Domenowy crawler kluczy publicznych i certyfikatów.

Mikołaj Koszowski 274392
Hubert Nowakowski 274415

Zad 3. wariant 5a
Kod źródłowy: github.com/Fisher16/KiBD

Celem zadania było stworzenie crawlera, który dla otrzymanej listy domen pobierze ich **klucze publiczne**, a następnie sprawdzenie czy otrzymana lista kluczy zawiera **duplikaty**.
Jako **źródło** domen wykorzystaliśmy zrzut **CommonCrawl**'a z lutego 2020, za pomocą dostępnego API wyciągnęliśmy wszystkie domeny kończące się na ***.pl**. i otrzymaliśmy listę ponad **250k** linków, dostępna na repo, a pobrane klucze udostępniamy na współdzielonym OneDrive.
Wykorzystaliśmy framework **scrapy**, który automatyzuje asynchroniczne wysyłanie i przetwarzanie zapytań. Do przechowywania pobranych certyfikatów wykorzystaliśmy dokumentową bazę danych **mongodb**. Po skutecznym pobraniu certyfikatu zapisujemy go do bazy, wraz z dodatkowymi informacjami, takimi jak **klucz publiczny** i jego **hash**. Wydajność crawlera wynosiła około **1k stron/min** i jak na nasze potrzeby była ona zadowalająca.
Udało się nam pobrać **135k** kluczy publicznych z czego **122k / 90.3%** stanowiły klucze RSA.

Długość/bits	% kluczy	# kluczy w tys
1024	0.1	0.1
2048	73.6	89.8
3072	0.3	0.4
4096	26.0	31.7

Podczas pobierania nie zapisywaliśmy stron podpisujących się nie swoim certyfikatem, których było ponad **4k** (strony takie wyświetlają się jako „Not secure" i wymagają potwierdzenia użytkownika na korzystanie z nich).
Do znalezienia duplikatów wykorzystaliśmy hashe, pomimo że staraliśmy się odfiltrowywać linki z tej samej domeny to znaleźliśmy **17.6k duplikatów**. Po dokładniejszym filtrowaniu stron z jednakowej domeny otrzymaliśmy **731 stron wykorzystujących ten sam klucz publiczny**. Przykładowe grupy stron wykorzystujące ten sam klucz poniżej:

- anielaolsztynek.pl ipresso.pl ranczokamienczyk.pl xpictures.pl applay.pl biorekredyt.pl brb.fundacjazendriving.pl fullstak.pl gsi-goszczynski.pl infobot.pl katowice.meetjs.pl ligagraczy.pl powtorkazpolskiego.pl sesja.dastudnia.pl www.smart.biz.pl tanieelewacje.pl times.edu.pl
- devfest.wroclaw.pl fmpmsa.pl misja-kerygma.pl portuj.pl repsol-car.pl weedweek.pl **chmurakrajowa.pl** esportgamesclub.pl kartawfrp.pl kwzgoda.pl caninto.space osiedleczekanow.pl standardexpress.pl
- akcesoria-dachowe.pl allekwiaty.pl galt.pl hardgirl.pl mok.opole.pl pix.mtlumaczenia.pl pysznica.pl grono.gda.pl osir.szczecinek.pl www.radosc.edu.pl www.sinprogres.pl

Spora część z tych stron to strony informacyjne przekierowujące do domeny głównej. Jednakże, gdy grupy są większe to wspólnym mianownikiem staje się CA Let's Encrypt nonprofit zapewniający darmowe certyfikaty TLS.

Źródła:
https://index.commoncrawl.org/
https://letsencrypt.org/
https://docs.scrapy.org/en/latest/topics/broad-crawls.html