Texterkennung mit Ocropy – Vom Bild zum Text

Nasarek, Robert
robert.nasarek@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Müller, Andreas
andreas.mueller@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Das OCR-Programm ocropy

Die optische Zeichenerkennung (engl. Optical Character Recognition – OCR) von historischen Texten weißt oftmals niedrige Erkennungsraten auf. Mit einem gekonnten Preprozessing und ocropy (auch ocropus), einem modular aufgebauten Kommandozeilenprogramm auf Basis eines neuronalen long short-term memory Netzes, ist es möglich, deutlich bessere Ergebnisse zu erzielen. (Springmann 2015, S. 3; Vanderkam 2015) Ocropy ist in Python geschrieben und enthält u. a. Module zur Binarisierung (Erzeugung einer Rastergrafik), zur Segmentierung (Dokumentaufspaltung in Zeilen), zur Korrektur fehlerhafter Erkennungstexte, zum Training neuer Zeichen und natürlich zur Erkennung von Dokumenten (siehe Abbildung 1). Ein bedeutender Vorteil dabei ist, dass jedes Modul eine Reihe von nachvollziehbaren Einstellungsmöglichkeiten hat, um auf die individuellen Herausforderungen jedes Dokumentes einzugehen. Zusätzlich besteht die Möglichkeit ocropy auf die Erkennung einer bestimmten Schriftart, bzw. eines Zeichensatzes zu trainieren.

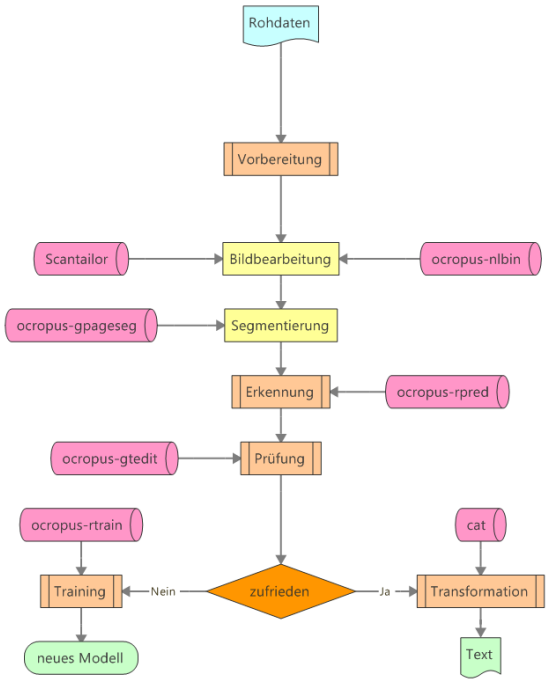

Abbildung 1. Überblick zum Prozessablauf der Texterkennung mit den grundlegenden Software-Modulen

Die Benutzung von ocropy als Kommandozeilenprogramms setzt jedoch den Umgang mit einer Consolen-Umgebung und eine grundlegende Kenntnis von Bash-Kommandos voraus: Für viele potenzielle NutzerInnen stellt dies eine erste Einstiegshürde dar, denn der NutzerInnenanteil von Linuxderivaten beträgt nur 3% (statista 2018), wobei die Gruppe an ShelluserInnen noch kleiner sein dürfte. Im Workshop wird diese Hürde abgebaut, indem alle Schritte „from zero to recognised textfile“ nachvollziehbar und zum Mitmachen aufzeigt wird. Insgesamt werden sechs Themengebiete behandelt, damit die TeilnehmerInnen des Workshops alle benötigten Informationen erhalten, um selbstständig Frakturschriften (oder andere Schriftarten) durch ocropy erkennen zu lassen.

Ubuntu in der VirtualBox

Für bisher ausschließliche NutzerInnen des Betriebssystem Windows oder Mac OS, ist es unverhältnismäßig, allein wegen ocropy Linux als Zweit- oder sogar Hauptsystem zu installieren. Durch die Verwendung einer VirtualBox und des Linux-Derivats Ubuntu kann dieser Schritt umgangen werden. Mit Hilfe einer virtuellen Maschine lässt sich ein Betriebssystem innerhalb eines anderen Betriebssystems emulieren. Das bringt den Vorteil mit sich, keine größeren Änderungen am System vornehmen zu müssen und die Software in einem geschützten virtuellen Rahmen testen zu können.

Das Einrichten einer virtuellen Maschine ist daher für die meisten NutzerInnen das Fundament (und vielleicht auch der Einstieg) in Unix-basierte Entwicklerumgebungen. Dabei sind diverse kleinere Einstellungen zu beachten, vom Einschalten der Virtualisierung im BIOS bis hin zur Installation von gemeinsam genutzten Ordnern zwischen Host und Gast. Ubuntu als „Einstiegslinux“ eignet sich hervorragend für die ersten Schritte, da es eine hohe Benutzerfreundlichkeit aufweist und trotzdem alle wichtigen Features mitbringt, die benötigt werden.

