Додаток 1
Форма відомостей про авторів матеріалу та описова інформація для видань ТНТУ

Авторська довідка

(кваліфікаційної роботи магістра)

Назва кваліфікаційної роботи магістра Кластеризація спам-доменів методами машинного навчання
назви записувати нижнім регістром (як у реченні)

Назва (англ.): Clustering of spam domains using machine learning methods
переклад англійською

Освітній ступінь : магістр

Шифр та назва спеціальності: 125 «Кібербезпека»
напр.:151 Автоматизація та комп'ютерно-інтегровані технології

Екзаменаційна комісія: Екзаменаційна комісія № 47

Установа захисту: Тернопільський національний технічний університет імені Івана Пулюя
напр.: Тернопільський національний технічний університет імені Івана Пулюя

Дата захисту: 21 грудня 2022 року **Місто:** Тернопіль

Сторінки:
Кількість сторінок роботи: 71

УДК: 004

Автор роботи
Прізвище, ім'я, по батькові (укр.): Грицюк Владислав Петрович

Прізвище, ім'я (англ.): Hrytsyuk Vladyslav

Місце навчання (установа, факультет, місто, країна): ТНТУ ім. І. Пулюя, Факультет комп'ютерно-інформаційних систем і програмної інженерії, Кафедра кібербезпеки, м.Тернопіль, Україна

Керівник
Прізвище, ім'я, по батькові (укр.): Стадник Марія Андріївна
повністю

Прізвище, ім'я (англ.): Mariia Stadnyk
використовувати паспортну транслітерацію (КМУ 2010)

Місце праці (установа, підрозділ, місто, країна): ТНТУ ім. І. Пулюя, Україна

Вчене звання, науковий ступінь, посада:

Рецензент
Прізвище, ім'я, по батькові (укр.): Никитюк В'ячеслав В'ячеславович

Прізвище, ім'я (англ.): Nykytiuk Viacheslav
використовувати паспортну транслітерацію (КМУ 2010)

Місце праці (установа, підрозділ, місто, країна): ТНТУ ім. І. Пулюя, Факультет комп'ютерно-інформаційних систем і програмної інженерії, Кафедра комп'ютерних наук, м.Тернопіль, Україна
Вчене звання, науковий ступінь, посада: доцент кафедри КН

Ключові слова

спам, спам-домен, кластеризація, електронна пошта

spam, spam domain, clusterization, email

Анотація

українською:

В кваліфікаційній роботі вирішується проблема кластеризації спам доменів з використанням k-means, LSH, групування з метою подальшого застосування при процесі фільтрації різноманітних листів електронної пошти.. В роботі наведено основні методи фільтрації від спаму, а також основні методології їх виникнення. Детально розглянуто основні методи кластеризації, такі як: k-means, групування, ієрархічні методи, дерева, LSH, DBSCAN. Наведено методи оцінки кластеризації.

Здійснено кластеризацію спам доменів на основі реального сформованого набору даних з використанням інформації з сайтів Alexa та stopforumspams.com. Здійснено оцінку результату кластеризації з використанням додатково штучно введених функцій при маркуванні набору даних.

англійською:

The qualification work solves the problem of clustering spam domains using k-means, LSH, grouping with the purpose of further application in the process of filtering various e-mails. The work provides the main methods of spam filtering, as well as the main methodologies of their occurrence. The main methods of clustering, such as: k-means, grouping, hierarchical methods, trees, LSH, DBSCAN, are considered in detail. Methods of clustering assessment are presented.

Clustering of spam domains was carried out on the basis of a real generated data set using information from the Alexa and stopforumspams.com sites. The result of clustering was evaluated using additionally artificially introduced functions when labeling the data set.

Перелік літератури:

1. Brad Templeton. “Reaction to the DEC Spam of 1978”. In: availabe at: http://www. templetons. com/brad/spamreact. html(accesed December 2010) (2003).
2. Ben Nahorney. Email Threats 2017. An ISTR Special Report. Oct. 2017. url: https://docs.broadcom.com/doc/istr-email-threats-2017-en
3. Scott S. Smith. Internet Crime Report. 2020. url: https://pdf.ic3. gov/2016_IC3Report.pdf
4. Moorthy J. (2022) 23 Email Spam Statistics to Know in 2022, Retrived from https://www.mailmodo.com/guides/email-spam-statistics/
5. Anders Wiehes. “Comparing anti spam methods”. MA thesis. 2005.
6. Meng Wong and Wayne Schlitt. Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1. Tech. rep. 2006.
7. Dave Crocker, Tony Hansen, and Murray Kucherawy. DomainKeys Identified Mail (DKIM) Signatures. Tech. rep. 2011.
8. John Levine. DNS blacklists and whitelists. Tech. rep. 2010. url: https: //tools.ietf.org/html/rfc5782.
9. Wikipedia contributors. Anti-spam techniques. (Accessed on 12/04/2018). 2018. url: https://en.wikipedia.org/wiki/Anti-spam_techniques.
10. Vipul Ved Prakash. Vipul’s Razor. 2018. url: http://razor.sourceforge. net/ (visited on 06/06/2018).
11. Ernesto Damiani et al. “An Open Digest-based Technique for Spam Detection.” In: ISCA PDCS 2004 (2004), pp. 559–564.
12. Spammer-X Spammer-X. Inside the SPAM Cartel: By Spammer-X. Elsevier, 2004.
13. Wael H Gomaa and Aly A Fahmy. “A survey of text similarity approaches”. In: International Journal of Computer Applications 68.13 (2013).
14. Ion Androutsopoulos et al. “An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages”. In: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 2000, pp. 160–167.

15. Mick Johnson. DSPAM Project Homepage. 2011. url: http://dspam. sourceforge.net/ (visited on 03/11/2018).
16. J. Klensin. Simple Mail Transfer Protocol. RFC 5321. IETF, Oct. 2008, pp. 1–95. url: https://tools.ietf.org/html/rfc5321.
17. Tom. P. (2008). 'Latent botnet discovery via spam clustering'. The Expanded MIT Spam Conference 2008. Mar. 27-28, 2008. Boston, MA.
18. Calais, P. H., Pires, D. E. V., Guedes, D. O., Meira, W. Jr., Hoepers, C. and Steding-Jessen, K. (2020). 'A Campaign-based Characterization of Spamming Strategies'. The 5th Conference on Email and Anti-Spam. Aug. 21-22, 2020. Mountain View, CA.
19. Webb, S., Caverlee, J. and Pu, C. (2006). 'Introducing the Webb Spam Corpus: Using email spam to identify web spam automatically'. The 3rd Conference on Email and Anti-Spam. Jul. 27-28, 2006. Mountain View, CA.
20 Wei, C, Sprague, A., Warner, G and Skjellum, A. (2009). 'Characterization of spam advertised website hosting strategy'. The 6th Conference on Email and Anti-Spam. Jul. 16-17, 2009. Mountain View, CA.
21 Chio, C.; Freeman, D. Machine Learning and Security, O'Relly Media, 2018, 125-180.
22 Mick Johnson. DSPAM Project Homepage. 2022. url: http://dspam. sourceforge.net/.
23 Andrew Rosenberg and Julia Hirschberg. "V-measure: A conditional entropy-based external cluster evaluation measure". In: Proceedings of the 2019 joint conference on empirical methods in natural language processing and computational natural language learning. 2019.
24 W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Journal of the American Statistical Association 66 (1971): 846–850.