

Klasifikácia webových stránok pomocou metód strojového učenia

Obhajoba bakalárskej práce
Martin Bača

Vedúci práce: doc. RNDr. Ľubomír Antoni, PhD.
Konzultant práce: RNDr. Šimon Horvát



Prírodovedecká fakulta
Univerzita Pavla Jozefa Šafárika v Košiciach

Košice, 21. jún 2023

Využitie klasifikácie webových stránok:

- Vyhľadávanie na Internete
- Analýza sentimentu
- Detekcia škodlivého obsahu

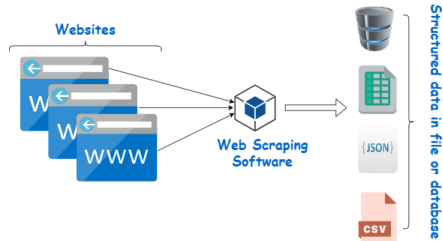


Požiadavka na konkrétnu aplikáciu:

- Kategorizácia zákazníkov spoločnosti GlobalLogic
- Analýza textu pomocou metód strojového učenia

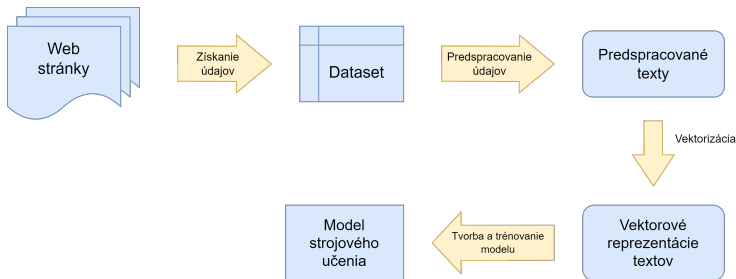
GlobalLogic®

- 1 Spracovať prehľad metód strojového učenia v oblasti **textovej analýzy**.
- 2 Navrhnuť 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

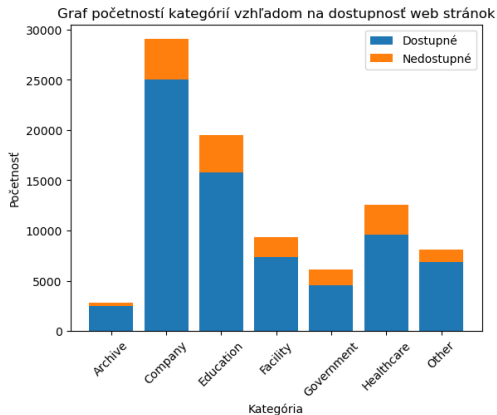
- 1 Vytvorenie datasetu scrapovaním údajov
- 2 Predspracovanie vstupov
- 3 Použitie metód na spracovanie prirodzeného jazyka (NLP)
- 4 Použitie metód strojového učenia
- 5 Porovnanie výsledkov pre použité metódy strojového učenia



Obr. 1: Diagram postupu riešenia (vlastný obrázok)

Zdroje webových stránok

- Register spoločností ROR Data 2022
- 70 000 dostupných inštitúcií v 7 kategóriách
- Rôzny počet inštitúcií v jednotlivých kategóriách



Tvorba 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\x99x9...
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

Agregované stĺpce:

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\tWelcome to The Alfred\n\t\t\t\t\t\n\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\t\n\t\t\t\t\tThe Alfred\n\t\t\t\t\tHave some feedback for us?

Use our to let us know. \n\t\t\t\t\t\n\t\t\t\t\t b' html \n\n Skip to main content

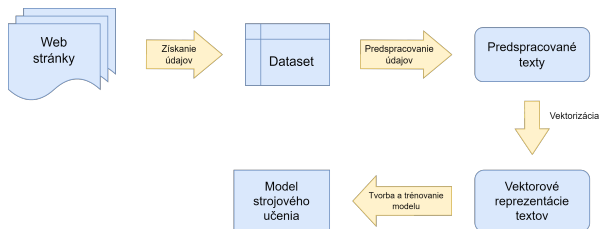
[illegible]

Charakteristika vstupov:

- Špeciálne znaky
- Číslice
- Rôzne jazyky

Transformácie datasetu

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



Obr. 2: Diagram postupu riešenia (vlastný obrázok)

- Úprava slov na ich základnú slovníkovú podobu
- Zjednotenie variantov slov
- Uľahčuje proces spracovania textu
- Odvodené slová majú rovnakú lemmu:
- Slová, ktoré nie sú odvodené majú rozdielne lemmy:

bežím => bežať

bežíš => bežať

bežia => bežať

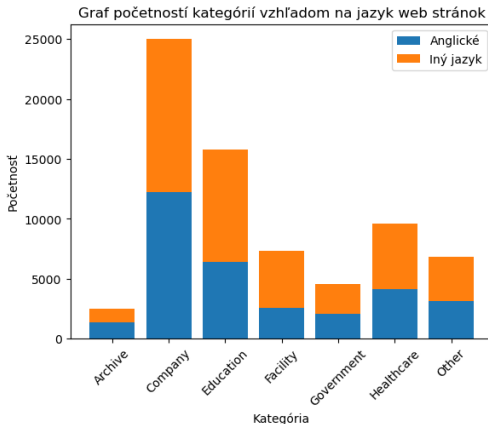
bežíš => bežať

chodia => chodiť



Pykes, K., 2023. Stemming and Lemmatization in Python

- **Výstup transformácie:** 40 000 anglických stránok zatriedených do 7 kategórií (najčastejší jazyk v datasete)



Vektorizácia textu

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

The quick brown fox jumped over the lazy brown dog.



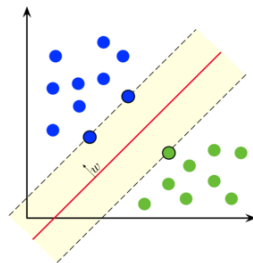
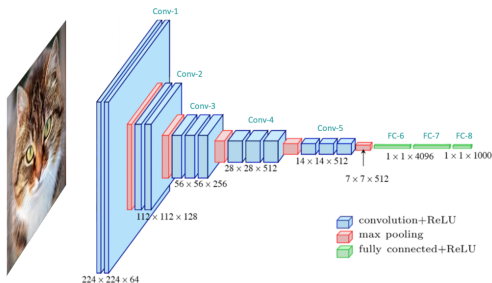
"The", "quick", "brown", "fox", "jumped", "over", "the", "lazy", "brown", "dog"



cat	the	quick	brown	fox	jumped	over	lazy	dog	house	bird
0	2	1	2	1	1	1	1	1	0	0

Použité metódy strojového učenia

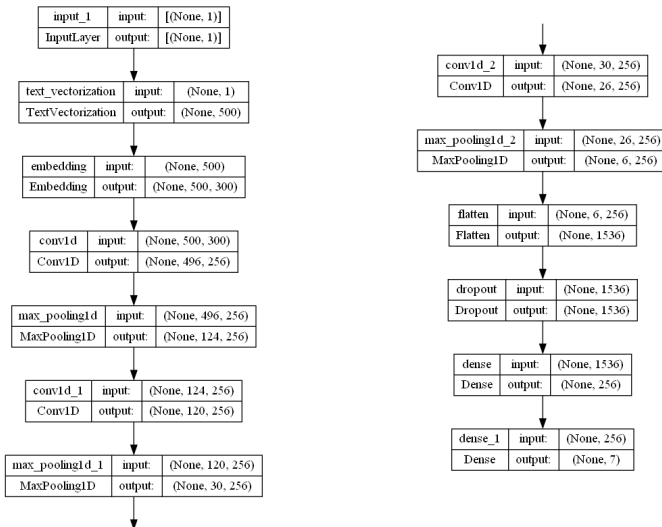
- Metóda naivného Bayesa
- Metóda podporných vektorov
- Konvolučné neurónové siete



Bayes' Theorem

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Architektúra konvolučnej neurónovej siete



