KLASIFIKÁCIA WEB STRÁNOK POMOCOU METÓD STROJOVÉHO UČENIA

Vedúci práce: RNDr. Ľubomír Antoni, PhD.

Konzultant: RNDr. Šimon Horvát

Študent: Martin Bača

MOTIVÁCIA

- Potreba firmy GlobalLogic kategorizovať svojich zákazníkov
- Využitie výsledkov práce na riešenie konkrétneho problému vo firme GlobalLogic

CIELE

1. Spracovať prehľad metód strojového učenia v oblasti textovej analýzy.
2. Navrhnúť a extrahovať vhodné atribúty na klasifikáciu webových stránok s využitím scrapovania stránok.
3. Implementovať metódy strojového učenia na klasifikáciu webových stránok podľa definovaných kategórií a porovnať dosiahnuté výsledky.

POSTUP RIEŠENIA

- Vytvorenie datasetu scrapovaním údajov
- Predspracovanie vstupov
- Použitie metód na spracovanie prirodzeného jazyka (NLP)
- Použitie metód strojového učenia
- Porovnanie výsledkov pre použité metódy strojového učenia

VYTVORENIE DATASETU

- Posielanie dopytov na servery stránok
- Scrapovanie údajov použitím knižnice BeautifulSoup
- Transformácia stiahnutých dát do DataFrame-u

	.	name	link	type	title	h1
1	0	Australian Nat…	http://www.anu.edu…	Education	ANU	<null>
2	1	Monash Univers…	http://www.monash.…	Education	Monash University…	Home
3	2	University of …	http://www.uq.edu.…	Education	The University of…	\n …
4	3	Macquarie Univ…	http://mq.edu.au/	Education	\n \n\n …	Macquarie Uni…
5	4	UNSW Sydney	https://www.unsw.e…	Education	UNSW Sydney | One…	Explore our p…
6	5	Newcastle Univ…	http://www.ncl.ac.…	Education	The things we do …	We\xe2\x80\x9…
7	6	University of …	https://www.uow.ed…	Education	Home - University…	<null>
8	7	University of …	http://www.unimelb…	Education	The University of…	\r\n …
9	8	University of …	http://www.utas.ed…	Education	University of Tas…	\r\n …
10	9	University of …	http://www.adelaid…	Education	The University of…	The Universit…
11	10	James Cook Uni…	http://www.jcu.edu…	Education	Study at James Co…	Pushing for s…
12	11	Flinders Unive…	http://www.flinder…	Education	Flinders Universi…	<null>
13	12	RMIT University	https://www.rmit.e…	Education	RMIT University -…	RMIT University
14	13	La Trobe Unive…	http://www.latrobe…	Education	La Trobe Universi…	<null>
15	14	Victoria Unive…	http://www.vu.edu.…	Education	Victoria Universi…	VU home
16	15	University of …	http://www.une.edu…	Education	Home - University…	Home
17	16	Deakin Univers…	http://www.deakin.…	Education	Home | Deakin	Put Deakin fi…
18	17	Griffith Unive…	http://www.griffit…	Education	Griffith Universi…	<null>
19	18	Central Queens…	https://www.cqu.ed…	Education	CQUniversity	<null>
20	19	University of …	https://www.unisa.…	Education	UniSA - Universit…	University of…
21	20	Swinburne Univ…	http://www.swinbur…	Education	Swinburne Univers…	\n R…
22	21	Bond University	http://bond.edu.au/	Education	Bond University |…	Experience Bo…
23	22	University of …	http://www.usc.edu…	Education	UniSC | Universit…	UniSC | Unive…
24	23	Charles Sturt …	http://www.csu.edu…	Education	Home - Charles St…	<null>
25	24	University of …	http://www.canberr…	Education	\n\nUniversity of…	\r\n Unive…
26	25	Federation Uni…	https://federation…	Education	Federation Univer…	\r\n \r\n

PRÍKLADY VSTUPOV

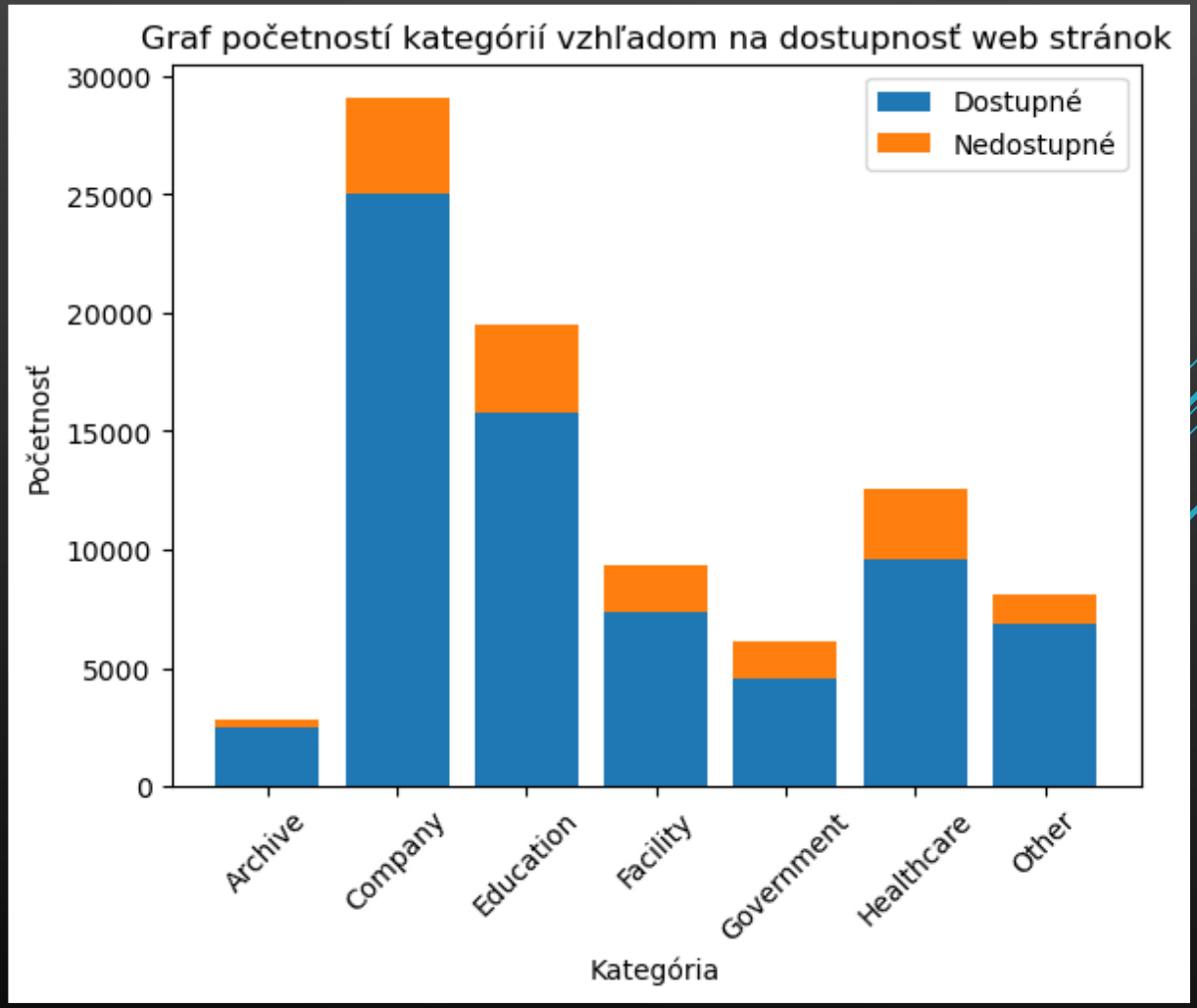
ΑρχικήΕΠΙΨΥ \n\t\t\t\t\t\t\t\t\tslide1\t\t\t\t\t\t\t\tΤο ΕΠΙΨΥ ιδρύθηκε το 1989 από τον Καθηγητή Κώστα Στεφανή σύμφωνα με τις διατάξεις των ν. 1514/85 και ν. 1771/88, άρθρο 10, αρχικά ως Ακαδημαϊκό Ερευνητικό Ινστιτούτο Ψυχικής Υγιεινής (Π.Δ. 268/89). Προσαρμόστηκε στις διατάξεις του άρθρου 17 του ν. 2083/92 με το Π.Δ. 217/96 και μετονομάστηκε σε Ερευνητικό Πανεπιστημιακό Ινστιτούτο Ψυχικής Υγιεινής (ΕΠΙΨΥ). Τα θέματα διοίκησης και λειτουργίας των ΕΠΙ ρυθμίζονται σήμερα με τους ν. 3685/2008 (άρθρο 11), ν. 3794/2009 (άρθρο 32) και ν. 4009/2011 (άρθρα 58 και 80).Το 2011 το ΕΠΙΨΥ εντάχθηκε από την ΕΛΣΤΑΤ στο μητρώο φορέων γενικής κυβέρνησης, υπαγόμενο στις διατάξεις του ν. 3429/2005, Κεφ. Α.\n

The University of Melbourne, Australia - Australia\'s best university and one of the world\'s finest \r\n Discover Melbourne at Dookie\r\n Featured articles Global pandemic therapeutics centre announced in Melbourne

The Alfred | Alfred Health \n\t\t\t\t\t\tWelcome to The Alfred\n\t\t\t\t\t \n\t\t\t\t\t\tOne of Australia\'s busiest hospitals, The Alfred provides specialised services and care for Victoria\'s most complex patients.\n\t\t\t\t\t \n\t\t\t\t\t Our services & clinics\n\t\t\t\t \n\t\t\t\t\t\t\t\t\tThe Alfred\n\t\t\t\t\t\t\t\t Have some feedback for us? Use our to let us know. \n\t\t\t\t\t\t\t \n\t\t\t\t\t\t b' html \n\n Skip to main content

ROZLOŽENIE PRÍKLADOV MEDZI KATEGÓRIAMI

- Dataset je nevyvážený



FILTROVANIE A PREDSPRACOVANIE VSTUPOV

- Odstránenie URL adries a špeciálnych znakov
- Odstránenie krátkych slov, t. j. menej ako 3 znaky (on, of,...)
- Odstránenie nežiaducich slov, ktoré nám nedávajú žiadnu informáciu (for, from, the,...)
- Lemmatizácia slov
- Filtrovanie webových stránok, ktoré nie sú v angličtine

LEMMATIZÁCIA

- Úprava slov na ich základnú slovníkovú podobu
- Zjednotenie variantov slov
- Uľahčuje proces spracovania textu

- Odvodené slová majú rovnakú lemmu:
  - beží**m** => bežať
  - beží**š** => bežať
  - beži**a** => bežať

- Slová, ktoré nie sú odvodené majú rozdielne lemmy:
  - beží**š** => bežať
  - chod**ia** => chodiť

VEKTORIZÁCIA TEXTU

- Prevod reťazcov na vektorové reprezentácie
- Spôsob ako modely dokážu pracovať s textom
- Reťazce s podobnými významami majú podobné vektory

