Podobnosť doménových mien

autor: Lukáš Horňáček
školiteľ: doc. RNDr. Martin Stanek, PhD.

Phishing

- Phishing je forma útoku s využitím metód tzv. sociálneho inžinierstva, pri ktorom sa zločinec vydáva za dôveryhodnú osobu alebo inštitúciu s cieľom získať od obete citlivé informácie.

  [https://www.eset.com/sk/phishing/]

exarnple.com

exaample.com

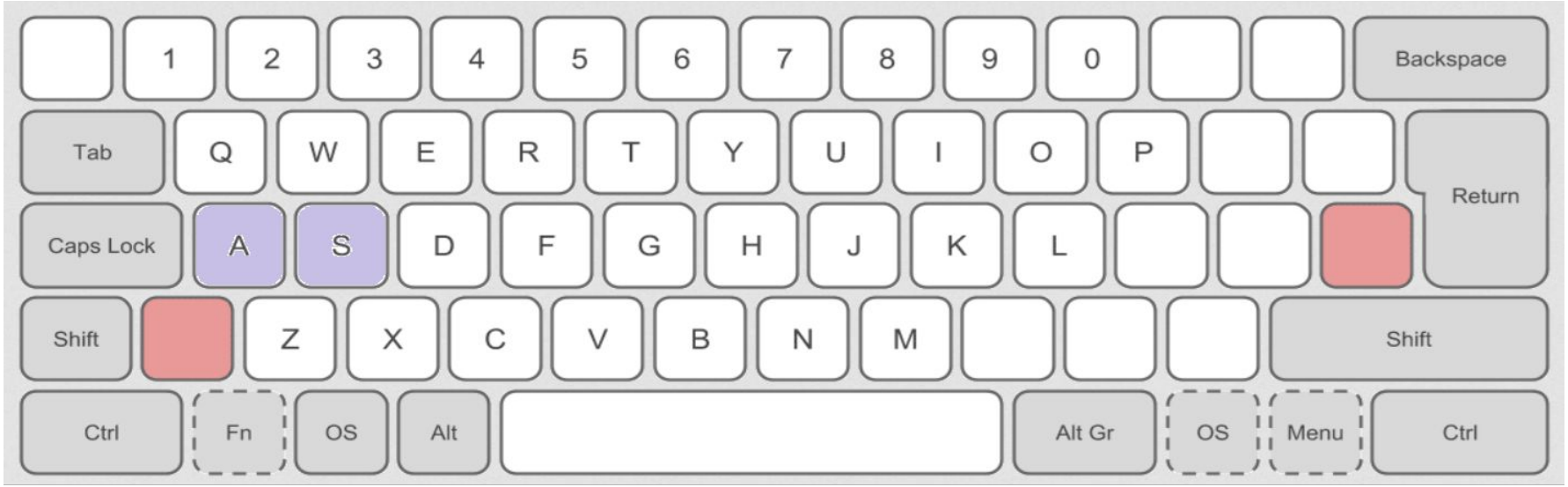
exmple.com

example-com.org

example.website.com

Typosquatting

example.com exsmple.com



Ciele práce

- detekcia phishingových a typosquattingových domén podľa ich podobnosti s legitímnou doménou
  - prehľad existujúcich funkcií na meranie podobnosti 2 domén
  - návrh 2 vlastných funkcií
  - porovnanie efektivity existujúcich a vlastných funkcií
- nástroj na generovanie podobných domén
  - návrh a implementácia
  - porovnanie s existujúcimi nástrojmi

Funkcie na meranie podobnosti

- Levenshtein distance
- Damerau-Levenshtein distance
- Jaro similarity
- Jaro-Winkler similarity
- Gestalt pattern matching
- …

Damerau-Levenshtein distance

- vzdialenosť medzi reťazcami *x* a *y* je najmenší počet vybraných operácií potrebných na transformáciu *x* na *y*
- operácie:
  - substitúcia jedného znaku za iný
  - výmena dvoch susedných znakov
  - pridanie znaku
  - odstránenie znaku

Damerau-Levenshtein distance

$$d(x_{1..i}, y_{1..j}) = \min \begin{cases} 0, & \text{if } i = j = 0 \\ d(x_{1..i-1}, y_{1..j}) + 1, & \text{if } i > 0 \\ d(x_{1..i}, y_{1..j-1}) + 1, & \text{if } j > 0 \\ d(x_{1..i-1}, y_{1..j-1}), & \text{if } i, j > 0 \wedge x_i = y_j \\ d(x_{1..i-1}, y_{1..j-1}) + 1, & \text{if } i, j > 0 \wedge x_i \neq y_j \\ d(x_{1..i-2}, y_{1..j-2}) + 1, & \text{if } i, j > 1 \wedge x_i = y_{j-1} \wedge x_{i-1} = y_j \end{cases}$$

Vanilla distance

- upravuje ceny operácií Damerau-Levenshtein distance, aby znížila vzdialenosti pre typosquattingové domény

Caramel distance

- upravuje Damerau-Levenshtein distance, aby detekovala phishingové domény, ktoré:
  - výrazne nemenia hlavnú časť originálnej domény
  - zmenia top-level doménu, prípadne pridajú prefix alebo suffix

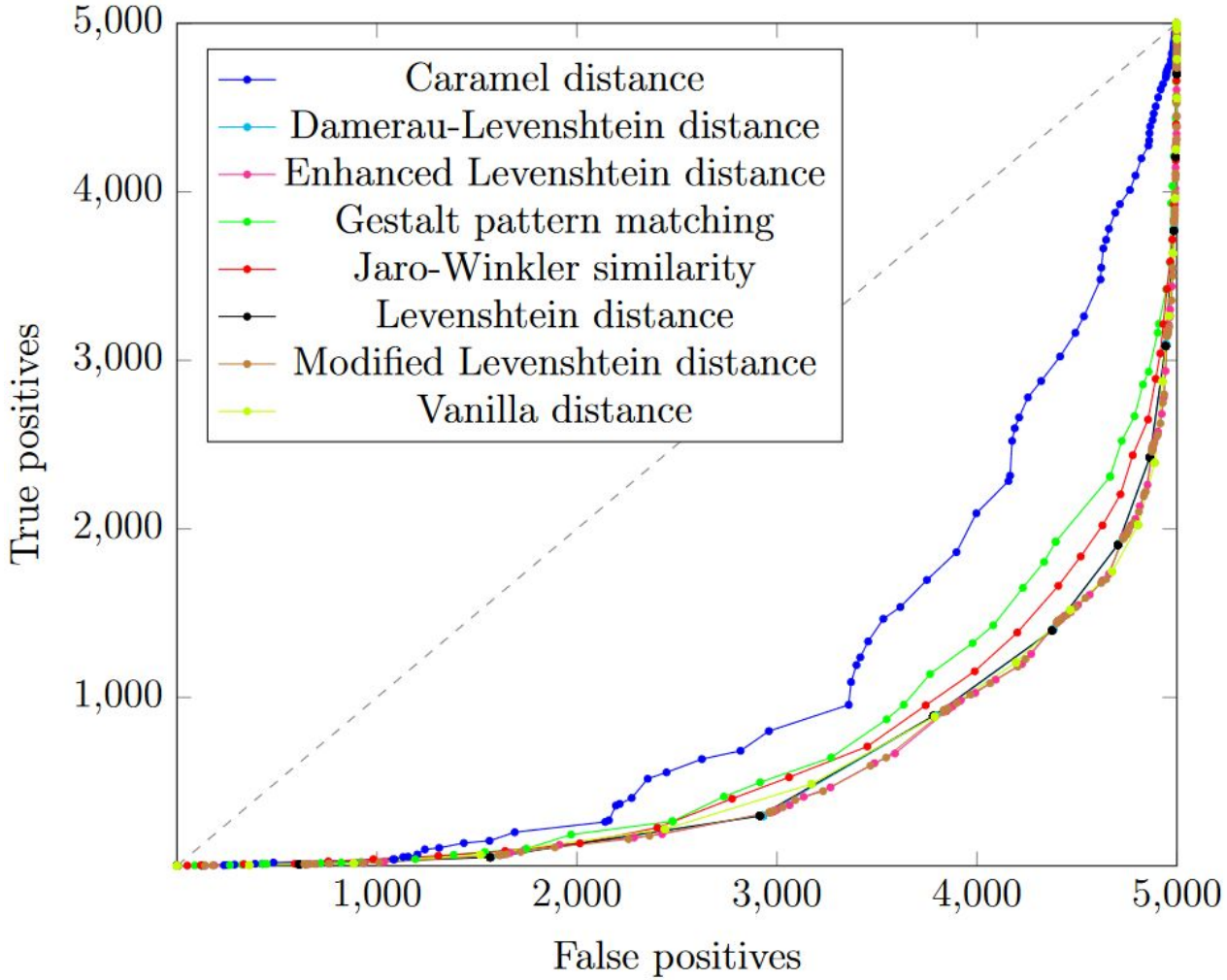
example.com example-site.org

Experiment

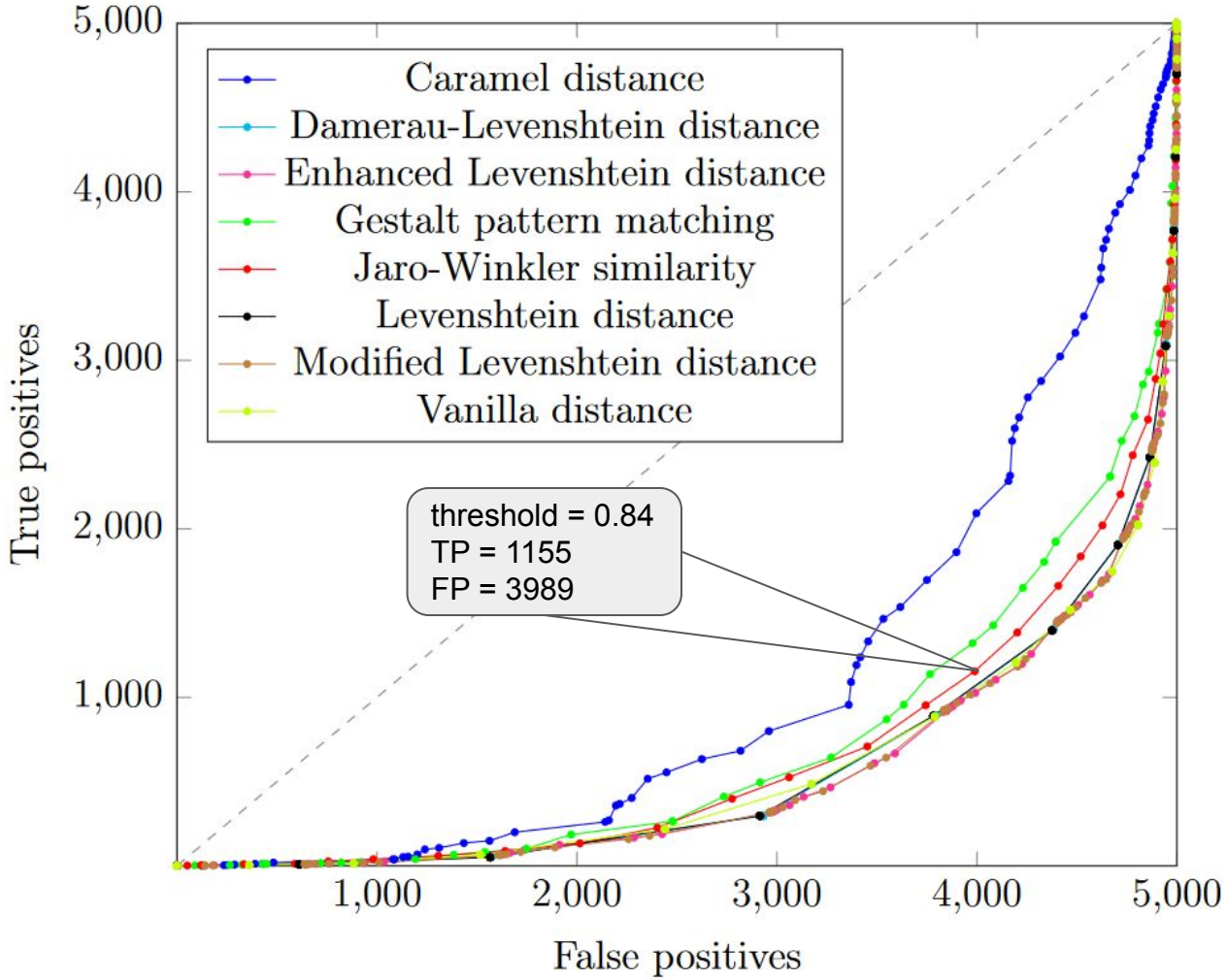
- získali sme dataset phishingových domén a dataset najpopulárnejších domén, ktoré sme označili ako legitímne
- dve domény sú podľa funkcie podobné práve vtedy, ak majú vzdialenosť menšiu ako nejaká zvolená hranica

$$TP = |\{x \in \text{phishing} \mid \exists y \in \text{legitimate}: d(x, y) < \text{threshold}\}|$$
$$FP = |\{x \in \text{legitimate} \mid \exists y \in \text{legitimate}: d(x, y) < \text{threshold} \wedge x \neq y\}|$$

Výsledky

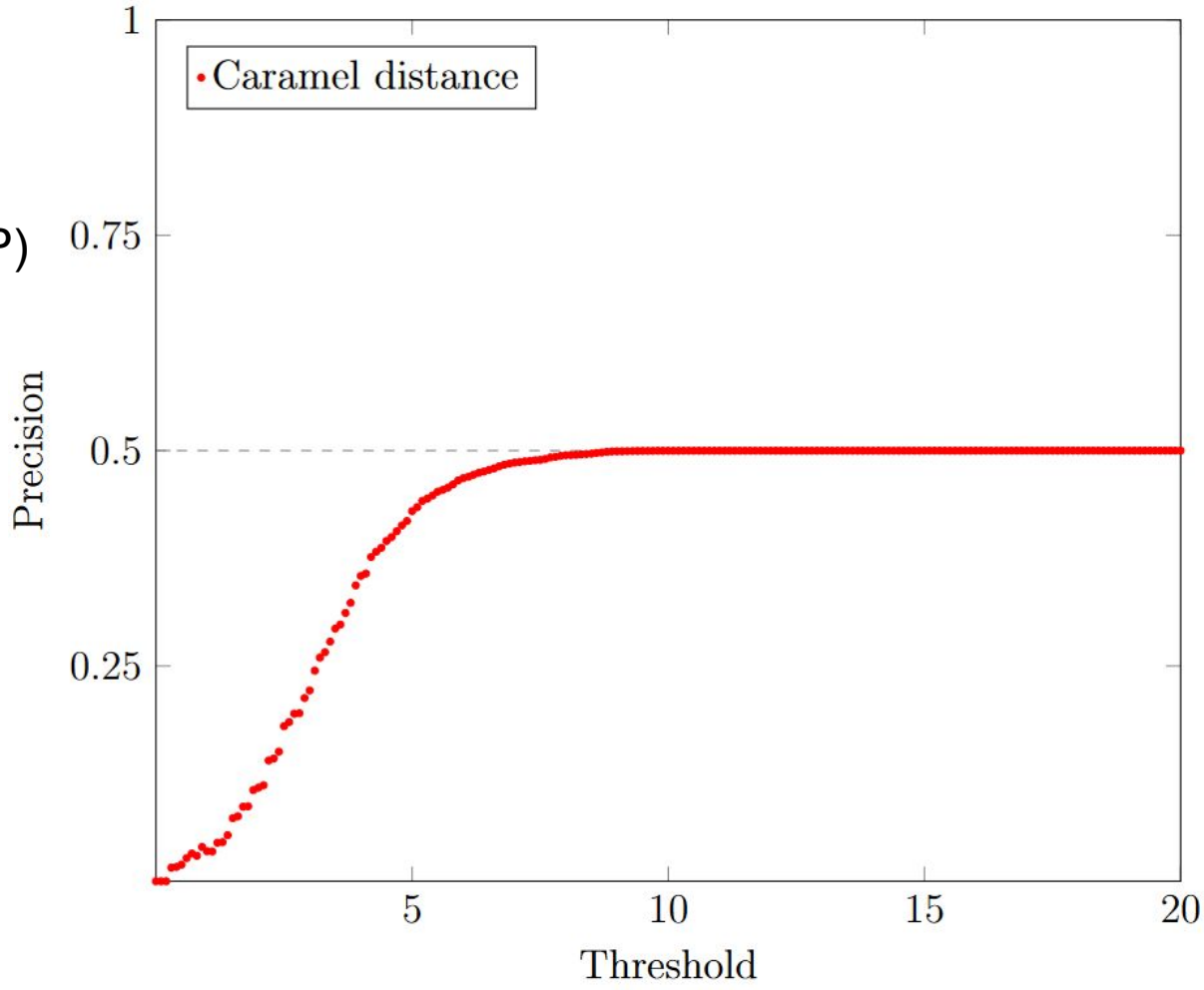

Caramel distance
Damerau-Levenshtein distance
Enhanced Levenshtein distance
Gestalt pattern matching
Jaro-Winkler similarity
Levenshtein distance
Modified Levenshtein distance
Vanilla distance
True positives
False positives
1,000
2,000
3,000
4,000
5,000

Výsledky


Caramel distance
Damerau-Levenshtein distance
Enhanced Levenshtein distance
Gestalt pattern matching
Jaro-Winkler similarity
Levenshtein distance
Modified Levenshtein distance
Vanilla distance
threshold = 0.84
TP = 1155
FP = 3989
True positives
False positives
1,000
2,000
3,000
4,000
5,000

Výsledky

Precision = TP / (TP + FP)



Experiment (2)

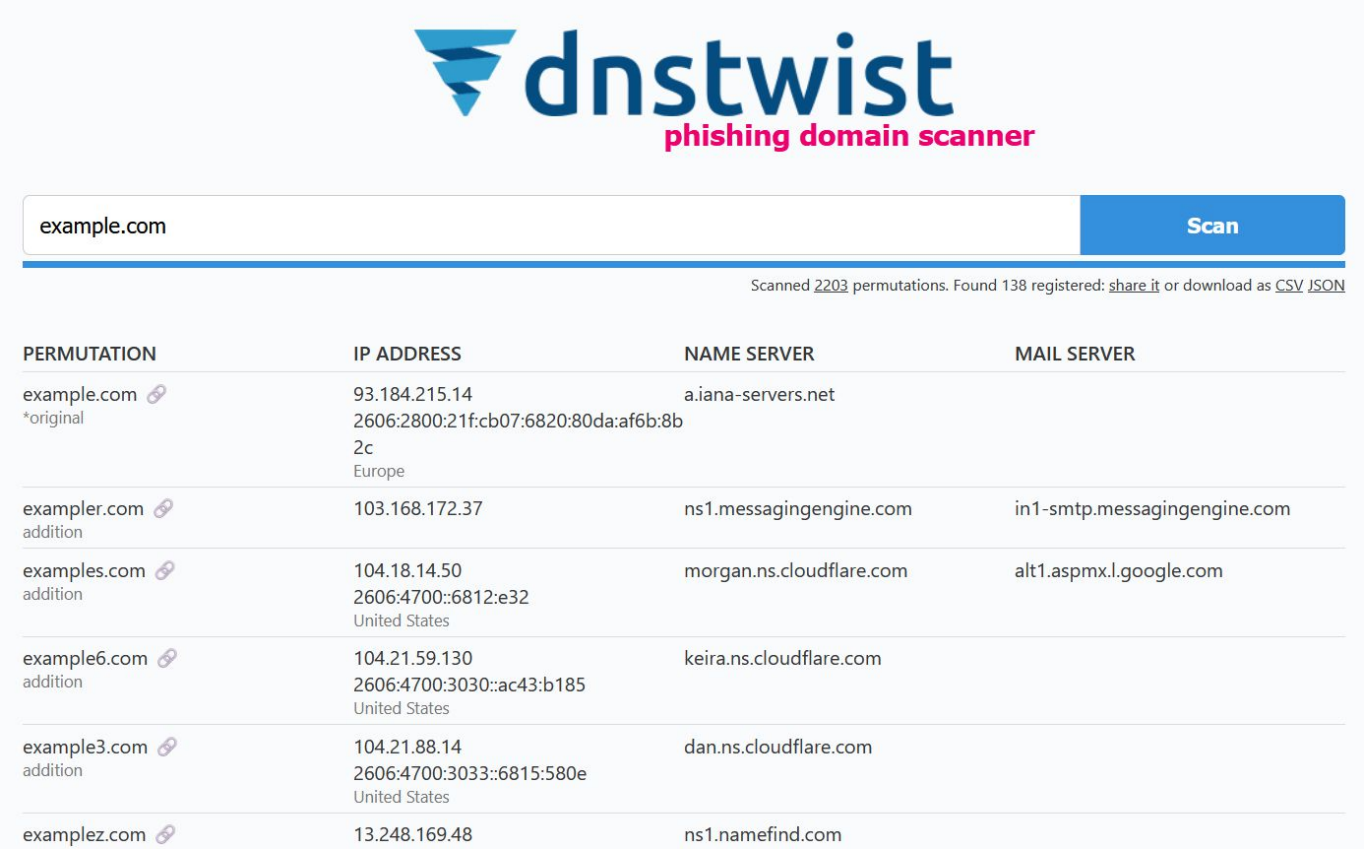
- vybrali sme 5 populárnych domén registrovaných pod doménou *sk*
- pre každú sme našli 20 najpodobnejších domén podľa rôznych funkcií
- manuálne sme overili obsah stránok na týchto doménach

Function	Malicious	Typosquatting	Legitimate	Neutral
Jaro-Winkler similarity	10	16	8	66
Gestalt pattern matching	11	14	7	68
Damerau-Levenshtein distance	14	15	9	62
Enhanced Levenshtein distance	13	14	10	63
Vanilla distance	13	15	6	66
Caramel distance	14	13	8	65

Generátor podobných domén

- namiesto analýzy zoznamov už navštívených domén môžu byť podobné domény generované preventívne
- vstup: doménové meno stránky spoločnosti
- výstup: zoznam podobných domén, ktoré môžu byť potenciálne použité útočníkmi snažiacimi sa imitovať danú spoločnosť

Existujúce nástroje



Ako zlepšiť existujúce nástroje

- viac vygenerovaných domén
  - viac spôsobov modifikácie originálnej domény
  - voľnejšie kombinovanie modifikácií
- kontrola nad veľkosťou výstupu
- domény zoradené podľa podobnosti

Implementácia

```
A → 'a' [0.00] | A 'a' [1.00] | A 'b' [1.00] | A 'c' [1.00] | A 'd' [1.00] ...
B → 'b' [0.00] | B 'a' [1.00] | B 'b' [1.00] | B 'c' [1.00] | B 'd' [1.00] ...
C → 'c' [0.00] | C 'a' [1.00] | C 'b' [1.00] | C 'c' [1.00] | C 'd' [1.00] ...
D → 'd' [0.00] | D 'a' [1.00] | D 'b' [1.00] | D 'c' [1.00] | D 'd' [1.00] ...
E → 'e' [0.00] | E 'a' [1.00] | E 'b' [1.00] | E 'c' [1.00] | E 'd' [1.00] ...
F → 'f' [0.00] | F 'a' [1.00] | F 'b' [1.00] | F 'c' [1.00] | F 'd' [1.00] ...
...
```

Porovnanie nástrojov

Tool	Generated domains	Time
dnstwist (default)	3256.6	0.45s
dnstwist (with lists)	20619.2	1.14s
Typosquatting Finder	10600	0.28s
URLCrazy	2030.6	0.35s
SDG Default (5000)	5000	7.98s
SDG General (5000)	5000	2.67s
SDG Default (10000)	10000	9.10s
SDG General (10000)	10000	4.45s
SDG Default (25000)	25000	23.23s
SDG General (25000)	25000	9.53s

Porovnanie nástrojov

Tool	Generated domains	Registered domains
dnstwist (default)	3256.6	17.8
dnstwist (with lists)	20619.2	31.2
Typosquatting Finder	10600.0	117.0
URLCrazy	2030.6	96.0
SDG Default (5000)	5000	26.0
SDG General (5000)	5000	38.0
SDG Default (10000)	10000	65.2
SDG General (10000)	10000	70.4
SDG Default (25000)	25000	271.0
SDG General (25000)	25000	115.6

Diskusia

Skupiny pravidiel

```
general:
    insertion
    deletion
    substitution
    transposition
typosquatting:
    repeat_insertion
    repeat_deletion
    adjacent_insertion
    adjacent_deletion
    adjacent_substitution
    adjacent_transposition
```

```
similar:
    similar_substitution
    similar_transposition
    similar_multicharacter_substitution
public_suffix:
    tld_insertion
    tld_deletion
    tld_substitution
    tld_transposition
extended_similar:
    user_prefix_suffix
    common_prefix_suffix
    custom_prefix_suffix
    level_transposition
    numeral_substitution
```