



11.01.2021

CSV File Importer

Dokumentation



Ulrich Stark
OTH AMBERG-WEIDEN

Inhalt

1	Grundlegendes.....	1
2	Funktionen	1
2.1	Importieren von CSV-Dateien.....	1
2.2	Importieren von XML-Dateien.....	2
2.3	Zusammenfügen	2
2.4	Exportieren	2
3	Grafische Anwendung	3
4	Programmierschnittstellen	3

1 Grundlegendes

Das Projekt teilt sich in das „src“ und das „example“ Verzeichnis auf. Im „src“-Verzeichnis befindet sich der Python-Quellcode für den CSV File Importer und einer grafischen Beispielanwendung, die mit dem Framework TK erstellt wurde. Das „example“-Verzeichnis erhält Beispieldateien zum Importieren. Diese teilen sich in CSV, XML und XSL Dateien auf. Alle der Aufgabe beigelegten Dateien wurden beigelegt und wurden um eigene Testdateien ergänzt.

Entwickelt wurde mit dem Versionsverwaltungstool Git und steht unter <https://github.com/ulrichstark/csv-importer> öffentlich verfügbar bereit. Die Commit-Historie ist ebenfalls dort nachzuvollziehen.

Dem Projekt liegt außerdem noch diese Dokumentation als Microsoft-Word und PDF-Datei bei. Über das PowerShell Skript „generateDoc.ps1“ kann zu jedem Modul im „src“-Verzeichnis mithilfe von pydoc eine einfache Dokumentation generiert werden.

Um die grafische Beispielanwendung oder das Konsolen-Testprogramm zu starten, müssen zuerst alle benötigten Python Module installiert sein. Dazu zählen die Module pandas, lxml, tkinter, chardet, csv, re und pandastable. Dann sollte mit dem Kommandozeilenbefehl „cd src“ das aktuelle Verzeichnis auf das Quellcode-Verzeichnis geändert werden. Um dann die grafische Beispielanwendung zu starten, reicht der Befehl „python gui.py“. Zum Starten des Konsolen-Testprogramms benutzen Sie bitte den Befehl „python main.py“

2 Funktionen

2.1 Importieren von CSV-Dateien

Es können CSV-Dateien mit beliebigem Format importiert werden. Dabei erkennt das Programm automatisch das Trenn- und Quotierungszeichen und die Kodierung einer Datei. Außerdem wird versucht, aus der ersten Datenzeile im Vergleich zu den anderen Zeilen zu schließen, ob es sich bei ihr um eine Überschriftenzeile handelt.

Dem Benutzer steht es dabei frei, diese Parameter im Nachhinein zu verändern. Zum Beispiel, wenn das Trennzeichen einer Datei falsch erkannt wurde und somit kein fehlerfreies Importieren gewährleistet werden kann.

2.2 Importieren von XML-Dateien

Zusätzlich dazu steht auch der Import von XML-Dateien bereit. Dazu muss eine XSL-Datei (= Extensible Stylesheet Language) angegeben werden, die diese zu importierende XML-Datei in das CSV-Format „transformieren“ kann.

Auch bei dieser Möglichkeit des Importierens wird die Dateikodierung und die speziellen Parameter des CSV-Formats automatisch erkannt. Die Ausgabe der XSL-Datei kann also ein beliebiges, aber geläufiges Trenn- und Quotierungszeichen wählen und sollte trotzdem ohne Fehler importiert werden können.

2.3 Zusammenfügen

Jede importierte Datei (CSV oder XML) wird in eine interne Tabelle im Speicher des Computers zusammengefügt. Dabei spielt die Reihenfolge der importierten Dateien und die Spaltenanzahl eine wichtige Rolle. Der erste Import setzt mit seiner Anzahl der Spalten die benötigte Spaltenanzahl aller folgenden Imports fest. Sollte eine Datei importiert werden, die sich von dieser Anzahl unterscheidet, wird sie übersprungen und ein Fehler wird ausgegeben.

Für die Beschriftung der Spalten wird die erste gültige Überschriftenzeile eines Imports herangezogen. Wenn keine der importierten Dateien eine derartige Kopfzeile besitzt, kann das Programm aus den Zelleninhalten in der Spalte auf den Datentyp schließen und setzt diesen mit einer fortlaufenden Nummer als Beschriftung.

Folgende Datentypen können mithilfe von regulären Ausdrücken erkannt werden: Geo-Koordinaten, E-Mail-Adressen, URLs, Datum kombiniert mit Uhrzeit, Datum, Uhrzeit, Dezimalzahlen, Ganzzahlen oder boolesche Ausdrücke. Sollte keiner dieser Datentypen zutreffen, wird der allgemeine Typ „Text“ angenommen.

2.4 Exportieren

Die kombinierten importierten Dateien lassen sich daraufhin als CSV- und XML-Datei exportieren. Bei beiden Möglichkeiten müssen der Name, Pfad und Kodierung der zu exportierenden Datei ausgewählt werden.

Während des Exportierens einer CSV-Datei kann zusätzlich dazu noch ein beliebiges Trenn- und Quotierungszeichen gewählt werden. Diese werden zur Generierung der Datei genutzt.

3 Grafische Anwendung

4 Programmierschnittstellen