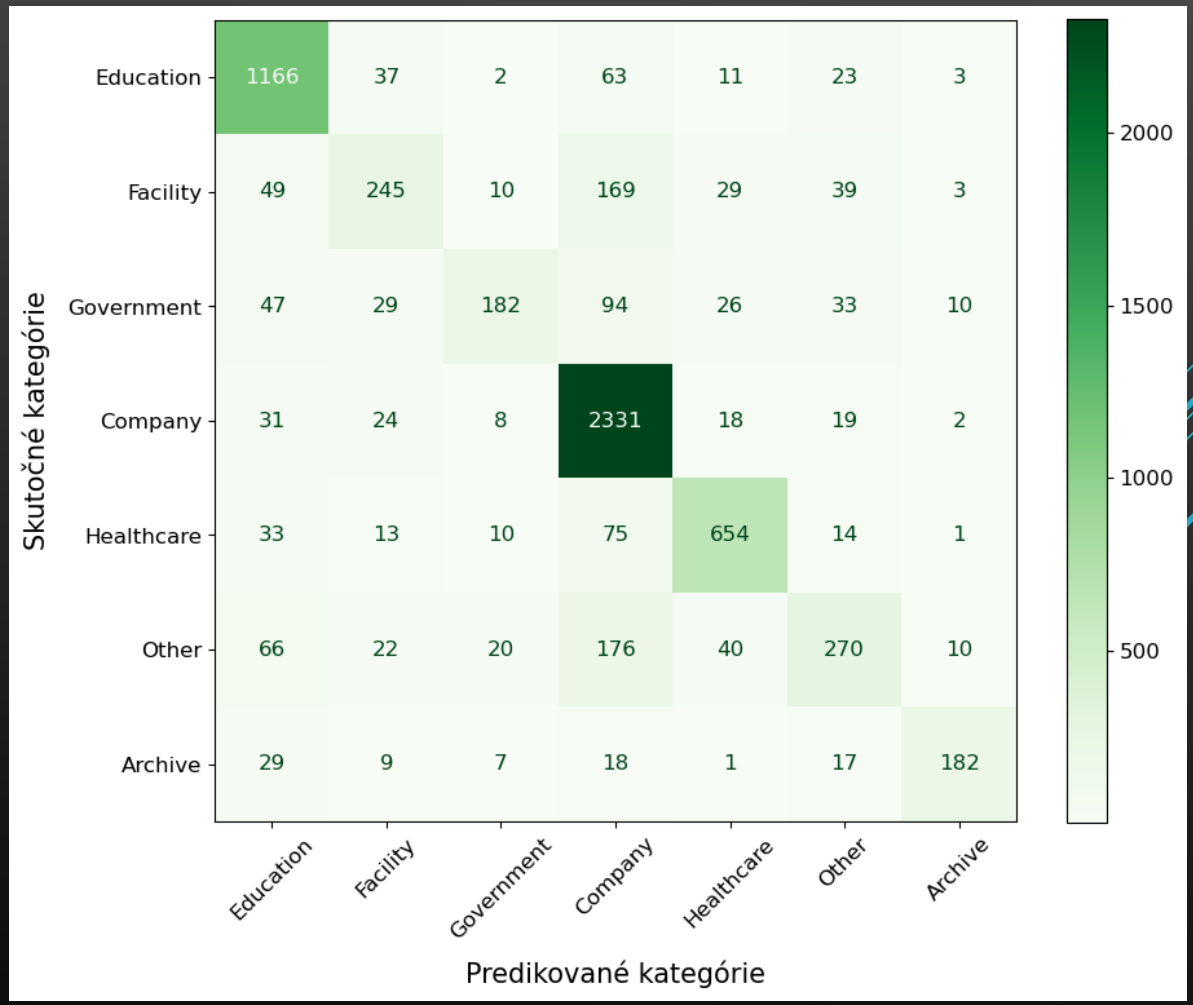
POUŽITÉ METÓDY STROJOVÉHO UČENIA

- Naivný Bayes
- Metóda podporných vektorov
- Konvolučné neurónové siete

VÝSLEDKY

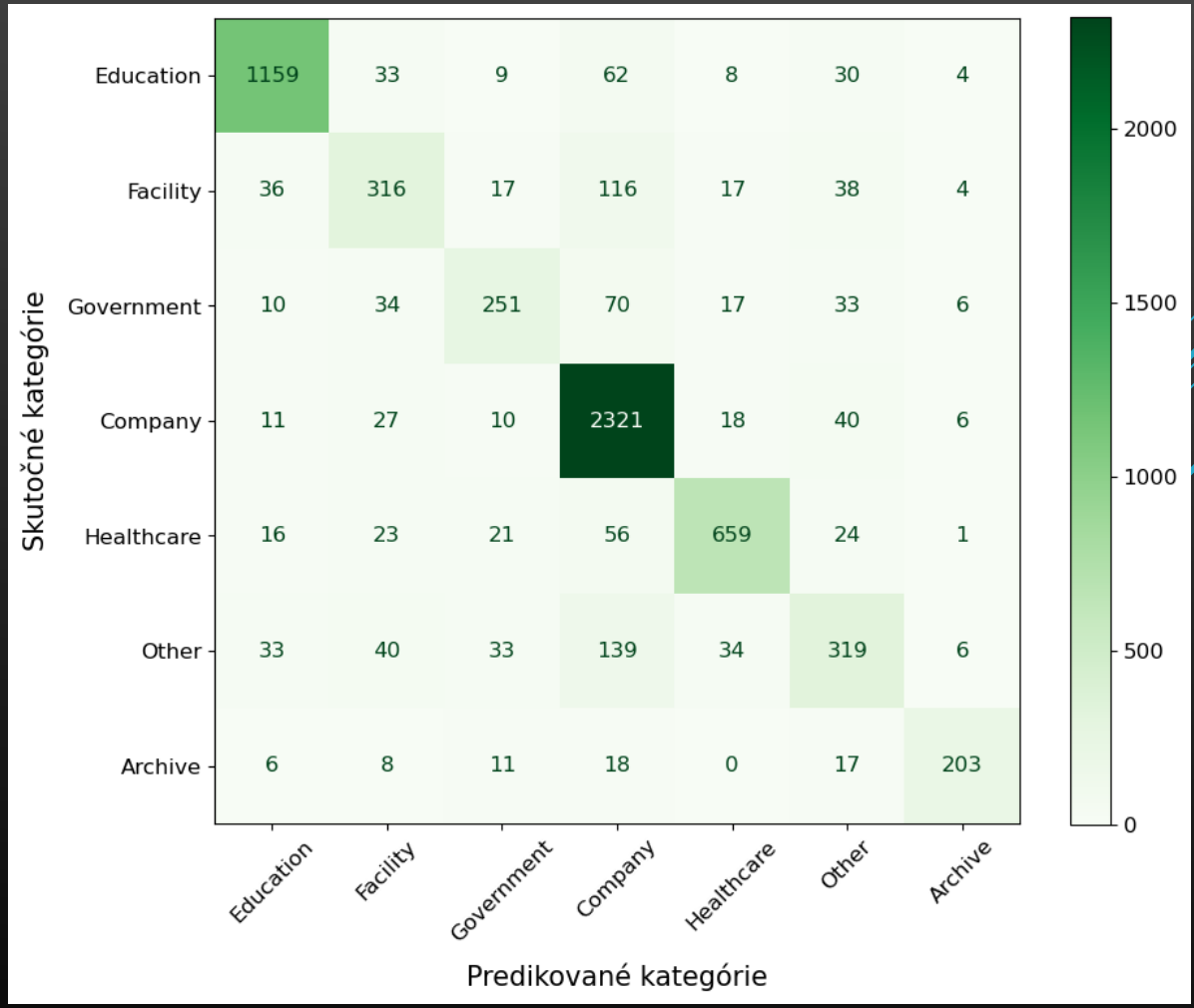
NAIVNÝ BAYES

Kategória	Správnosť	Presnosť	Citlivosť	F1-miera
Education	0.8935	0.8205	0.8935	0.8555
Facility	0.4504	0.6464	0.4504	0.5309
Government	0.4323	0.7615	0.4323	0.5515
Company	0.9581	0.7967	0.9581	0.8699
Healthcare	0.8175	0.8395	0.8175	0.8284
Other	0.4470	0.6506	0.4470	0.5299
Archive	0.6920	0.8626	0.6920	0.7679
Celkovo	0.7896	0.7807	0.7896	0.7753



