ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

POSUDEK OPONENTA ZÁVĚREČNÉ PRÁCE

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Shlukování a modelování chování uživatelů založené na datech z webového prohlížeče
Jméno autora:	Bc. Jan Žaloudek
Typ práce:	diplomová
Fakulta/ústav:	Fakulta elektrotechnická (FEL)
Katedra/ústav:	KP
Oponent práce:	Gustav Šourek
Pracoviště oponenta:	KP

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání — náročnější

Hodnocení náročnosti zadání závěrečné práce.

Téma je velmi atraktivni, nijak zvlášť odborně náročné, avšak zahrnuje složitější práci s daty.

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 problému splněno.

Zvolený postup řešení — vynikající

Posuďte, zda student zvolil správný postup nebo metody řešení.

Postupu není co vytknout, student prozkoumal množství zajímavých alternativ.

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.

Student skvěle využil získané znalosti ze strojového učení, které jsou v práci v široké škále promítnuté. Je to spíše aplikační práce, z odborného hlediska je standardní.

Formální a jazyková úroveň, rozsah práce — A - výborně

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

Text práce je velmi profesionální, srozumitelný, přehledně členěný. Chválím i TEXovou šablonu. V práci je jen pár chybějících slov a překlepů (např. nadpis Experimnets).

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.

V pořádku (až na první citaci a pár dalších neúplných záznamů).

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.

Cílem předložené práce je učení efektivní vektorové reprezentace chování uživatelů na základě zaznamenaných sekvencí jejich interakcí na webu, k čemuž měl student reálná data společnosti Smartlook. Jedná se o atraktivní, aktuální a praktické téma. Student k řešení přistupuje pomocí analogií s vybranými NLP technikami, což je celkem bežný, relevantní a validní postup. Pro použití vybraných technik, tj. převod na problém vektorové reprezentace lineárního textu, použil student množství zjednodušujících předpokladů tak, aby odpovídajícím způsobem změnil charakter dat.

Do samotného postupu pak student zapojil širokou škálu metod strojového učení včetně zajímavých NLP technik. Vše je velmi srozumitelně a věcně popsáno - struktura teoretického úvodu, metodologie, dílčí výsledky, interpretace a závěry – jsou skvěle, profesionálně členěné.

Konkrétní poznámky, které lze chápat i jako otázky:

- mluvíte o struktuře vstupních záznamů, ale ta je ve výsledku ignorována, stejně jako struktura vět v použitých NLP metodách
  - např. URL není jen sekvence slov, je to orientovaný strom, stejně jako DOM
  - a nakonec i samotná sekvenčnost je v použitých modelech ignorována (pouze samplujete jednotlivá slova/eventy z kontextu)
- doc2vec/paragraph2vec byla, alespoň v čase zveřejnění a portování do Gensimu, dost kontroverzní práce s nereproducibilními výsledky (řekl sám Mikolov)
  - osobně jsem ji zkoušel a lepší výsledky mělo i jednoduché průměrování embeddings slov
- Navíc zvolené „neurální architektury“ (w2v,d2v) v principu pouze aproximují výpočet maticového rozkladu nad co-occurence maticí slov (viz např. Levy&Goldberg,2014)
  - čili oba zvolené postupy (frekventistický a neurální) dělají v principu to samé

Jako future work, když už experimentujete s ANNs, bych doporučoval spíš více využít strukturu vstupu, alespoň tu sekvenčnost (např. RNNs), nebo lépe celý strom (např. TreeNNs).

Předloženou závěrečnou práci hodnotím klasifikačním stupněm A - výborně.

Datum: 31.05.18 Podpis: