Gymnázium, Brno, třída Kapitána Jaroše 14

Vývoj aplikace na kontrolování změn na webových stránkách

Závěrečná maturitní práce

Vedoucí práce:
Mgr. Marek Blaha

Jiří Kalvoda

Blansko 2020

TODO poděkování

Prohlašuji, že jsem tuto práci vyřešil samostatně s použitím literatury, kterou uvádím v seznamu

V Blansku dne 28. prosince 2019 .....

Abstract

Kalvoda, J.

Development an aplication for checking changes on web pagesabstract

Abstrakt

Kalvoda, J. Vývoj aplikace na kontrolování změn na webových stránkách

Tato závěrečná práce popisuje vývoj a použití aplikace na monitorování změn na webových stránkách. Aplikace je vyvíjena v jazyce c++ pomocí knihovny Qt. Díky tomu se jedná o multiplatformní software. Je dostupná včetně zdrojového kódu pod licencí GNU LGPL. Tato závěrečná práce obsahuje popis jejího fungování, implementace a použitého softwaru při jejím vývoji.

Obsah

1 Úvod a cíl práce

1.1 Úvod do problematiky

Webové stránky se mohou neustále měnit, proto je dobré automaticky monitorovat jejich aktualizace. Tato aplikace umožňuje automatizovat tento problém a tím uživateli ušetřit čas a eliminovat lidský chybový faktor.

Aplikace podporuje různé tolerance při načítání a porovnávání změn stránek. Například na stránce, kde se část neustále mění, mohu tuto část vypustit, nebo přímo porovnávat jen nějaké části a podobně. Díky práci s cookies je možné navázat i složitější spojení se serverem a provést definovanou sekvenci úkolů (např. přihlásit se a načíst nějaký soukromý obsah). Historie stránek se může ukládat a pak lze v ní vyhledávat a zjišťovat rozdíly mezi verzemi pomocí grafického porovnávání napojeného na uživatelův oblíbený prohlížeč.

1.2 Cíl práce

Cílem této práce je vyvinout funkční aplikaci umožnující zjišťování aktualizací, archivaci a porovnávání webových stránek (případně i jiných dokumentů) a publikovat ji cílovým uživatelům na různých operačních systémech. K aplikaci také bude vypracována rozsáhlá uživatelská i technická dokumentace, která umožní její další vývoj. Umožním tedy dalším programátorům tuto aplikaci pohodlně modifikovat a upravovat dle svých potřeb. Cílem této práce je také aplikaci rozšířit mezi skupinu testovacích uživatelů a použít jejich připomínky a problémy k dalšímu vývoji a stabilizaci aplikace.

2 Přehled literatury

3 Popis fungování a ovládání aplikace

3.1 Instalace

Linux

Windows

macOS

3.2 Seznam stránek na kontrolu

Při spuštění si aplikace načte seznam stránek ke kontrole. Ten je obsažen v souboru `pages.json`, který musí být umístěn v adresáři aplikace (respektive v adresáři, kde se aplikace spouští). Soubor musí být validní json. V případě, že soubor neexistuje nebo není validní, aplikace vypíše upozornění. Očekává se, že soubor obsahuje pole struktur. Každá z nich obsahuje informace o jedné stránce, která se má kontrolovat.

Adresa stránky a název

Každou stránku je nutné pojmenovat jednoznačným identifikátorem. Proto je nutné u každé stránky definovat položku označenou klíčem `name`. Toto jméno se pak zobrazuje v seznamu změn a také je nutné při vyhledávání v historii. Jméno se také používá jako identifikátor v databázi i jako identifikátor souborů jednotlivých verzí stránky.

Dále je nutné u každé stránky definovat její umístění na webu, tedy její adresu. K tomu slouží položky `server`, `dir`, `file`. Ovšem není zapotřebí definovat všechny tyto položky. Pod klíčem `server` by měla být definována adresa serveru, na kterém je požadovaná stránka včetně přístupového protokolu. Tedy například `https://is.jaroska.cz` nebo pomocí ip adresy může zápis vypadat následovně `http://195.178.65.1` V případě, že není požadovaná stránka na serveru umísténa v jeho kořenové složce, je doporučeno název složky (popřípadě celou cestu několika zanořených složek) umístit do položky `dir`. `file` obsahuje jméno požadovaného souboru včetně přípony V případě, že je požadován přístup základní soubor (index) na dané adrese, není potřeba `file` uvádět. Dále zde také můžou být obsaženy parametry stránky, které se předávají metodou `GET`. Stačí je uvézt za otazník. Takový zápis může vypadat následovně: `index.php?akce=42&akcicka=0`.

V případě, že je nutné předat stránce informace pomocí protokolu `POST` (typicky při přihlašování na stránky), je možné v zápisu stránky užít klíče `post`.

Nejjednodušší způsob, jak zjistit tyto informace pro požadovanou stránku je využít prohlížeč. Většina prohlížečů totiž umožnuje zobrazit informace o navázaném spojení. Odsud stačí potřebné informace jen zkopírovat. V prohlížečích založených na jádře Chromium lze se lze pomocí klávesy `F12` dostat do vývojářského panelu. V záložce `Network` je možné najít příslušný soubor, jehož hlavičku je třeba použít.

Způsoby kontroly změn

U některých stránek často mění její část, i když sledovaný obsah se nezmění. Pro omezení kontrolování je proto vhodné užít v zápisu položky `diff`. Ta by měla obsahovat speciální strukturu popisující způsob ignorování změn[1]. Tato struktura může obsahovat následující položky:
Klíč `ignore` s libovolnou hodnotou znamená, že stránka se vůbec nebude kontrolovat na změny. Toto je vhodné například když se jedná pouze o přihlašovací stránku, která neobsahuje žádaná data.
Tag `ignoreSector` může obsahovat pole struktur. Každá z nich může obsahovat informace o jednom nutném vynechání pomocí tagů `start`, `end` a `countOfEnd`, které znamenají, že text od `start` po `countOfEnd`-tý výskyt řetězce `end` bude při porovnávání vynechán. Začátek i konec může být definován pomocí regulárního výrazu. Na přesnou implementaci regulárních výrazů lze nahlédnout do dokumentace Qt na adrese `https://doc.qt.io/qt-5/qregexp.html`. Tagy `end` a `countOfEnd` nemusí být uvedeny. V takovém případě je `end` nastaveno na konec řádku a `countOfEnd` na 1.
Obdobným způsobem lze použít tag `onlySector`, který také může obsahovat pole struktur složených z `start`, `end` a `countOfEnd`. Při použití tohoto tagu budou porovnávány pouze změny v těchto úsecích (budou ignorovány úseky od začátku k prvnímu výskytu, mezi nimi a od posledního na konec).
Další možností je využít klíče `permutation` s libovolnou hodnotou. Jeho použití znamená, že libovolné permutace znaků budou považovány za stejné. Toto je vhodné v případě, že se na stránce některé objekty náhodně prohazují.
Tyto omezení porovnávání se provádí v zde uvedeném pořadí.

Skupiny stránek, přihlášení a práce s cookie

Například v případě, že je nutné pro získáni informací se na nějakou stránku přihlásit, je možné využít v zápisu stránky položku `cookie`. Pomocí ní lze spojit více stránek do skupiny, ve které si stránky mezi sebou ukládají cookies. Stačí pouze u všech stránek vyplnit stejnou hodnotou tagu `cookie`. Když se některou z stránek nepodaří načíst, v daném průchodu se nebudou načítat ani další stránky ze stejné skupiny.

[1] Případně, že struktura může obsahovat řetězec `"ignore"`, který má stejný efekt jako `{"ignore":1}`

3.3 Spuštění kontroly, tabulka změn, informační konzole

3.4 Otevření lokální kopie stránky, její napojení

3.5 Historie změn a její procházení

3.6 Grafické porovnávání verzí stránek

3.7 Klávesové zkratky

4 Implementace aplikace

4.1 Použitý software

Qt

Git

4.2 Objektový model, rozdělení problému

4.3 Pozadí aplikace

4.4 Grafické uživatelské rozhraní