejich podobnosti, vzešlá z použitého modelu, je okomentována takto: "výsledná podobnost textov, titulkov aj perexov prekonáva 99% kvantil porovnania častí rozných _náhodne_vybratých_článkov_".

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
5. Formální úroveň práce	*80 (B)*

Popis kritéria:
Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku ZP, viz Směrnice děkana č. 26/2017, článek 3.

Komentář:
Úroveň slovenského textu si netroufám posoudit. Překlepů a očividných prohřešků ale není více, než je obvyklé.
Chybějící či chybné popisy u obrázků a vzorců jsou podstatnějším prohřeškem.

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:
Práce odkazuje na 20 referencí. Všechny se zdají být relevatní, ale na text mining by se jistě našla
lepší reference než náhodné slajdy přednášek z muni.cz ([5]). Jmémo autora je někdy iniciálou za příjmením,
někdy plným jménem před příjmením. Měsíc je u čísla zahraničního časopsu někdy česky, někdy anglicky.

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
7. Hodnocení výsledků, publikační výstupy a ocenění	*70 (C)*

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:
Hlavním výsledkem je myslím crawler stahující samotné články.
Kód spouštějící knihovnu GENSIM nad staženými texty je pak přímočarou aplikací.

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:
Crawler stahující články z iDnes a Novinek lze použít při budování korpusu takových článků.
Naproti tomu čísla získaná aplikací sémantického enginu nad tímto korpusem myslím k ničemu dalšímu využít nelze.

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:
Kolik je v češtině slov, a kolik z nich se vyskytlo ve stažených článcích?
Proč ukládáte stažené články do CSV souborů místo do nějaké relační databáze?
V čem je lepší trénovat sémantický engine na wikipedii, místo na (starších) článcích z týchž novin?
Jak vzniknou v modelu GENSIM reprezentující vektory délky 100? Jaký vektor je
přiřazen textu "Palestina a Izrael zahájí další kolo mírových jednání" a co oněch
100 reálných čísel _znamená_ ve vztahu k tomuto textu?
Kolik je slov, která se vyskytla ve článcích, ale model je nezná? Jakou část všech slov tvoří?
Kolik jich je nejvíce/průměrně v jednom článku? Kolika článků se to týká, jakou část tyto články tvoří?
Co jsou body na obrázku 3.1, přesahující vysoko/hluboko do extrému?

Hodnotící kritérium:	*Způsob hodnocení - bodové hodnocení 0 až 100 bodů (známka A až F):*
10. Celkové hodnocení	*70 (C)*

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:
Můj celkový dojem je ten, že si řešitel vyzkoušel spuštění sémantického enginu. Možná je tak zadání myšleno.
Výsledkem je pak sada čísel, o kterých nevíme, co přesně znamenají, i když na uvedených ukázkách mohou dávat jakýsi smysl:
subjektivně podobnější texty jsou skutečně označeny jako podobnější. Jak a proč se to děje se ovšem nedozvíme.

Podpis oponenta práce: