POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Extrakce klíčových slov z textových shluků
Jméno autora:	Diana Korladinova
Typ práce:	bakalářská
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	KK
Oponent práce:	Gustav Šír
Pracoviště oponenta práce:	KP

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Zadani se jevi jako adekvatne narocne pro bakalarskou praci.

Splnění zadání **splněno**

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.

Bez vyhrad splneno.

Zvolený postup řešení **vynikající**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postup je vyborny, studentka zacina obsahlou resersi a experimentuje s sirokou skalou vybranych metod.

Odborná úroveň **B - velmi dobře**

Posuďte úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Studentka prozkoumava a pracuje s nejruznejsimi metodami na rozhrani NLP a strojoveho uceni, z velke casti za hranicemi bakalarskeho studia. Misty je lehce znat, ze porozumeni temto metodam je spise intuitivni a lehce nepresne, ale to je zcela pochopitelne. Trochu mi chybi nejake explicitnejsi porovnani metod a nejaky kvantifikovany zaver, ale je to opet pochopitelne vzhledem k unsupervised povaze ulohy.

Formální a jazyková úroveň, rozsah práce **A - výborně**

Posuďte správnost používání formálních zápisů obsažených v práci. Posuďte typografickou a jazykovou stránku.

Textova cast prace je velmi povedena, jen anglictina je misty lehce počešťená, ale citelnosti (pro cecha) to nijak neubira. Po typograficke strance velmi pekna, zpracovana se zjevnou peclivosti a citem pro detail (i barvy).

Výběr zdrojů, korektnost citací **A - výborně**

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.

Bez problemu.

Další komentáře a hodnocení

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.

Viz celkove hodnoceni.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

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.

Jedna se o peknou bakalarskou praci s cilem predevsim prozkoumat metody NLP v oblasti detekce shluku temat a jejich anotace na datasetu ceskeho Twitteru. Jedna se zjevne o pocatecni explorativni analyzu, ktera byla pro studentku nova, a poprala se s ni velmi pekne. Ocenuji predevsim siri metod, se kterymi se musela efektivne seznamit (od k-means az po transformery).

Poznamky/otazky:

Trochu mi chybi nejaky zaver/diskuze z hlediska vysledku, jakou metodu byste hodnotila jako nejslibnejsi?

5.1.1 – experiment A vs. B – z hlediska detekce slov, ktere maji charakterizovat jednotlive clustery (a ne tweety) mi dava varianta A vyrazne vetsi smysl, nemyslite? Pokud ano, cim si vysvetlujete ze vychazi podobne jako varianta B?

Vysledky dvou stezejnich postupu pomoci k-means a LDA by mozna bylo pekne nejak blize porovnat (neni mi jasne jak se doslo k zaveru ze k-means je lepsi).

+ par veci, ktere mi prisly vytrzene z kontextu (evokujici limitovane porozumeni danym metodam):

2.1.2: Methods that embed the vectors into an R n space and consider semantic relations and meaning (word order is taken into account)

- Jak souvisi poradi slov s delenim na counting/embedding?

Podobne:
the conditional distribution is a quotient of the joint probability distribution of z, β, and θ across all w and the probability of observing D across all topics

A high-dimensional space means multiple relations embedded for a single word.

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A - výborně.**

Datum: 10.6.2023 Podpis: