

Údaje o bakalářské práci

Osobní číslo:	A17B0318P	Datum zadání:	7. října 2019
Jméno a příjmení:	Jan PAŠEK	Plánované datum odevzdání:	7. května 2020
Obor/kombinace:	Informatika (INF)	Datum odevzdání:	
Zadané téma:	Využití duplicitních otázek na Stackoverflow pro učení reprezentace významu vět		
Stav práce:	Rozpracovaná práce		

Údaje o kvalifikační práci

1. Hlavní téma

Využití duplicitních otázek na Stackoverflow pro učení reprezentace významu vět

2. Hlavní téma v angličtině

Learning of sentence encoding by using duplicate questions from Stackoverflow.

3. Název dle studenta

Learning of Sentence Encoding by Using Duplicate Questions from Stackoverflow

4. Název dle studenta v angličtině

Learning of Sentence Encoding by Using Duplicate Questions from Stackoverflow

5. Souběžný název

6. Podnázev

7. Anotace (krátký popis práce)

Tato bakalářská práce se zabývá vývojem neuronové sítě pro porozumění textu v odborném jazyce. Výstupy této práce mohou zlepšit výsledky úloh jako je získávání informací či generování zdrojového kódu. Pro vyřešení této úlohy představujeme novou architekturu neuronové sítě založenou na využití enkodéru kódů společně s textovým enkodérem. Architektura dále využívá nepříliš známou f1 loss, která významně zlepšuje dosažené výsledky. Důležitým výstupem této práce je vektorová reprezentace vět, která se nalézá ve skrytých vrstvách neuronové sítě. Navržený přístup je demonstrován na využití duplicitních otázek ze stránky Stackoverflow, ze kterých jsme připravili nový dataset použitelný nad rámec této práce. Pomocí navržené architektury bylo na datasetu dosaženo f1 score 74.1 %, což představuje zlepšení o 5.1 % v porovnání s výchozí architekturou založenou na sčítání reprezentací slov.

8. Klíčová slova (oddělujte čárkou)

strojové učení, zpracování přirozeného jazyka, sémantická podobnost, Stackoverflow, neuronové sítě

9. Anotace v angličtině (krátký popis práce)

This bachelor thesis aims to create a neural network for natural language understanding in expert domains. Our outcome can significantly improve tasks such as information retrieval or code generation. The work proposes a neural network architecture utilizing a code encoder in parallel with a commonly used text encoder. Furthermore, the architecture uses a not widely known f1 loss, significantly improving results. An important outcome of this work is a vector representation of text stored in hidden layers of the network. We demonstrate our approach on Stackoverflow data utilizing duplicate questions to create a novel dataset, usable beyond the scope of this work. Our architecture achieved f1 score of 74.1%, which is a 5.1% improvement compared to a baseline model based on word embedding summation.

10. Anglická klíčová slova (oddělujte čárkou)

machine learning, natural language processing, semantic similarity, Stackoverflow, neural networks

11. Přílohy volně vložené

1 CD ROM se zdrojovými kódů

Údaje o bakalářské práci

Osobní číslo:	A17B0318P	Datum zadání:	7. října 2019
Jméno a příjmení:	Jan PAŠEK	Plánované datum odevzdání:	7. května 2020
Obor/kombinace:	Informatika (INF)	Datum odevzdání:	
Zadané téma:	Využití duplicitních otázek na Stackoverflow pro učení reprezentace významu vět		
Stav práce:	Rozpracovaná práce		

12. Přílohy vázané v práci
tabulky

13. Rozsah práce
81 s. (87 600 znaků)

14. Jazyk práce
AN

15. Záznam průběhu obhajoby

16. Zásady pro vypracování

1. Seznamte se se základními principy hlubokého učení, neuronových sítí a nástroje Tensorflow. Naprogramujte základní úlohy v nástroji Tensorflow.
2. Seznamte se s formátem dat použitých v offline kopii Stackoverflow (nebo obdobného zdroje dat) a extrahuje duplicitní otázky.
3. Aplikujte metody pro učení reprezentace významu vět na získaná data.
4. Změřte úspěšnost a proveďte kritické zhodnocení dosažených výsledků.

17. Seznam doporučené literatury

Dodá vedoucí bakalářské práce.

18. Osoby VŠKP

Vedoucí bakalářské práce: Ing. Miloslav Konopík, Ph.D.
Katedra informatiky a výpočetní techniky

Oponent bakalářské práce: Ing. Ondřej Pražák
Nové technologie pro informační společnost

Elektronická forma kvalifikační práce

Soubor je vložen, ale není zveřejněn (práce nebyla obhájena...)

Posudky kvalifikační práce

Posudek(y) oponenta:

Hodnocení vedoucího:

Soubor s průběhem obhajoby:

Údaje o bakalářské práci

Osobní číslo:	A17B0318P	Datum zadání:	7. října 2019
Jméno a příjmení:	Jan PAŠEK	Plánované datum odevzdání:	7. května 2020
Obor/kombinace:	Informatika (INF)	Datum odevzdání:	
Zadané téma:	Využití duplicitních otázek na Stackoverflow pro učení reprezentace významu vět		
Stav práce:	Rozpracovaná práce		

Potvrzuji správnost vložených údajů a potvrzuji plnou shodu elektronické verze s odevzdávanou listinnou verzí VŠKP.

Datum: 6.5.2020

Podpis:

