

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Automatické generování názvu a popisek webových stránek s neznámým obsahem
Jméno autora:	Jonáš Amrich
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra kybernetiky
Oponent práce:	RNDr. Jiří Materna, Ph.D.
Pracoviště oponenta práce:	Seznam.cz, a.s.

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	náročnější
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
Předmětem práce byla syntéza textů v přirozeném jazyce, což je v obecnosti velmi obtížná úloha, která dosud v dostupné literatuře nebyla uspokojivě vyřešena.	

Splnění zadání	splněno s menšími výhradami
<i>Posuďte, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.</i>	
Úkolem práce bylo zejména navrhnout algoritmy pro automatické generování titulků a popisů stránek. Generování popisů bylo splněno bez výhrad. Generované titulky jsou však pouze ve formě výčtu klíčových slov. V zadání práce není specifikováno, jak má titulek vypadat, ale očekával bych víceslovný výraz nebo větu.	

Zvolený postup řešení	správný
<i>Posuďte, zda student zvolil správný postup nebo metody řešení.</i>	
Ke zvolenému postupu řešení nemám žádné výhrady.	

Odborná úroveň	A - výborně
<i>Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.</i>	
Student prokázal odbornou úroveň odpovídající bakalářskému stupni vzdělání.	

Formální a jazyková úroveň, rozsah práce	A - výborně
<i>Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.</i>	
Přestože se v práci vyskytuje několik drobných jazykových chyb, není to na závadu čitelnosti nebo srozumitelnosti. Velice oceňuji, že byla práce napsána v anglickém jazyce, a to na velmi dobré úrovni. Text je dobře strukturovaný a typograficky v pořádku. Rozsah textu odpovídá nárokům na bakalářskou práci.	

Výběr zdrojů, korektnost citací	A - výborně
<i>Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. 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.</i>	
K výběru zdrojů a korektnosti citací nemám žádné výhrady.	

Další komentáře a hodnocení
<i>Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.</i>

Zdrojové kódy vytvořených programů jsou podle autora umístěné na serveru Github.com. Uvedená webová adresa však neexistuje nebo není veřejně dostupná.

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Práce řeší netriviální problém na pomezí oblastí zpracování přirozeného jazyka, strojového učení a extrakce informací. Student vyšel z existujících přístupů popsaných v odborné literatuře, které doplnil o vlastní nápady a uspokojujivě tak naplnil zadání práce.

K práci mám následující připomínky a otázky:

- Vzorec 3.3 na straně 11 definuje způsob vytváření vektoru reprezentujícího kontext jako součet vektorů jednotlivých dokumentů. Velikost takového vektoru však bude ovlivněna počtem dokumentů, ze kterých je kontextový vektor sestaven. Vektory různých kontextů tudíž nemusí být porovnatelné v případě, že se bude počet dílčích dokumentů výrazně lišit. To může mít za následek horší generalizaci algoritmu. Nebylo by vhodné kontextové vektory nějakým způsobem normalizovat?
- Použitá experimentální data byla rozdělena na trénovací a testovací část. Z textu však není patrné, pomocí které z těchto množin byly určeny parametry použitých algoritmů. Pro hledání parametrů se doporučuje používat třetí – validační – datovou množinu a výsledky reportovat na testovací množině, která nebyla použita ani pro trénování ani pro hledání optimálních parametrů.
- Pro předzpracování textových dat nebyly použity žádné pokročilejší techniky jako lemmatizace, detekce jmenných entit apod. Domnívá se autor, že by jejich použití mohlo vést k lepším výsledkům?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **B - velmi dobře**.

Datum: 4.6.2015

Podpis: