Handbuch für

Modul zum Import verschiedener Dateiformate

Mark Unger und Siegfried Kienzle

13. November 2016

Erklärung

Die in diesem Projekt verwendete Software unterliegt den rechtlich jeweiligen Bestimmungen der einzelnen Organisationen und Firmen.

Inhaltsverzeichnis

1 Modul

1.1 Über die Software

Dieses Modul dient zur Extrahierung von Text aus Dateien. Sie können dieses Modul für folgende Endungen verwenden:

- .doc
- .docx
- .odt
- .pdf
- rtf

Es wurde für Python 3.4.3 entwickelt und unter Ubuntu 14.04.05 LTS getestet. Zur Installation liegt ein Bash-Script vor.

1.2 Über das Handbuch

Dieses Handbuch beschreibt die Installation und die Handhabung mit dem Modul.

2 Grundlagen

2.1 Installation

1. Installationsscript mittels ./inst.sh aufrufen:

Abbildung 1: Nach Aufruf des Installationscriptes ./inst.sh

2. sudo-Passwort eintippen und die Enter-Taste drücken.
3. Es werden nun einige Abhängigkeiten installiert, die zur Ausführung dieses Moduls benötigt werden.

4. Geben Sie nun den Pfad an, in den das Modul installiert werden soll. Sollte der Pfad nicht existieren, werden Sie wie in Abbildung 4 gefragt ob der Pfad erstellt werden soll. Existiert der Pfad, entfallen die Schritte 6 bis 8.

Abbildung 2: Nach Aufruf des Installationscriptes ./inst.sh

Abbildung 3: Nach Eingabe des Installationspfads

5. Wenn der OK-Button blau hinterlegt ist, können Sie mit der Enter-Taste den Pfad bestätigen.

6. Sollte kein Pfad existieren, erscheint folgendes Fenster:

Abbildung 4: Pfad erstellen?

7. Wählen Sie nun mit den Pfeiltasten aus, ob Sie den Pfad erstellen möchten oder nicht und drücken Sie dann die Enter-Taste.

Abbildung 5: Pfad wurde erstellt

8. Es wurde nun der Pfad erstellt. Drücken Sie nun die Enter-Taste, um die Dateien in das entsprechend vorher erstellte Verzeichnis, zu entpacken.

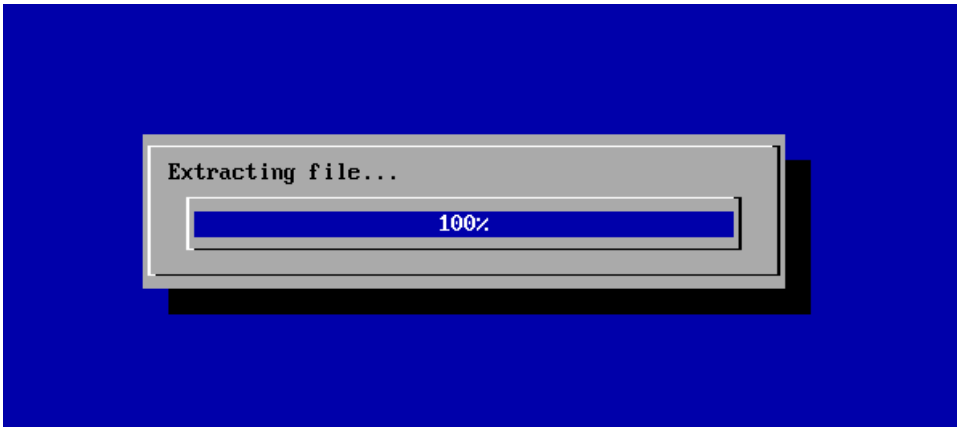

Abbildung 6: Pfad wurde erstellt

9. Die Installation ist nun abgeschlossen. Prüfen Sie nun bitte ob alle Dateien installiert wurden. Eine genaue Auflistung finden Sie unter dem Punkt 2.3.

2.2 Bestandteile Installationspaket

Datei	Beschreibung
inst.sh	Bash-Script für die Ausführung als Super-User (sudo) unter Ubuntu
ubuntu.sh	Bash-Script für die Installation unter Ubuntu
moduls.tar	Tar-Datei, die die Python-Module enthält

Tabelle 1: Bestandteile Installationspaket

2.3 Modulbestandteile

Datei	Verwendung
convertToTxt.py	Hauptdatei zur Extrahierung von Text
docTxt.py	Modul für die Dateiendung doc
docxTxt.py	Modul für die Dateiendung docx
odtTxt.py	Modul für die Dateiendung odt
pdfTxt.py	Modul für die Dateiendung pdf
rtfTxt.py	Modul für die Dateiendung rtf

Tabelle 2: Modulbestandteile

2.4 Erste Schritte

3 Technischer Hintergrund

3.1 Aufbau

3.2 Verwendete Fremdsoftware

4 Kontaktdaten

Name	E-Mail
Mark Unger	mrk.unger@gmail.com
Siegfried Kienzle	siegfried.kienzle@gmx.de