Repositorien für brauchbare Digitalisate

OCR-Software erzielt bessere Ergebnisse mit hochauflösenden und fehlerfreien Digitalisaten. Bilddateien sollten mindestens eine Auflösung von 300 DPI besitzen und nicht bereits durch Programme oder Algorithmen bearbeitet worden sein. Google Books referenziert zwar auf viele gescannte Werke, diese liegen aber meist in schlechter Qualität und verlustreich binarisiert vor. Bekannte Repositorien wie das *zentrale Verzeichnis digitalisierter Drucke* oder das *Münchener DigitalisierungsZentrum* bieten exzellente Anlaufstellen zur Beschaffung digitalisierter Drucke; aber auch Sammlungen wie das *Verzeichnis der im deutschen Sprachbereich erschienen Drucke des 16. - 19. Jahrhunderts* der Universität- und Landesbibliothek Sachsen-Anhalt verfügen über frei zugängliche Digitalisate mit einer Auflösung bis zu 600 DPI.

Installation von ocropy

Ocropy ist nicht in den nativen Quellen von den bekanntesten Linux-Derivaten enthalten, sondern muss von Github heruntergeladen und über ein Script installiert werden. Dabei ist die Version der Programmiersprache Python 2.7 zu beachten und die Abhängigkeiten einiger benötigter Module. Im Workshop wird die Installation begleitet und ein bereits auf Drucke des 18. Jahrhundert trainiertes Erkennungsmodul zur Verfügung gestellt.

Preprocessing mit ScanTailor

Eine Texterkennung ist nur so gut wie das Preprocessing des Digitalisates. Bilder, Initiale oder Flecken im Bild stören die Texterkennung und müssen entfernt werden. Darüber hinaus benötigt ocropy binarisierte (schwarz/ weiß gerasterte) oder normalisierte Graustufenbilder zur Verarbeitung. Obwohl ocropy mit dem Modul ocropus-nlbin eine eigene Lösung zur Binarisierung von Bilddateien anbietet, hilft dies nicht in Bezug auf Nicht-Text-Elemente, wie Bilder oder schräge Spaltenlinien. Bearbeitungssoftware wie Gimp beinhaltet zwar alle benötigten Funktionen, ist jedoch in Bezug auf die serielle Verwendung bei Textdigitalisaten ineffizent. Im Workshop wird die Software ScanTailor als passgenaues Preprocessing-Tool zur Vorbereitung der Digitalisate favorisiert. ScanTailor ist wie dafür gemacht gescannte Texte in eine einheitliche Form zu bringen und beinhaltet (zum Teil vollständig automatisierte) Funktionen wie

- der Aufsplittung von Spalten oder Seiten
- das Ausrichten der Seite
- des Auswählens des Inhalts
- der Möglichkeit Bereich zu füllen
- der Entzerrung gekrümmter Seiten und
- der Anpassung des Schwellwertes (threshold) bei der Binarisierung.

Außerdem werden Hinweise zu den grundlegenden Eigenschaften eines guten Eingangsbildes gegeben, z. B. in Bezug auf Schwellwert oder DPI-Zahl.

Entwicklung einer Pipeline zur Texterkennung

Die ocropy-Module funktionieren am effizientesten innerhalb einer Pipeline. Ausgehend von der Konvertierung unpassender Dateiformate der Roh-Digitalisate bis hin zur Erstellung einer Korrektur-HTML für die Verbesserung der falsch erkannten Zeichen bietet die Linux-Shell zusammen mit ocropy und dem Programm ImageMagick alle benötigten Werkzeuge. So lassen sich auch große Mengen an Bilddateien stapelweise verarbeiten. In einem Script werden Befehle zur Bildkonvertierung, Zeilenauftrennung, Texterkennung und Textkonvertierung in Reihe geschaltet, um eine stapelhafte Verarbeitung zu ermöglichen. Der Workshop bietet zwei vorgefertigte Scripte zum Gebrauch an und erklärt ihren Ablauf, um eventuelle Anpassungen an die eigenen Bedürfnisse vornehmen zu können.

Training unbekannter Schriftarten

Die eigentliche Stärke von ocropy ist die Möglichkeit Erkennungsmodule für Schriftarten zu trainieren. Die dazu bereitgestellte Ground Truth Data bestimmt maßgeblich die Leistungsfähigkeit der Erkennungsmodule. Dabei stellt sich die Frage, wie eine gute Ground Truth im wörtlichen Sinne auszusehen hat? Wie „schmutzig“ dürfen die Daten sein? Sind abgeschnittene Serifen, fehlende Bögen oder i-Punkte ein Problem? Welche Zeichen sollten verwendet werden, um Abbreviationen oder Abkürzungszeichen zu kodieren? Darüber hinaus trainiert ocropy sich nicht permanent besser, sondern baut das neurale Netz zeitweise mit negativen Auswirkungen für die Erkennungsraten um (siehe Abbildung 2). Im Workshop wird ein Script zur Identifikation des besten Trainingsmoduls vorgestellt, um das Beste aus ocropy herauszuholen.

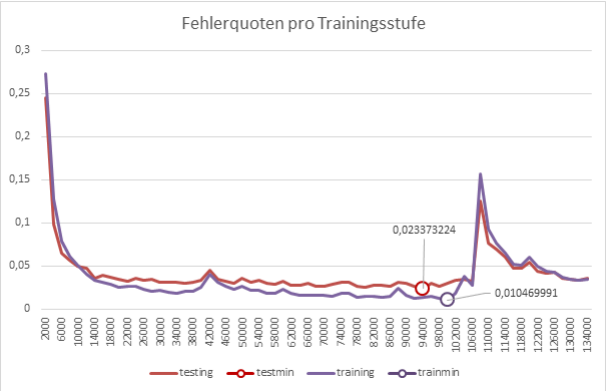

Abbildung 2. Trainingsprozess von ocropus-rtrain mit Ground Truth von Zedlers Universallexikon. training = Ground Truth anhand derer das Modul trainiert wurde, testing = unbekanne Ground Truth zum Test der Performance.

Ablauf

Der Workshop richtet sich vorrangig an Anfänger und leicht fortgeschrittene NutzerInnen im Umgang mit Linux und der Console. Es werden keine Vorkenntnisse in Bash oder Python benötigt und alle im Kurs vorgestellte Software, Scripte und Daten stehen frei zur Verfügung. Der Workshop möchte alle an Interessierten da abholen, wo sie stehen und versucht durch ein schrittweises Vorgehen an die Vorzüge der Consolen-Benutzung und kommandozeilenbasierte Software heranzuführen. Teilnehmer sollten ihr eigenes Notebook mitbringen, auf dem sie auch Administrator-Rechte besitzen. Des Weiteren wird ein Internetzugang benötigt, um fehlende Software oder Abhängigkeiten herunterladen zu können. Größere Softwarepakete (VirtualBox, Ubuntu) werden auch auf USB-Sticks zur Verfügung gestellt, sollten aber nach Möglichkeit vorher selbstständig heruntergeladen werden. Es können je nach Erfahrungsstand der TeilnehmerInnen mit Console und Linux 20 bis 25 Personen betreut werden. Der Workshop dauert drei bis vier Stunden.

Bibliographie

ImageMagick (2018): *Convert, Edit, Or Compose Bitmap Images @ ImageMagick*, URL: https://www.imagemagick.org/ , [zuletzt besucht am 14.10.2018].

MDZ (2018): *Münchner DigitalisierungsZentrum*, Bayerische Staatsbibliothek, München, URL: https://www.digitale-sammlungen.de/ , [zuletzt besucht am 12.10.2018].

ocropy (2018): *Python-based tools for document analysis and OCR*, URL: https://github.com/tmbdev/ocropy , [zuletzt besucht am 14.10.2018].

ScanTailor (2018): *ScanTailor*, http://scantailor.org/ , [zuletzt besucht am 14.10.2018].

Springman, Uwe (2015): *Ocrosis. A high accuracy OCR method to convert early printings into digital text, Center for Information and Language Processing (CIS)*, Ludwig-Maximilians-University, Munich, URL: http://cistern.cis.lmu.de/ocrocis/tutorial.pdfm [zuletzt besucht am 14.10.2018].

statista, Marktanteile der führenden Betriebssysteme in Deutschland von Januar 2009 bis Juli 2018, URL: https://de.statista.com/statistik/daten/studie/158102/umfrage/marktanteile-von-betriebssystemen-in-deutschland-seit-2009/, [zuletzt besucht am 10.10.2018].

Vanderkam, Dan (2015): *Extracting text from an image using Ocropus*, URL: http://www.danvk.org/2015/01/09/extracting-text-from-an-image-using-ocropus.html, [zuletzt besucht am 10.10.2018].

VD (2018): *Digitale Sammlungen des 16. bis 19. Jahrhunderts*, Universitäts- und Landesbibliothek Sachsen-Anhalt, Halle (Saale), URL: http://digitale.bibliothek.uni-halle.de/ , [zuletzt besucht am 14.10.2018].

VirtualBox (2018): *Oracle VM VirtualBox*, URL: https://www.virtualbox.org/ , [zuletzt besucht am 14.10.2018].

Ubuntu (2018): *The leading operating system for PCs, IoT devices, servers and the cloud | Ubuntu*, URL: https://www.ubuntu.com/ , [zuletzt besucht am 14.10.2018].

ZVDD (2018): *Zentrales Verzeichnis Digitalisierter Drucke*, Georg August Universität Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Göttingen, URL: http://www.zvdd.de/ , [zuletzt besucht am 12.10.2018].