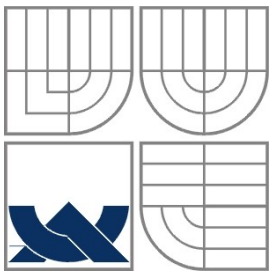
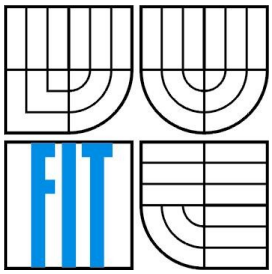


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

IPP 2016/17 – DOKUMENTACE K PROJEKTU 2 : CSV

AUTOR Lukáš Drahník
XDRAHN00@STUD.FIT.VUTBR.CZ

BRNO 2017

Obsah

Obsah....1
1 Úvod....2
2 Návrh....3
3 Základní implementace....5
3.1 Parsování argumentů....7
3.2 Načtení a validování CSV....7
3.3 Generování XML....7
4 Rozšíření....5
4.1 PAD....7
4.2 VLC....7
5 Závěr....8

1 Úvod

Cílem druhého projektu a varianty CSV bylo vytvořit skript v jazyce Python3.6 pro konverzi formátu CSV (viz RFC 4180) do XML.

2 Návrh

Na parsování argumentů z příkazové řádky byla použita standartní knihovna argparse, kterou doplňuje ruční ověřování a donastavování každého argumentu podle zadání.

CSV soubor je načítán pro zpracování i pro výpis do XML souboru pomocí standartní knihovny pro práci s csv, přísnější validace na kterou navazuje i rozšíření VLC je prováděna ručně znak po znaku.

Zápis struktury do XML souboru je prováděn ručně bez pomoci externí knihovny.

Zdrojový kód (názvy argumentů, funkcí, proměnných) je v angličtině z důvodu konzistence se zadáním a tedy snadnějšího orientování v kódu. Komentáře kódu včetně nápovědy k argumentům a dokumentace jsou psány v českém jazyce.

Veškeré načítání souborů bylo prováděno s kódováním UTF-8. Při dodatečné kontrole byly znaky převáděny na ascii hodnoty, aby CSV formát byl validován vůči striktnímu výkladu RFC 4180.

3 Základní implementace

3.1 Parsování argumentů

Parsování argumentů z příkazové řádky se provede hned po zavolání skriptu a je vykonáno pomocí knihovny argparse. Funkci této knihovny bylo potřeba doplnit o validování argumentů a dodatečné donastavování z důvodu dodržení zadání.

Implementace parsování argumentů probíhá ve třech krocích. Začíná použitím knihovny argparse (základní parsování a validování). Pokračuje vlastním validováním argumentů z důvodu dodržení zadání (některé argumenty nemohou být nataveny současně, některé musí mít specifickou hodnotu) a končí donastavováním defaultních hodnot argumentů (knihovna nepodporuje nastavení defaultní hodnoty pokud argument není vůbec zadán).

Nastavení argumentu --help nemohlo být provedeno skrze objekt knihovny, protože argument --help nemohl být použit současně s jakkýmkoliv argumentem, musela být provedena vlastní validace a až poté bylo možné zavolat ručně nápovědu knihovny argparse.

3.2 Načtení a validování CSV

Prvotní načtení souboru je provedeno standartní knihovnou pro práci s csv (mmj. slouží jako první filtr, který odhalí základní chyby spojené s nevalidním souborem, tedy pokud dojde k chybě vrací error s návratovým kódem 4). Validování pracuje již bez knihovny pro práci s csv, tedy s načteným souborem převedeným do jednoho řádku (soubor je převeden do jednoho řádku z důvodu snadnější validace složitějších pravidel - například zadávání několikařádkových položek v ohraničujících uvozovkách). Validování slouží pouze pro validaci. Není zde tedy proveden převod znaků dvojitých uvozovek na jeden nebo vynechání ohraničujících dvojtých uvozovek.

I bez vypnutého argumentu --validate zde dochází k validování z důvodu nedostatečné kvality standartní knihovny pro práci s csv. Knihovna striktně nekontroluje ukočení řádků pomocí CLRF, protože považuje za korektní i ukončení pomocí LF, na posledním řádku toleruje použití oddělovače

řádků, knihovna také toleruje znaky mezi separátorem a ohraničujícími uvozovkami. Pro již zmíněné případy program vrací návratový kód 39.

3.3 Generování XML

Výsledné generování XML souboru je provedeno manuálně, ale vychází z parsování CSV souboru pomocí standartní knihovny. K tomu bylo potřeba doimplementovat převádění problematických znaků s menším ascii kódem než 127 (`<`, `>`, `&`), s větším než 127 převáděny dle zadání nebyly.

Kontrola validnosti názvu XML elementu vychází z doporučení o XML[1]. Zvlášť se validují první a zbylé charaktery.

4 Rozšíření

4.1 PAD

Rozšíření se aktivuje pomocí argumentu --padding. Pro rozšíření bylo přidáno počítání potřebného odsazení (tzn. počet přidaných 0 jako prefix počítadla sloupců) při prvním zpracování souboru pomocí standartní knihovny pro práci s csv. Vypočítané hodnoty jsou použity při převodu do XML.

4.2 VLC

Rozšíření rozšiřuje omezenou validaci poskytnutou knihovnou pro standartní zpracovaní csv. Aktivace rozšíření je podmíněna zadáním argumentu --validate. Validování znak po znaku je odvozeno z RFC 4180 a uvedené ABNF gramatiky.

Přijde se tedy na chyby v počtu zadaných uvozovek vedle sebe (v uvozovkované položce musí být dvojitá uvozovka doprovázená další), na nepovolené znaky v bloku ohraničeném nebo neohraničeném dvojitými uvozovkami. Jako chyba, s kterou si standartní knihovna neporadí, vůči striktnímu výkladu RFC je také považováno umístění znaku mezi separátory a ohraničujícími uvozovkami. Pro zmíněné případy poté skript vrací návratový kód 39.

Ve výstupním XML souboru není nijak speciálně nakládáno (nepřidává se indent na aktuální zarovnání sloupce) se znaky CR, LF, CRLF uvnitř ohraničujících dvojitých uvozovek.

5 Závěr

Testování bylo provedeno pomocí poskytnutých školních testů a u rozšířeních byl výstup testován ručně.

1: John Cowan, Google, Andrew Fang, PTC-Arbortext, Paul Grosso, PTC-Arbortext (Co-Chair), Konrad Lanz, A-SIT, Glenn Marcy, IBM, Henry Thompson, W3C (Staff Contact), Richard Tobin, University of Edinburgh, Daniel Veillard, Norman Walsh, Mark Logic (Co-Chair), François Yergeau, Extensible Markup Language (XML) 1.0 (Fifth Edition),