Dokumentace úlohy CSV: CSV2XML v Python 3 do IPP 2015/2016
Jméno a příjmení: Klára Nečasová
Login: xnecas24

1. Úvod

Cílem projektu bylo vytvořit skript `csv.py` v jazyce Python 3, který převede vstupní soubor ve formátu CSV do formátu XML. Činnost skriptu `csv.py` lze rozdělit do následujících částí:

- zpracování parametrů příkazové řádky,
- zpracování vstupního CSV souboru,
- ošetření problematických a nepovolených znaků,
- generování výstupního XML souboru.

2. Zpracování parametrů příkazové řádky

Parametry skriptu jsou zpracovány funkcí `getParams()` pomocí modulu `argparse` (třída `ArgumentParser`). Funkce `getParams()` rovněž zajišťuje ošetření všech chybových stavů a nastavení implicitních hodnot u některých přepínačů. Nedostatkem třídy `ArgumentParser` je vyvolání výjimky v případě zadání nesprávných parametrů příkazové řádky. Bylo nutné zdědit tuto třídu a redefinovat metodu `error()` tak, aby vracela chybový kód odpovídající zadání. Také bylo nutné vypořádat se s problémy při importování modulu `csv`, neboť se jeho název shoduje s názvem skriptu, do kterého se modul importuje. Problém s importováním zmíněného modulu byl vyřešen úpravou proměnné `sys.path`.

3. Zpracování vstupního CSV souboru

K načtení vstupního CSV souboru je využit modul `csv`, konkrétně metoda `reader()`. Tato metoda načte vstupní soubor a vrátí každý načtený řádek CSV souboru jako seznam řetězců.
Nejdříve je načten první řádek CSV souboru, na základě kterého je získána informace o počtu sloupců, poté je načten zbytek obsahu CSV souboru. Přepínač `-s=separator` specifikuje oddělovač záznamů CSV souboru, implicitní hodnota je znak pomlčky. Hodnota přepínače `-s` musí být jednoznaková.

4. Ošetření problematických a nepovolených znaků

V případě přepínače `-h` je provedena náhrada *nepovolených znaků*, a to buď řetězcem `subst` (v případě volby `-h=subst`) nebo znakem "-" (pomlčka, což je implicitní hodnota přepínače `-h`). Za nepovolené znaky jsou považovány znaky, které nejsou definovány jako platné (sekce 2.2 v dokumentu W3C [1]), bílé znaky a znaky čárka (`,`), středník (`;`), svislá čára (`|`), stříška (`^`) a vlnovka (`~`).
Pokud i po náhradě nepovolených znaků vznikne invalidní XML element, skript skončí s příslušným chybovým kódem. Validitu XML elementu zajišťuje funkce `validateXMLElement()`, detekce nepovolených znaků je zpracována též na základě dokumentů W3C [1], sekce 2.3. Tato funkce je využita i při validaci ostatních XML elementů daných přepínači `-r`, `-l` a `-c`. U těchto přepínačů ovšem nedochází k překódování nepovolených znaků, a pokud je zadána hodnota vedoucí na nevalidní XML element, skript je ukončen s příslušným chybovým kódem. Při detekci nepovolených znaků a validaci XML elementů jsou využity regulární výrazy (modul `re`).

Problematické znaky XML jsou znaky, které mají UTF-8 kód menší než 128. Takové znaky jsou nahrazeny pomocí zápisů obsahující metaznaky `&`, o což se stará funkce `convertChars()`. Mezi problematické znaky se řadí znaky ampersand (`&`), menší než (<), větší než (>), dvojitá uvozovka (`"`) a apostrof (`'`) (dokument W3C [1], sekce 2.4). Problematické znaky je potřeba řešit v obsahových buňkách CSV souboru a také v případě volby přepínače `--missing-field=val`, kde `val` je hodnota, která se do výstupního XML souboru doplní, pokud chybí nějaká vstupní buňka.

5. Generování výstupního XML souboru

Převod CSV souboru do XML formátu zastřešuje funkce `CSVToXML()`. Zpracování CSV souboru probíhá ve dvou cyklech. První cyklus zajišťuje vytištění řádkového elementu a případně tisk atributu `index` s hodnotou danou přepínačem `--start=n`. Řádkový element je definován přepínačem `-l=row-element`, kde `row-element` je řetězec, kterým budou řádky označeny. Implicitní hodnota přepínače `-l` je řetězec `row`.

Druhý cyklus zajišťuje zpracování dat nacházejících se uvnitř řádkového elementu – tisknou se tedy jednotlivé XML elementy označující sloupce a samotná data získaná z CSV souboru. Názvy sloupců jsou buď odvozeny z prvního řádku CSV souboru (pokud je zadán přepínač `-h`) nebo jsou definovány pomocí přepínače `-c=column-elementX`, kde `column-element` je řetězec, kterým budou sloupce označeny, a `X` je čítač sloupců, implicitně nastaven na hodnotu 1. Implicitní hodnota přepínače `-c` je řetězec `col`.

5.1 Zotavení z chybného počtu sloupců ve vstupním CSV souboru

Pokud nějaký řádek vstupního CSV souboru obsahuje neodpovídající počet sloupců, je skript ukončen s příslušným chybovým kódem. V případě, že je zadán parametr `-e` nebo `--error-recovery`, zajistí skript doplnění chybějících sloupců – vygeneruje se prázdný XML element, např. `<tag></tag>`. Je nutné zjistit, kolik prázdných elementů bude potřeba doplnit. Tento údaj je vypočten jako rozdíl počtu sloupců odvozeného z prvního řádku CSV souboru a počtu sloupců na aktuálním zpracovávaném řádku.

Pokud je zadán přepínač `--missing-field=val`, je namísto prázdného XML elementu vygenerován XML element ve tvaru `<tag>val</tag>`, kde `val` je hodnota, která se doplní v případě, že nějaká vstupní buňka chybí. Tento přepínač je opět nutné kombinovat s přepínači `-e` nebo `--error-recovery`. Sloupce, které jsou ve vstupním CSV souboru navíc, jsou ignorovány.

5.2 Vytištění všech sloupců CSV souboru

Pomocí přepínače `--all-columns` lze zajistit, aby výstupní XML soubor obsahoval všechny sloupce ze vstupního CSV souboru bez ohledu na jejich počet. Pokud se tedy na některém řádku CSV souboru nachází více sloupců, než je specifikováno prvním řádkem CSV souboru, jsou i tyto sloupce součástí výstupu. Sloupce, které jsou navíc, jsou označeny dle hodnoty přepínače `-c=column-elementX`. Popsaný přepínač je možné využít pouze v kombinaci s přepínači `-e` nebo `--error-recovery`.

Literatura

[1] BRAY, Tim, et al. Extensible markup language (XML). *World Wide Web Consortium Recommendation REC-xml-19980210. http://www. w3. org/TR/1998/REC-xml-19980210*, 1998, 16.

[2] SHAFRANOVICH, Yakov. Common format and MIME type for Comma-Separated Values (CSV) files. 2005.