KLASIFIKÁCIA WEB STRÁNOK POMOCOU METÓD STROJOVÉHO UČENIA

Vedúci práce: RNDr. Ľubomír Antoni, PhD.

Konzultant: RNDr. Stanislav Hrivňak, PhD.

Študent: Martin Bača

MOTIVÁCIA

- Vytvoriť model, ktorý dokáže webovú stránku zaradiť do jednej z vopred definovaných kategórií na základe jej obsahu

CIELE

- Spracovať prehľad metód strojového učenia na klasifikáciu webových stránok.
- Navrhnúť a extrahovať vhodné atribúty na klasifikáciu webových stránok s využitím scrapovania stránok.
- Implementovať metódy strojového učenia na klasifikáciu webových stránok podľa definovaných kategórií.
- Porovnať dosiahnuté výsledky s inými dostupnými štúdiami.

POSTUP RIEŠENIA

- Vytvorenie datasetu scrapovaním údajov
- Predspracovanie vstupov
- Použitie metód NLP
- Použitie metód strojového učenia
- Porovnanie výstupov pre použité metódy strojového učenia

AKTUÁLNY STAV

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

FILTROVANIE A PREDSPRACOVANIE VSTUPOV

- Filtrovanie webových stránok, ktoré nie sú v angličtine
- Odstránenie špeciálnych znakov
- Odstránenie krátkych slov (< 3 znaky)
- Odstránenie bežných slov, ktoré nám nedávajú žiadnu informáciu
- Stemming slov

STEMMING

- Hľadanie slovotvorného základu slova
- Spôsob vyjadrenia príbuznosti slov

- Podobné slová majú rovnaký stem:
  - likes => like
  - liked => like
  - likely => like

- Slová, ktoré nemajú veľa spoločného majú rozdielny stem:
  - likes => like
  - hates => hate

POUŽITIE METÓD NLP

- Bag of Words
- N-gram model
- Word2Vec

BAG OF WORDS

Reťazec: `John likes to watch movies. Mary likes movies too.`

Reťazec rozdelíme na tokeny (jednotlivé slová):

```
"John","likes","to","watch","movies","Mary","likes","movies","too"
```

Reprezentácia:

```
{"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
```

```
[1, 2, 1, 1, 2, 1, 1]
```

N-GRAM MODEL

Reťazec: `John likes to watch movies. Mary likes movies too.`

Reťazec rozdelíme na tokeny (jednotlivé slová):

```
"John","likes","to","watch","movies","Mary","likes","movies","too"
```

Reprezentácia:

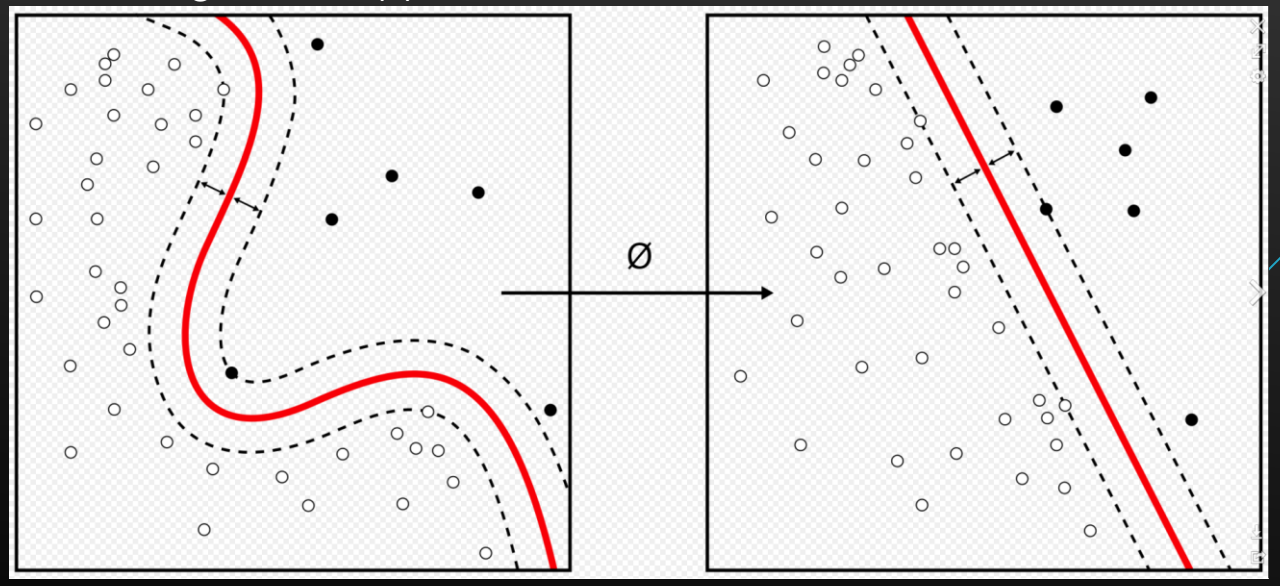
2-GRAM:

```
"John likes",
"likes to",
"to watch",
"watch movies",
"Mary likes",
"likes movies",
"movies too",
```

WORD2VEC

- Každý reťazec reprezentujeme jedinečným vektorom rovnakej dĺžky
- Vektory podobných reťazcov budú v priestorovej reprezentácii blízko pri sebe
- Použitím vhodného algoritmu vieme nájsť hranicu

Použitie algoritmu Support Vector Machine



https://en.wikipedia.org/wiki/Word2vec#/media/File:Kernel_Machine.svg

NAJBLIŽŠIE KROKY

- Preštudovanie odporúčanej literatúry
- Vytvorenie modelov s použitím rôznych metód strojového učenia
- Vyhodnotenie a porovnanie výsledkov

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