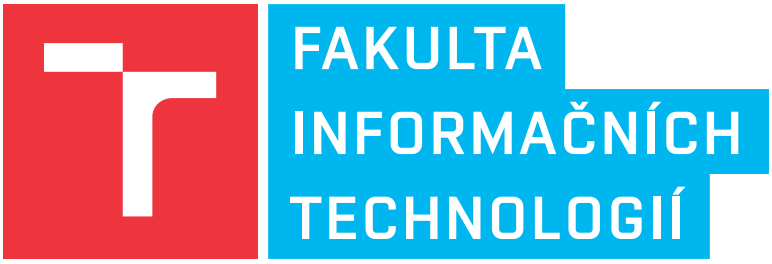

Projekt do předmětu PDS – Přenos dat, počítačové sítě a protokoly

Parser bitcoinových adres z HTML stránky

27. dubna 2018

Řešitel: Michal Moravec (xmorav28@stud.fit.vutbr.cz)
Fakulta Informačních Technologií
Vysoké Učení Technické v Brně

1 Zadání

- nalezení veřejně dostupných stránek obsahujících dostatečné množství kryptoadres;
- vytvoření PHP skriptu (za použití libovolných knihoven k ulehčení práce), který bude z těchto stránek (např., https://bitcoinwhoswho.com/blog/2016/02/29/bitcoin-donation-address-rankings/) schopen parsovat kryptoměnové adresy a ukládat je spolu s metadaty (identifikátor, zdroj původu, URL) do PostgreSQL databáze.

2 Použité technologie

- PHP 7
- Laravel 5.6
- PostgreSQL

3 Návrh řešení

Projekt jsem implementoval se zaměřením na rozšiřitelnost a integrovatelnost do stávajícího projektu v PHP frameworku Laravel. Vstupem je pole s URL adresami, které obsahují libovolné množství kryptoadres. Tyto adresy se mohou na dané stránce vyskytovat na libovolném místě. Výstupem je pak asociativní pole, které jako klíče obsahuje kryptoadresy, ke kterým je jako hodnota přiřazeno asociativní pole obsahující dodatečné informace o dané adrese:

- URL adresa, ze které je adresa získána
- Title stránky, vyparsovaný z hlavičky HTML tagu title
- Metadata získaná ze stránky, na které se daná kryptoadresa nachází
- Metadata získaná ze stránky blockchain.info
- Typ dané adresy (např. zdali se jedná o BTC, LTC, aj.)

Skript nejprve stáhne celý obsah HTML stránky, kde je nejprve třeba vyparsovat title přes jednoduchý XPath. Dále je oříznut obsah stránky tak, aby obsahoval jen vnitřek tagu body. Toto je ochrana proti potenciálnímu zachycení nesmyslných kryptoadres, které by mohly být získány z hlavičky HTML stránky, která obsahuje např. bezpečnostní tokeny, které by mohly projít regexem.

Jednotlivé kryptoměny jsou zadefinovány regulárními výrazy (regexy), je tedy možné skript doplnit o libovolné množství nových typů kryptoměn. Pro každý typ kryptoadresy se projde celý obsah stránky. Může se stát, že některé adresy patří do více druhů kryptoměn. Toto je vyřešeno na konci skriptu při sestavování výsledného asociativního pole tak, že se do parametru typ přidají všechny typy kryptoměn, do kterých daná adresa spadá. Adresa však může obsahovat v sobě substring, který může být vyhodnocen jako kryptoadresa jiného typu, která je však nesmyslná. Zde je proto zavedena filtrace těchto adres pomocí kontroly, zdali se nově nalezená adresa nenachází jako substring v již některé stávající adrese, která byla extrahována.

Po sesbírání všech kryptoadres ze stránky následuje fáze extrakce metadat. Nejprve se extrahují dodatečná metadata z webu blockchain.info, která mohou obsahovat název/jméno potenciálního vlastníka dané adresy. Potom následuje pokus o extrakci metadat z webu, ze kterého parsujeme

samotné kryptoadresy. Zde je princip takový, že přes XPath je nalezen element, který obsahuje jako text jednu z kryptoadres. Od tohoto elementu se získá jeho rodič, který může potenciálně obsahovat ve svých potomcích nějaké dodatečné informace o dané adrese. Jelikož ale většina těchto informací může být zbytečná, obsahuje skript několik filtrů. Nejdůležitější je filtrace potomka, který obsahuje danou kryptoadresu, kterou máme již extrahovanou a není potřeba ji ukládat do metadat. Další filtry jsou čistě textové: filtrují se prázdné řetězce a řetězce obsahující pouze číslice. Může se také stát, že rodič daného elementu obsahuje větší množství potomků, které vůbec nejsou relevantní (např. je rodičem div, který obsahuje velký obsah webu spolu s elementy obsahujícími kryptoadresy). Potomci jsou tedy procházení pouze do určitého daného omezeného množství (konkrétně 5, tato hodnota je libovolně upravitelná).

4 Implementace

Skript je implementován jako konzolový příkaz v Laravelu. Vstupní pole s URL adresami obsahujícími kryptoadresy jsou pak předány pomocí parametru příkazové řádky. Výstupní asociativní pole s metadaty je na konci skriptu procházeno v cyklu a jednotlivé hodnoty jsou uloženy do databáze. Pro danou tabulku v databázi je vytvořen model, kde v cyklu ho je třeba v každém průchodu instancovat, naplnit daty z asociativního pole a uložit do DB.

Z důvodu chybového hlášení knihovny `DOMDocument`, které se může vyskytnout, když HTML stránka obsahuje HTML5 elementy, ale nemá nastaven DOCTYPE na HTML5, je vypnuto chybové hlášení v PHP pomocí funkce `error_reporting`.

5 Příklad spuštění

Skript se dá spustit z příkazové řádky tímto způsobem:

```
php artisan bitcoin:parse "adresa1" "adresa2" ... "adresaN"
```

Všechny adresy poskytnuté uživatelem jsou pak procházeny a na výstupu v příkazové řádce je zobrazen progress bar.

6 Nalezené URL adresy obsahující kryptoadresy

V rámci zadání projektu bylo nalézt URL adresy obsahující kryptoadresy. Jejich výčet je zde, dále ho lze nalézt v přiloženém souboru addresses.txt.

- https://bitcoinwhoswho.com/blog/2016/02/29/bitcoin-donation-address-rankings/
- https://bitinfocharts.com/top-100-richest-bitcoin-addresses.html
- http://www.theopenledger.com/9-most-famous-bitcoin-addresses/
- https://bitinfocharts.com/top-100-richest-litecoin-addresses.html
- https://99bitcoins.com/bitcoin-rich-list-top1000/
- https://blockchain.info/popular-addresses
- https://btc.com/stats/rich-list
- https://bitcoinchain.com/block_explorer/catalog

- http://bitcoinforcharity.com/bitcoin-charity-list/
- https://github.com/pw2393/crypto-deanonymization/wiki/Famous-Bitcoin-Public-Address

7 Rozšiřitelnost

Skript byl implementován s důrazem na rozšiřitelnost. Všechny připomínky se budou týkat zdrojového souboru /app/Console/Commands/ParseBitcoin.php.

- URL adresy jsou předávány parametrem příkazové řádky.
- Stahování zdrojového kódu HTML stránky lze v skriptu přepínat. Na výběr je stahování pomocí knihovny curl nebo pomocí vestavěné PHP funkce `file_get_contents`. Do funkce `getUrlContent` lze doimplementovat libovolnou další metodu stahování.
- Na začátku skriptu (řádky 39 - 44) jsou definovány regulární výrazy, které skript používá k extrakci kryptoadres. Tyto regexy se dají libovolně rozšířit.
- Na řádku 33 jsou definovány weby, ze kterých se pro každý typ kryptoadresy dají extrahovat dodatečná metadata. Tyto weby mohou fungovat i pro ověření validity dané adresy.
- Funkce `getPrimaryMetadata` obsahuje XPath, který nalezne rodiče elementu obsahujícího procházenou kryptoadresu (řádek 316). Samotná pravidla filtrace extrahovaných potenciálních metadat jsou na řádcích 324 - 327. Zde je možnost přidat další pravidla, upravit stávající nebo je odebrat.