Vysoké učení technické v Brně
Fakulta informačních technologií

ISA projekt - manuál

Čtečka novinek ve formátu Atom a RSS s podporou TLS

Pavel Hurdálek (xhurda01)

10. listopadu 2022

Obsah

Úvod

Cílem projektu je navrhnout a implementovat jednoduchý přenositelný konzolový program, který vypisuje informace uvedené ve stažených zdrojích (feed) ve formátu Atom a RSS 2.0. Program po spuštění stáhne zadané zdroje a na standardní výstup vypíše informace požadované uživatelem podle zadaných přepínačů. Přičemž program automaticky vypíše název zdroje uvozený znaky "*** "a ukončený znaky "***", spolu s nadpisy jednotlivých dílčích prvků (např. nadpisy článků). Dále je možné vypsat čas poslední aktualizace, jméno/email autora a asociované URL (pokud jsou informace uvedené). [4]

Pokud má zadaná webová stránka digitální certifikát SSL/TLS je nutné ho ještě před navázáním spojení zkontrolovat.

Návrh programu

Program je implementován v jazyce C a orientován na platformu *Linux*. Je rozdělen na pět modulů (atom_parser, rss_parse, ssl, feedreader_logic a error_handling), které řeší příslušné aspekty funkcionality (viz Popis implementace). Uložení a předávání dat potřebných v rámci celých modulů je realizováno pomocí globálních proměnných.

Pro práci s certifikáty, návázaní spojení (https i http) a je použita knihovna openssl [3].

Na zpracování xml souborů je využita knihovna libxml2 [5].

Zpracování parametrů není implementováno pomocí žádné knihovny a proto není možné při spuštění kombinovat přepínače (např. -Tau).

Při drobných chybách jako: vícenásobné uvedení stejného přepínače, chybný zdroj dat ve feedfile atd., není program ihned ukončen, ale pokračuje dál, přičemž na konci své činnosti vrací příslušný chybový kód.

Popis implementace

Pro zjednodušení programu a minimalizaci chyb zde není dynamická alokace paměti. Proto je jsou veškeré url adresy, požadavky na server atd. načítány do bufferu o velikosti 1024 znaků (pojme tedy 1023 znaků bez koncové nuly).

Data ve formátu XML od serveru jsou z obdobného důvodu uložena do dočasného souboru "_tmp.feedreader", ze kterého se potom načítají a dále zpracovávají.

Struktura

Samotná aplikace se skládá z modulů, které jsou následující:

- atom_parser - zpracování souboru ve formátu atom
  - atom_parser.c
  - atom_parser.h
- rss_parser - zpracování souboru ve formátu RSS 2.0
  - rss_parser.c
  - rss_parser.h
- ssl - navázání spojení, ověření cetifikátu SSL/TLS, stažení dat
  - ssl.c
  - ssl.h
- feedreader_logic - zpracování parametrů (přepínačů), načítaní xml souboru
  - feedreader_logic.c
  - feedreader_logic.h
- error_handling - definování chybových hlášek, návratových hodnot
  - error_handling.h

Další soubory:

- tests/* - složka s testy
- feedreader_tester.sh - testovací script (bash)
- feedreader.c - funkce main
- makefile - překlad, spuštění testů
- README.md - základní informace o projektu
- manual.pdf - tento manuál

Překlad

Pro překlad je využit překladač gcc a standart gnu99, zejména kvůli zkušenostem z minulých projektů.

Pro odchycení co největšího množství chyb jsou použity následující c flagy: -Wall -Wextra -Werror

Pro překlad libxml2 se musí dynamicky zjistit flagy a umístění knihoven. Proto se v souboru makefile volá shell: shell xml2-config –cflags –libs

Pro správný překlad knihovny openssl se na konec přidávají flagy: -lssl -lcrypto

Knihovna openssl

Modul ssl, který používá knihovnu openssl, je do velké míry inspirován tutoriálem o jejím používání [2]. Je zde popsáno, jak komunikovat se nezabezpečenou i zabezpečenou webovou stránkou a jakým způsobem případně ověřit platnost certifikátů SSL/TLS.

Dalším zdrojem pro tvorbu tohoto modulu byl článek na OpenSSL Wikipedii o tvorbě SSL/TLS klienta[7], ve kterém popsaný celý tento proces a k tomu jsou i přiloženy části kódu, popř. možnost si celý tento program stáhnout a vyzkoušet. Článek je chráněn OpenSSL licencí.

Použitý kód

Při načítání URL adres ze souboru feedfile se používá funkce fgets(), která načte řádek i s koncovým znakem '\n'. Pro odstranění tohoto znaku je použit kód ze stackoverflow.com.

```
url[strcspn(url, "\n")] = 0;
```

URL otázky: removing-trailing-newline-character-from-fgets-input
URL autora odpovědi: Tim Čas

Základní informace o programu

Průběh

Nejdříve se zpracují parametry a ošetří se neplatné kombinace, či chybějící vstupy.

Poté se již inicializuje knihovna openssl a podle vstupů se začne zpracovávat URL adresa nebo feedfile. Zpracování feedfile se liší pouze tím, že se v cyklu načítají řádky souboru, u kterých se určí zda se jedná o URL adresu, komentář nebo prázdný řádek. Poté už probíhá zpracování stejně.

U zpracování URL adres se nejdříve adresa identifikuje protokol, nasledovně rozdělí cesta a samotná doména. Pomocí těchto informací se vytvoří spojení se serverem, případně ověří platnost cerfifikátů. Nakonec se odešle požadavek na získání dat, které ukládá do dočasného souboru (_tmp.feedreader).

V dalším kroku se tento soubor načítá a zpracovává pomocí knihovny libxml2 [5]. Zde zjistí o jaký formát se jedná a podle toho se dále zpracovává Atom nebo RSS 2.0.

Zpracování formátu Atom

Při zpracování se vychází zejména ze stránky Wikipedie Atom (web standard) [1].

Zde se rovnou prochází děti kořenového uzlu, přičemž uzel 'title' vypíše a uzel 'entry' se dále zpracovává. Tento uzel se prochází a ukládá si ukazatele na důležité poduzly. Ty se po průchodu celým uzlem 'entry' podle zadaných parametrů zpracují/vypíší, pokud je v uzlu obsažen titulek, jinak se jde na další.
U uzlu 'author' se vypíše jméno i email (pokud jsou uvedeny), pokud není uvedeno ani jméno ani email, ačkoliv existuje uzel 'author', vypíše se: "Autor: není uvedeno".
U uzlu 'link' se vypíše asociovaná URL adresa (pokud je uvedena), pokud není, ačkoliv existuje uzel 'link', vypíše se: "URL: není uvedena".

Zpracování formátu RSS 2.0

Při zpracování se vychází zejména ze specifikace formátu RSS 2.0. [6].

Zde se zkontroluje zda kořenový uzel obsahuje uzel 'channel', který se dále zpracovává. Tento uzel prochází své děti, poduzel 'title' se vypíše a hledá poduzly se jménem 'item'.
Děti uzlu 'item' se postupně prochází a ukládá si ukazatele na důležité poduzly. Ty se po průchodu celým uzlem 'item' podle zadaných parametrů zpracují/vypíší, pokud je v uzlu obsažen titulek, jinak se jde na další.

Návratové hodnoty

Při drobných chybách jako: vícenásobné uvedení stejného přepínače, chybný zdroj dat ve feedfile atd., není program ihned ukončen, ale pokračuje dál, přičemž na konci své činnosti vrací příslušný chybový kód. Při vážnějších chybách je program ihned ukončen.

- 0 - průběh bez chyby
- 1 - chyba při zpracování parametrů
- 2 - chyba při zpracování URL adresy
- 3 - chyba při navazování spojení pomocí openssl
- 4 - chyba při zpracování XML
- 9 - nedefinovaná chyba
- 42 - interní chyba - pouze pro předávání chyb v rámci programu

Návod na použití

Instalace

Instalace/reinstalace je velice jednoduchá, stačí v terminálu ve složce s projektem zadat příkaz: **make**. Tento příkaz zkompiluje zdrojový kód a vytvoří spustitelný **feedreader**.

Odinstalaci provedeme obdobně pomocí příkazu: **make clean**, který tento soubor odstraní.

Testy spustíte pomocí příkazu: **make test**

Spuštění

Máme dvě možnosti jak program spustit. Můžeme si nechat vypsat nápovědu nebo zadat zdroj dat a nechat si vypsat požadované informace:

```
./feedreader <URL | -f <feedfile>> [-c <certfile>] [-C <certaddr>] [-T] [-a] [-u]
./feedreader <-h | --help>
```

POZOR: není možné kombinovat přepínače, tedy např.: -Tau

Parametry (nezáleží na pořadí):

- -h | –help - Výpis nápovědy
- URL - Url adresa zdroje
- -f <feedfile> - název souboru s url adresami zdrojů
- -c <certfile> - soubor s certifikáty pro ověření platnosti certifikátu SSL/TLS
- -C <certaddr> - adresář, ve kterém se mají vyhledávat certifikáty (SSL/TLS)
- -T - pro každý záznam zobrazí navíc informace o čase změny záznamu
- -a - pro každý záznam zobrazí jméno autora, či jeho e-mailová adresa
- -u - pro každý záznam zobrazí asociované URL

Příklad spuštění

```
$ ./feedreader https://what-if.xkcd.com/feed.atom -T

*** what if? ***
Transatlantic Car Rental
Aktualizace: 2022-09-06T00:00:00Z

Hailstones
Aktualizace: 2022-07-06T00:00:00Z

Hot Banana
Aktualizace: 2022-05-04T00:00:00Z

Earth-Moon Fire Pole
Aktualizace: 2018-05-21T00:00:00Z

Electrofishing for Whales
Aktualizace: 2017-03-09T00:00:00Z
```

Testování

K projektu jsou vytvořené testy, které se dají spustit pomocí příkazu:

```
make test
```

Příkaz spustí bash script, který provede testy definované soubory ve složce tests/. Vždy načte příkaz s předem definovanými parametry ze souboru */command.txt, přičemž výstup (stdin) je přesměrován do */output.txt, poté porovná návratovou hodnotu s hodnotou v */return.txt a nakonec porovná výstup se správným výstupem v souboru */corr_output.txt.
Soubor */output.txt je smazán, pokud je test úspěšný, jinak soubor zůstává, abychom se mohli podívat na reálný výstup.

Testy jsou rozděleny do 4 kategorií:

1. testování parametrů - zejména chybné vstupy, jako: nezadán zdroj, přepínač bez povinné doplňující informace atd.
2. testování chybně zadané URL adresy - nepodporovaný protokol nebo zadán pouze protokol bez adresy
3. testování programu, při zadání URL adresy jako zdroje - různé kombinace přepínačů
    - testy formátu Atom
    - testy formátu RSS 2.0
4. testování programu, při zadání feedfile souboru jako zdroje - různé kombinace přepínačů, pouze komentáře a prázdné řádky, chybné zdroje a kombinace zmíněných

Na konec vypíše testovací script počet úspěšných a neúspěšných testů.

POZOR: 3. a 4. kategorie testů závisí na obsahu webových stránek https://what-if.xkcd.com/feed.atom a https://www.rssboard.org//files/sample-rss-2.xml, slouží tedy hlavně pro testování při vývoji.

Mimo tyto testy, jsem program manuálně testoval na těchto adresách:

- http://www.theregister.com/headlines.atom
- https://www.theregister.com/headlines.atom
- http://www.theregister.com/software/headlines.atom
- https://www.theregister.com/software/headlines.atom
- https://en.wikipedia.org/w/api.php? hide-bots=1&days=7&limit=50&hidewikidata=1&action=feedrecentchanges&feedformat=atom

Odkazy

[1] *Atom (web standard)*. URL: https://en.wikipedia.org/wiki/Atom_(web_standard). (accessed: 10.11.2022).

[2] Kenneth Ballard. *Secure programming with the OpenSSL API*. URL: https://developer.ibm.com/tutorials/l-openssl/. (accessed: 06.11.2022).

[3] *OpenSSL Cryptography and SSL/TLS Toolkit*. URL: https://www.openssl.org/. (accessed: 06.11.2022).

[4] Libor Polčák. *ISA - zadání projektu: Čtečka novinek ve formátu Atom a RSS s podporou TLS*. URL: https://www.vut.cz/studis/student.phtml?script_name=zadani_detail&apid=231021&zid=50242. (accessed: 06.11.2022).

[5] *Reference Manual for libxml2*. URL: http://xmlsoft.org/html/. (accessed: 06.11.2022).

[6] *RSS 2.0 Specification*. URL: https://www.rssboard.org/rss-specification. (accessed: 10.11.2022).

[7] *SSL/TLS Client*. URL: https://wiki.openssl.org/index.php/SSL/TLS_Client. (accessed: 06.11.2022).