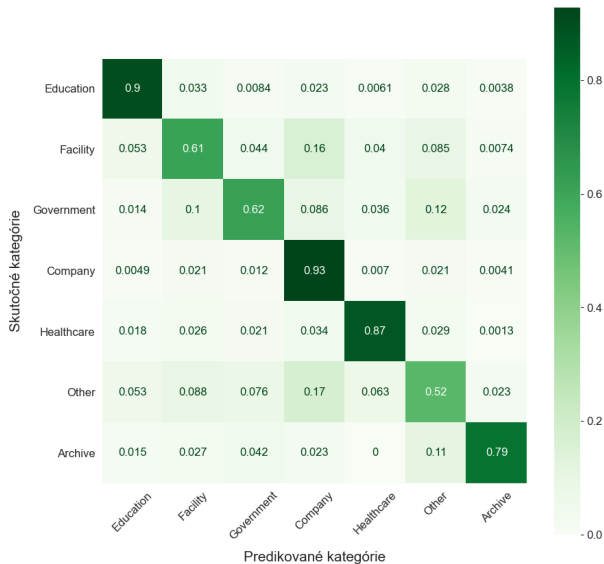
Výsledky modelu naivného Bayesa

Správnosť: 0.7896

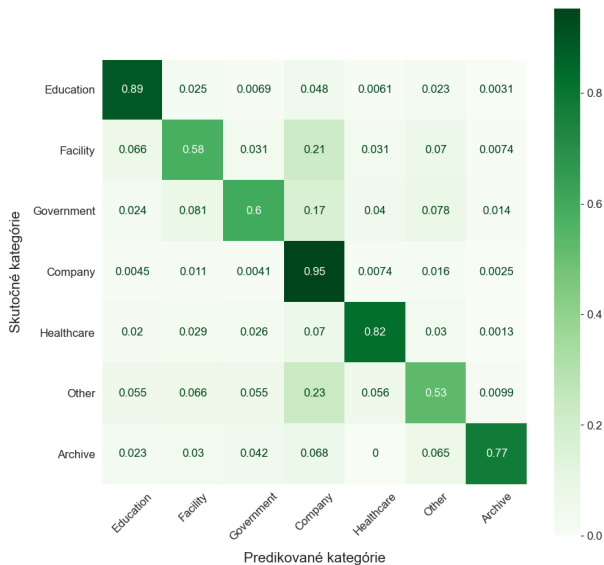
Presnosť: 0.7807

Citlivosť: 0.7896

F1-miera: 0.7753



Výsledky modelu podporných vektorov



Správnosť: 0.8207

Presnosť: 0.8154

Citlivosť: 0.8207

F1-miera: 0.8152

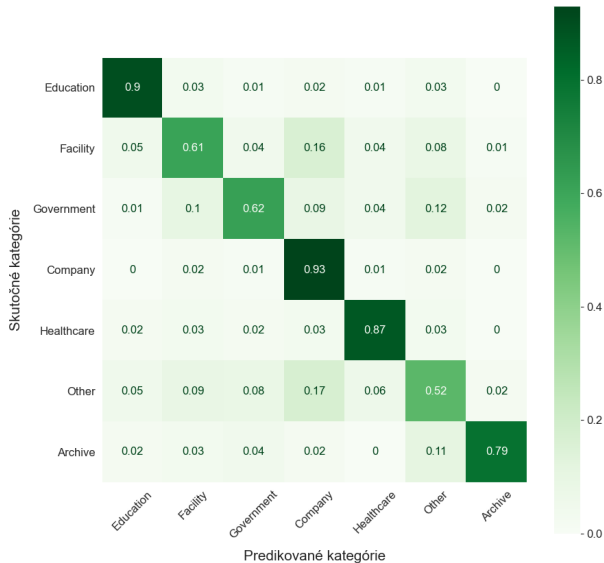
Výsledky modelu konvolučných nerónových sietí

Správnosť: 0.8235

Presnosť: 0.8207

Citlivosť: 0.8235

F1-miera: 0.8217



Porovnanie výsledkov s inými prácami

Článok	F1-miera	Špecifiká
Náš model	82 %	Konvolučné neurónové siete, 7 kategórií, lemmatizácia
Buber & Diri	85 %	Rekurentné neurónové siete, transfer learning
Razali a kol.	97 %	Rekurentné neurónové siete, 5 kategórií, stemming
Feng a kol.	99 %	LSTM a CNN, binárna klasifikácia

Tabuľka 1: Porovnanie výsledkov s inými článkami

Aplikácie metód strojového učenia majú veľký potenciál odhaliť skryté súvislosti v dátach a priniesť pokrok v rôznych oblastiach.

Ďalšia práca:

- Preloženie webstránok namiesto ich odstránenia z datasetu
- Experimentovanie s ďalšími technikami predspracovania údajov
- Aplikovanie ďalších metód strojového učenia

- <https://api.time.com/wp-content/uploads/2022/05/Google-Search-Personal-Information.jpg>
- <https://www.webharvy.com/images/web%20scraping.png>
- <https://learnopencv.com/wp-content/uploads/2023/01/tensorflow-keras-cnn-vgg-architecture-1024x611.png>
- <https://cdn1.byjus.com/wp-content/uploads/2020/10/Bayes-Theorem.png>

Ďakujem za pozornosť

Uvažovali ste aj o kategorizácii slovenských webových stránok?

Uvažovali ste aj použitie rekurentných NS ako Razali et al.?

- Časová náročnosť tréningu
- Použitie iných architektúr

Popíšte možné úlohy, ktoré by bolo možné riešiť v tejto oblasti v budúcnosti.

Aké ďalšie aplikácie strojového učenia v oblasti textovej analýzy majú v súčasnosti veľký potenciál?

- Analýza sentimentu, chatboty, automatická sumarizácia textu