POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	**Sémantické shlukování dat z Twitteru**
Jméno autora:	**Jan Petrov**
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	Katedra počítačů
Oponent práce:	Ing. Jiří Kubalík, Ph.D.
Pracoviště oponenta práce:	CIIRC, ČVUT

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání **průměrně náročné**

Hodnocení náročnosti zadání závěrečné práce.

Tato práce se zabývá úlohou sémantického shlukování dat z Twitteru. Výstupem je SW nástroj, který kombinuje několik pokročilých metod strojového učení, jako jsou modely typu BERT a Set Transformery. Dále se analyzují klasické shlukovací algoritmy a shlukovací algoritmus založený na neuronových sítích, Deep Amortized Clustering. Podle mého názoru jde o zajímavé zadání o standardní náročnosti.

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.

Student splnil zadání ve všech bodech.

Zvolený postup řešení **správný**

Posuďte, zda student zvolil správný postup nebo metody řešení.

Student postupoval přesně dle zadání. Nastudoval relevantní literaturu se zaměřením na metody zpracování přirozeného jazyka založené na moderních modelech typu Set Transformer. Stejně tak nastudoval oblast shlukovacích algoritmů, zaměřil se na klasické K-means a DBSCAN plus algoritmus Deep Amortized Clustering založený na neuronové síti. Naimplementoval anotační nástroj, který kombinuje několik metod sémantického vyhledávání. S využitím tohoto nástroje vytvořil z více než 200 tis. tweetů 65 anotovaných shluků, které zrevidoval na vzorku 18 tis. tweetů. Nakonec experimentálně otestoval shlukovací metody a dosažené výsledky kriticky zhotnotil.

Odborná úroveň **A - výborně**

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

Odbornou úroveň hodnotím velice kladně. Student prokázal schopnost nastudovat, pochopit a použít state-of-the-art metody strojového učení.

Formální a jazyková úroveň, rozsah práce **Zvolte položku.**

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

Po jazykové stránce nemám výhrad. Práce je napsána velice dobrou a čtivou angličtinou, s malým množstvím chyb a překlepů. Pozitivně hodnotím i pěkné ilustrativní obrázky, které napomáhají pochopení problematiky.
Výtku mám pouze k odkazování na obrázky a k jejich popisu. Na některé obrázky není v textu žádný odkaz. Například na obrázek 7.1. Tento obrázek (i některé další) by si také zasloužil podrobný popis. Chvíli mi trvalo, než jsem jej pochopil.

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.
Počet citovaných zdrojů je nadprůměrný. Jedná se o velice relevantní a aktuální zdroje.

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.
Student implementoval robustní nástroj pro semi-automatickou anotaci a sémantické vyhledávání. Po stránce SW inženýrské, je to pěkné dílo s příjemným uživatelským rozhraním a pěknými interaktivními grafickými výstupy.

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.

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

Do diskuze mám následující otázky:

1. Máte nějakou hypotézu, proč DBSCAN fungoval mnohem hůř než K-means?
2. Ještě ke K-means a k poznatku, že K-means produkuje shluky, které nekorespondují s ručně anotovanými skupinami, viz poslední odstavec sekce 7.1. Nešlo by algoritmu nějak „napovědět“, jaká témata by neměl opomenout? A zbývající ať si dohledá.

Datum: 9.6.2022 Podpis: