Algorytmy uczenia maszynowego a użycie nazw w domenach. Jak robot widzi domenę .pl?

Piotr Studziński-Raczyński
Starszy Specjalista ds. DNS

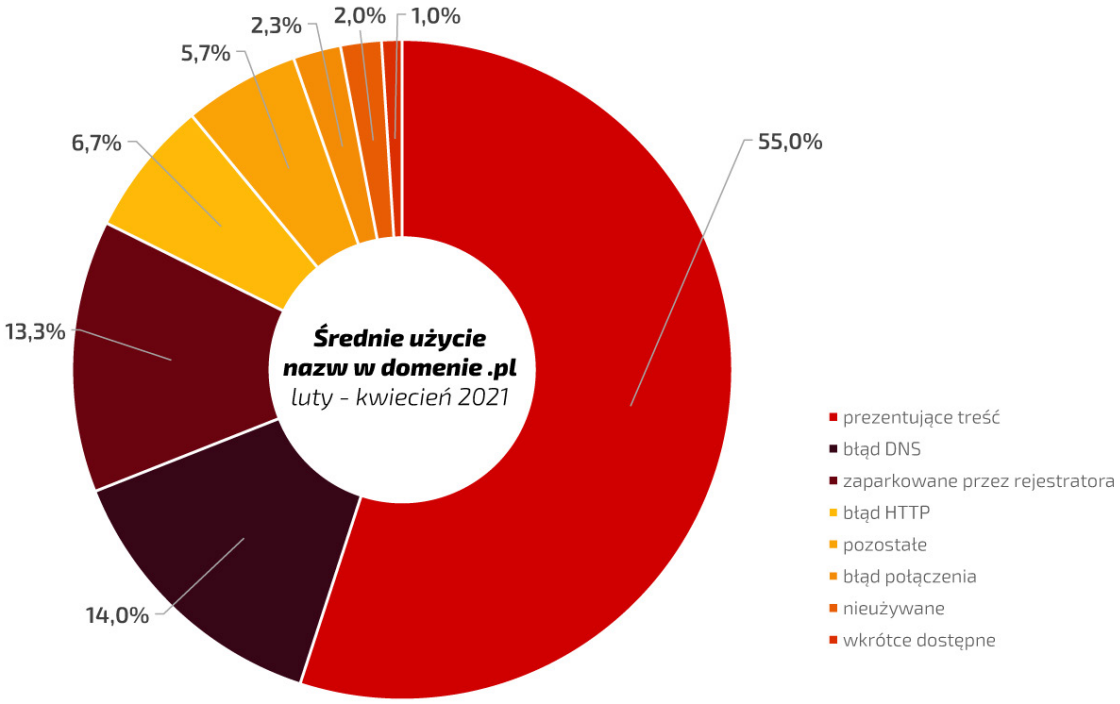
Aby zbadać i skategoryzować wykorzystanie domen internetowych, w ramach prac prowadzonych w CENTR opracowany został robot internetowy „crawler" *signs of life*. Jest to narzędzie indeksujące, pozwalające na comiesięczne uzyskiwanie danych na podstawie losowej próby domen (50k), udostępnianej przez uczestniczących w tej inicjatywie członków, wśród których znajduje się również Rejestr domeny .pl.

Wiedza o sposobie wykorzystania domen internetowych jest niezwykle cenna, pozwala lepiej zrozumieć, co dzieje się z zarejestrowanymi domenami, np. jaka jest skala występowania poszczególnych błędów skutkujących niepoprawnym dostarczeniem usługi lub dla jakiej części domen abonenci nie znaleźli jeszcze zastosowania. Otrzymany obraz może być pomocny w budowaniu strategii, wytyczaniu celów oraz ocenie ryzyka biznesowego. Odsetek domen „zaparkowanych", z błędami, przekierowanych lub przedstawiających strony internetowe o istotnej zawartości może dać pogląd na ogólne zainteresowania i potrzeby abonentów. Może też pomóc w badaniu bieżących trendów na rynku domenowym, jak również wesprzeć rejestry, rejestratorów oraz samych użytkowników domen internetowych w jak najefektywniejszym i płynnym wykorzystaniu zasobów sieci.

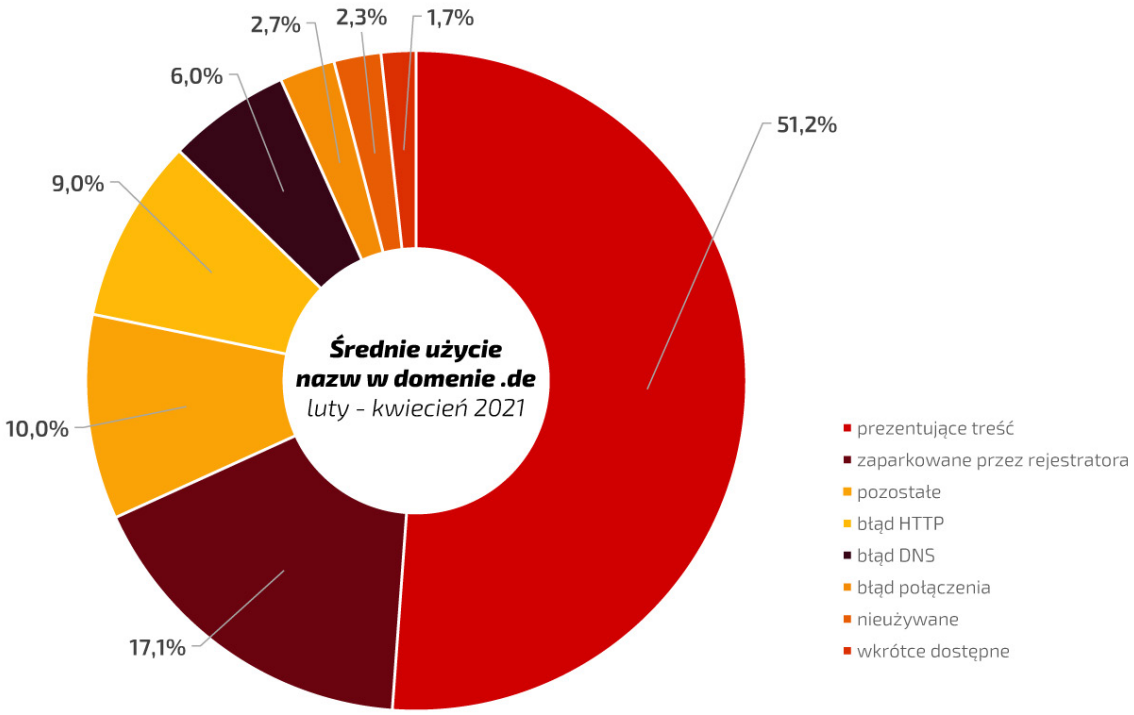
Platforma *signs of life*, oparta o algorytm uczenia maszynowego, skanuje i klasyfikuje dane wejściowe w zdefiniowanych obszarach technicznych. Dzięki temu można określić, jak duża część domen najwyższego poziomu jest używana do prezentowania treści w Internecie. Klasyfikacja podzielona została na kategorie pod kątem szczegółowości. Na pierwszym poziomie szczegółowości rozróżniamy, czy badana domena odnosi się do jakiejś zawartości. Domeny nie odnoszące się do żadnej zwartości to te, których działanie (ładowanie się strony internetowej) blokowane jest wskutek jakiegoś błędu, np. błędu DNS, błędu połączenia (błąd serwera, timeout) czy błędu http o wskazanym kodzie. Domeny odnoszące się do określonej zawartości podzielone z kolei zostały na te, które prezentują treści oraz te, które zawierają informację o ich kategorii (trzeci poziom szczegółowości), np. „zaparkowana" przez rejestratora, zablokowana, nieużywana, wkrótce dostępna, porzucona. Ostatni, czwarty poziom szczegółowości precyzuje kategorię domen określoną na poprzednim poziomie, co w niektórych przypadkach związane jest z dalszą i ostateczną subkategoryzacją. Przykładem może być podział domen nieużywanych na te odnoszące się do indeksu strony lub do pustej strony lub dalszy podział domen z kategorii wkrótce dostępna na prezentujące informację „strona w budowie" oraz przedstawiające początkową stronę kreatora stron internetowych.

Poza badaniem domen pod względem ich zawartości, crawler weryfikuje też w jakim stopniu są one przekierowywane na zewnętrzne adresacje i ile spośród analizowanych domen jest powiązanych z usługą poczty elektronicznej. Jeśli domena nie zwraca strony internetowej, nie musi to oznaczać, że nie jest używana, może być powiązana z serwerem pocztowym. Aby uzyskać te informacje każda domena jest odpytywana o rekordy MX (Mail Exchange) za pomocą biblioteki dnspython. Przykładowo w maju 2021 roku wskaźnik rekordu MX dla domeny .pl, podobnie jak dla domeny .fr, wyniósł 88% i był to drugi pod względem wysokości wynik po domenach .de i .si, gdzie usługa Mail Exchange występowała w 89% przebadanych nazw domen. Dla porównania, średnia wartość tego wskaźnika dla innych objętych badaniem domen ccTLD wynosiła 76%.

Poniżej, na wykresie nr 1, przedstawiono uśrednione dane dotyczące użycia nazw w domenie .pl w okresie od lutego do kwietnia 2021. Analogiczne dane za ten sam okres zostały przedstawione na wykresie nr 2 dla domeny .de. Na podstawie otrzymanych danych można zaobserwować, iż odsetek domen odnoszący się do zawartości, gdzie prezentowana jest treść, w przypadku obu analizowanych domen jest podobny i przekracza 50%, z kilkuprocentową przewagą Rejestru domeny .pl. Wyniki te mogą świadczyć o dobrej efektywności wykorzystania zarejestrowanych nazw domen przez ich abonentów. Innym kryterium, które może zwrócić uwagę obserwatora, jest występowanie błędu DNS, tj. niepoprawnego skonfigurowania delegacji domeny. Pod tym względem widać, że częściej poprawną delegację wskazaną mają nazwy zarejestrowane u naszych zachodnich sąsiadów.

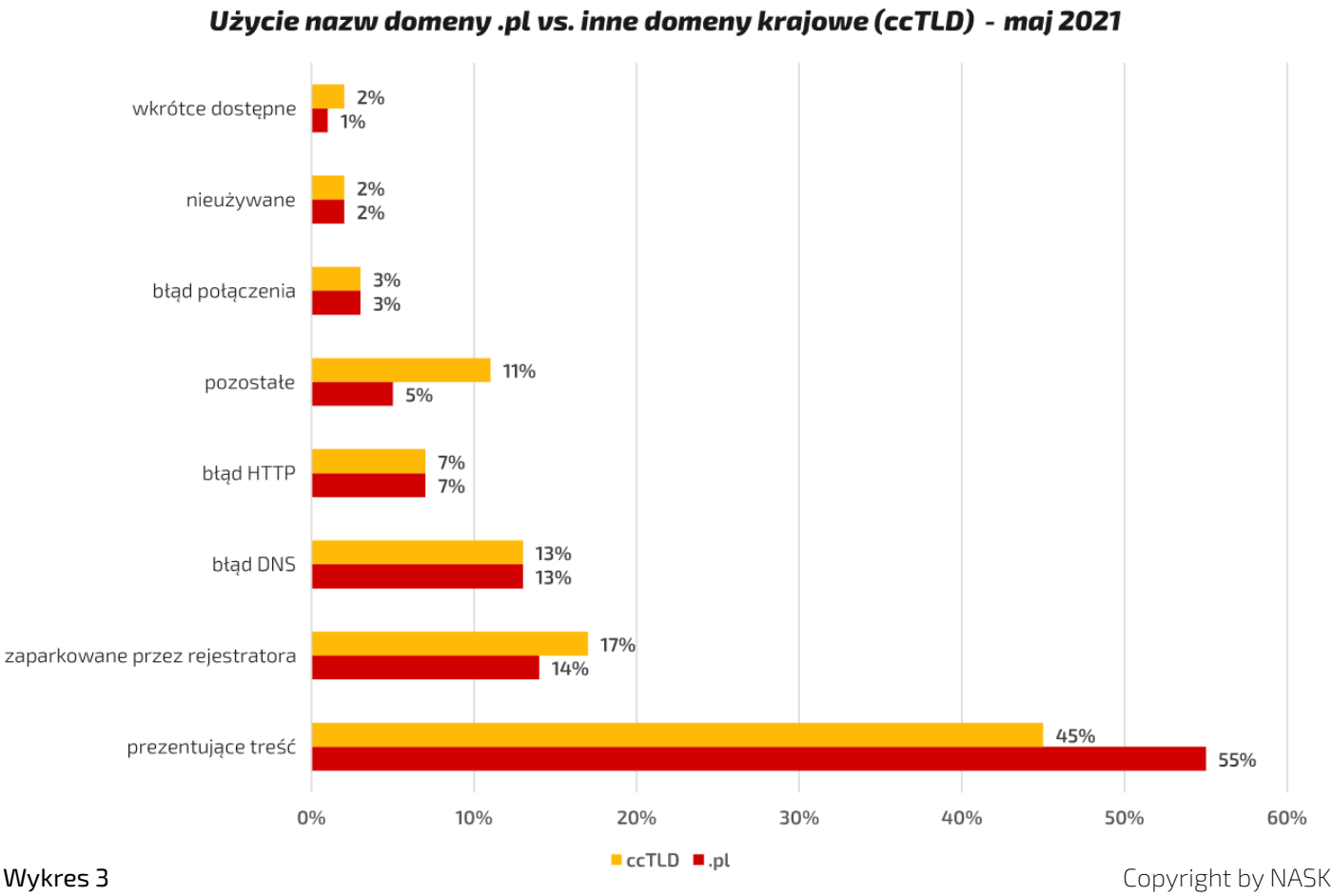


Wykres 1



Wykres 2

Patrząc z kolei bardziej globalnie, a więc zestawiając wyniki uzyskane w maju 2021 roku dla nazw w domenie .pl ze średnimi wartościami dla innych domen krajowych najwyższego poziomu biorących udział w badaniu (Wyk. nr 3), zaobserwować można, że odsetek efektywnego wykorzystania nazw w domenie .pl, wynoszący 55%, znacznie przekracza średnią dla objętych badaniem rejestrów ccTLD, gdzie wartość ta utrzymuje się na poziomie 45%. W tej kategorii przykładem wybitnie dobrego wyniku, gdzie wartość wskaźnika sięgnęła 60%, jest domena .si. Warto przy okazji zauważyć, że słoweńska domena krajowa (.si) odnotowuje niski wskaźnik domen „zaparkowanych" przez rejestratorów (3%) w porównaniu ze średnią dla innych domen krajowych, wynoszącą 17%. W maju 2021 roku dla domeny .pl wartość ta utrzymywała się na poziomie 14%. W przypadku pozostałych kategorii uwzględnionych w badaniu, wartości odnotowane przez domenę .pl są równe bądź zbliżone do średniej uzyskanej w zestawieniu przez inne domeny ccTLD.



Wykres 3

Wśród niektórych przedstawicieli rejestrów europejskich panuje pogląd, iż możliwości jakie oferuje crawler, stanowią pierwszy krok w kierunku bardziej szczegółowej charakterystyki zawartości, do której odnoszą się domeny internetowe. Podzielana jest również opinia, co do zalet analizy z wykorzystaniem uczenia maszynowego oraz wielostronnych korzyści płynących z uzyskanych w ten sposób informacji. Część rejestrów już prowadzi takie analizy lub planuje ich rozpoczęcie. Wskazując plusy związane z tego rodzaju mapowaniem strefy, np. te związane z wykorzystaniem otrzymanych danych do oceny siły i jakości strefy, przedstawiciele rejestrów ccTLD przy CENTR podkreślają jednocześnie istotę racjonalnej oceny danych, dzięki której możliwe jest wyciągnięcie wniosków pozwalających na jeszcze efektywniejsze i bezpieczniejsze wykorzystanie możliwości, jakie domeny oferują użytkownikom Internetu.