Introduction

* As malware industry grows, so does the means of infecting a computer or device.
* Internet is the entry point of infection
* URLs plays the main part in this because we connect to anything outside our network through URL in a browser.
* Identity theft and identity fraud are referred as two sides of cybercrime
* It helps to steal consumers' personal data such as user names, account numbers, passwords and other financial account credentials
* Malicious URL, a.k.a. malicious website, is a com- mon and serious threat to cybersecurity.

[1]A. Gabriel, D. Gavrilut and P. Stefan, "Detecting Malicious URLs: A Semi-Supervised Machine Learning System Approach - IEEE Xplore Document", *Ieeexplore.ieee.org*, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7829617/. [Accessed: 26- Sep- 2016].

[2]D. Sahoo, C. Liu and S. Hoi, "Malicious URL Detection using Machine Learning: A Survey", *Arxiv.org*, 2017. [Online]. Available: https://arxiv.org/abs/1701.07179. [Accessed: 13- Mar- 2017].

MA, J., Saul, L. K., Savage, S., and Voelker, G. M. 2011. Learning to Detect Malicious URLs. ACM Trans. Intell. Syst. Technol. 2, 3, Article 30 (April 2011), 24 pages. DOI = 10.1145/1961189.1961202 http://doi.acm.org/10.1145/1961189.1961202

Lexical features

Domain token count

Path token count

Longest path token length

Average domain token length

Longest domain token length

Average path token length

Length of url:

Length of primary domain:

Number of dots

Contain ip

Average word length

Longest word length

Whole url will be divided into three components

Hostname, path, parameters

**Length**- hostname, first-directory, URL, path,parameters, tld, second level domain, longest token in host, path, parameters and whole URL

**Counting features** – occurrence of sequence or character, - , @, ? , . , = , %, digits, numbers, letters, tokens, non-alphanumeric, directories, number of valid dictionary words (

**Pattern features-**

**Ratio features** - ratio between different types of characters or tokens:

Vowel/consonant, digit/letter ratio, average length of tokens for each subsection and the entire URL (12)

**Binary features -** DamerauLevenshtein algorithm to check the edit distance

Eg: www.millionairelottery.com/www.google.com

Although this information has the potential to improve classification accuracy, we exclude it for a variety of reasons. First, avoiding downloading page content is strictly safer for users. Second, classifying a URL with a trained model is a lightweight operation compared to first downloading the page and then using its contents for classification. Third, focusing on URL features makes the classifier applicable to any context in which URLs are found (Web pages, email, chat, calendars, games, etc.), rather than dependent on a particular application setting. Finally, reliably ob- taining the malicious version of a page for both training and testing can become a difficult practical issue.

LENGTH

1. URL 🡪 done
2. Length of hostname 🡪 done
3. First-directory 🡪 done
4. Parameters 🡪 done
5. Path 🡪 done
6. Top level domain 🡪 done
7. Second level domain 🡪 done
8. Longest token in the host 🡪 done
9. Longest token in path 🡪 done
10. Longest token in parameter 🡪 done
11. Longest token in URL 🡪 done

Count

1. – in url 🡪 done
2. – in hostname 🡪 done
3. \_ in url 🡪 done
4. \_ in hostname 🡪 done
5. @ in url 🡪 done
6. ? in url 🡪 done
7. . in url 🡪 done
8. = in url 🡪 done
9. % in url 🡪 done
10. digits in url 🡪 done
11. digits in hostname 🡪 done
12. letters in url 🡪 done
13. small letters in url 🡪 done
14. capital letters in url 🡪 done
15. letters in hostname 🡪 done
16. small letters in hostname 🡪 done
17. capital letters in hostname 🡪 done
18. tokens in url 🡪 done
19. tokens in hostname 🡪 done
20. tokens in path 🡪 done
21. tokens in parameter 🡪 done
22. non alphanumeric characters 🡪 done
23. directories 🡪 done
24. parameters 🡪 done

ratio features

1. vowelByConsonant ratio url 🡪 done
2. vowelByConsonant ratio hostname 🡪 done
3. vowelByConsonant ratio path 🡪 done
4. vowelByConsonant ratio parameter 🡪 done
5. digitByLetter ratio url 🡪 done
6. digitByLetter ratio hostname 🡪 done
7. digitByLetter ratio path 🡪 done
8. digitByLetter ratio parameter 🡪 done
9. average length of tokens URL 🡪 done
10. average length of tokens hostname 🡪 done
11. average length of tokens path 🡪 done
12. average length of tokens parameter 🡪 done

Binary features

1. .com in path 🡪
2. google in path
3. facebook in path
4. ebay in path
5. amazon in path
6. youtube in path
7. baidu in path
8. yahoo in path
9. twitter in path
10. Instagram in path
11. Microsoft in path

misc

1. wordgood?

errors

https-paypal.verifications-updates.com/

24026

26454

39214