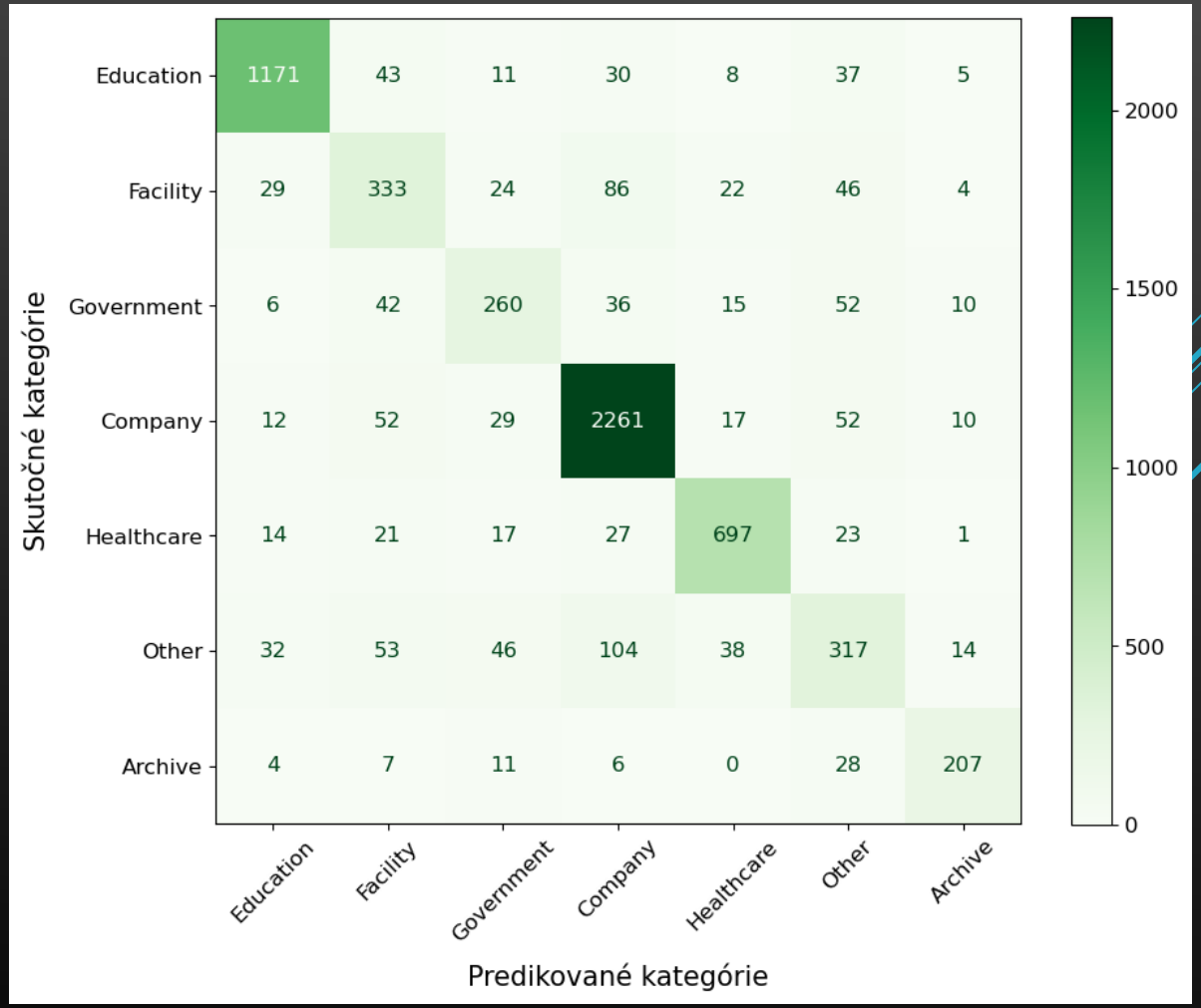
MODEL PODPORNÝCH VEKTOROV

Kategória	Správnosť	Presnosť	Citlivosť	F1-miera
Education	0.8881	0.9119	0.8881	0.8998
Facility	0.5809	0.6570	0.5809	0.6166
Government	0.5962	0.7131	0.5962	0.6494
Company	0.9540	0.8343	0.9540	0.8901
Healthcare	0.8238	0.8752	0.8238	0.8487
Other	0.5281	0.6367	0.5281	0.5774
Archive	0.7719	0.8826	0.7719	0.8235
Celkovo	0.8207	0.8154	0.8207	0.8152



KONVOLUČNÉ NEURÓNOVÉ SIETE

Kategória	Správnosť	Presnosť	Citlivosť	F1-miera
Education	0.8973	0.9235	0.8973	0.9102
Facility	0.6121	0.6044	0.6121	0.6082
Government	0.6176	0.6533	0.6176	0.6349
Company	0.9293	0.8867	0.9293	0.9075
Healthcare	0.8712	0.8745	0.8712	0.8729
Other	0.5248	0.5712	0.5248	0.5470
Archive	0.7871	0.8247	0.7871	0.8054
Celkovo	0.8235	0.8207	0.8235	0.8217



INTERPRETÁCIA VÝSLEDKOV

- Konvolučné neurónové siete vykazujú najlepšie výsledky
- Model podporných vektorov je mierne horší
- Naivný Bayes podľa očakávaní zaostáva

ZÁVER

- Podarilo sa nám splniť všetky ciele práce
- Výsledkom práce je model, ktorý má dobrú výkonnosť
- Ďalší výskum:
  - Preloženie web stránok namiesto ich odstraňovania
  - Experimentovanie s ďalšími technikami predspracovania údajov
  - Aplikovanie ďalších metód strojového učenia

ODPORÚČANÁ LITERATÚRA

- 1. Kazemian, H. B., & Ahmed, S. (2015). Comparisons of machine learning techniques for detecting malicious webpages. Expert Systems with Applications, 42(3), 1166-1177.
- 2. Chen, H., & Chau, M. (2003). Web Mining: Machine Learning for Web. Annual Review of Information Science and Technology 2004, 38, 289.
- 3. Raschka, S., & Mirjalili, V. (2019). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd.

ĎAKUJEM ZA POZORNOSŤ