READ
Recognition and Enrichment of Archival Documents

Vorhandene Transkriptionen zum Trainieren eines Handschriftenerkennungs-modells verwenden

Version 1.4.0
Letzte Aktualisierung dieses Guides: 08.06.2018

Dies ist eine Anleitung für Nutzer, die bereits über Transkriptionen verfügen und diese verwenden möchten, um ein Handschriftenerkennungsmodell (HTR Modell) zu trainieren. Das funktioniert mit dem neuen t2i (text2Image) Tool ganz einfach.

Laden Sie den Transkribus Expert Client herunter oder stellen Sie sicher, dass Sie die neueste Version verwenden:

- https://transkribus.eu/

Besuchen Sie das Transkribus Wiki für mehr Informationen und weitere How to Guides:

- https://transkribus.eu/wiki/

Transkribus und die zugrundeliegende Technologie werden durch die folgenden Projekten und Plattformen verfügbar gemacht:

- https://read.transkribus.eu/
- https://transcriptorium.eu/
- https://github.com/transkribus/

Kontakt

- Das Transkribus Team: email@transkribus.eu

Inhalt

Das READ Projekt wird durch das Horizon 2020 Forschungs- und Innovationsprogramm im Rahmen des Fördervertrags Nr. 674943 finanziert.

Einleitung

Die Transkribus Plattform macht es Nutzern möglich ein HTR Modell zu trainieren, das Dokumente automatisch erkennt. Das Modell wird trainiert, indem ihm Bilder von Dokumenten mit einer bestimmten Handschrift und deren exakte Transkription gezeigt werden.

In den letzten 20 Jahren wurden viele wissenschaftliche Transkriptionsprojekte durchgeführt. Einige von ihnen sind heute noch aktiv. Eine große Menge von Dokumenten wurde transkribiert und ist jetzt in elektronischer Form verfügbar. Alle diese Transkriptionen können ganz einfach als Trainingsmaterial für HTR verwendet werden.

Das t2i-Tool, das vom CITlab an der Universität Rostock entwickelt wurde, erstellt Trainingsdaten basierend auf den vorhandenen Transkriptionen. Das Tool verwendet einen Algorithmus, der automatisch die Transkriptionen mit den Bildern der handschriftlichen Dokumente verknüpft. In weiterer Folge kann damit ein HTR-Modell trainiert werden.

Anstatt Trainingsdaten für HTR manuell in Transkribus zu erstellen, können Sie also einfach ihre vorhandenen Transkriptionen verwenden, um die Technologie auszuprobieren. So können zuverlässige Transkriptionen erstellt werden, ohne dass der Workflow eines Projekts verändert werden muss.

Vorbereitungen

- Wenn Sie mit dem t2i-Tool arbeiten möchten, brauchen Sie Zugriff auf die digitalisierten Bilder und Transkripte ihrer Dokumente
- Diese Dateien müssen laut den nachfolgenden Anleitungen vorbereitet werden, bevor sie mit vom t2i-Tool verarbeitet werden können.

Seitenmenge

- Wir empfehlen Ihnen den Trainingsprozess mit mindestens 20.000 Wörtern (ca. 100 Seiten) transkribierten Materials zu starten.
- T2i ist speziell bei einer großen Menge von verfügbaren Transkripten interessant (500 oder mehr Seiten)
- Diese Technologie kann schnell große Mengen von Transkripten verarbeiten (100.000 Seiten und mehr)
- Die neuralen Netzwerke in HTR lernen schnell. Je mehr Trainingsdaten verfügbar sind, desto besser werden die Resultate.

Bilddateien

- Alle Arten von Bilddateien können verarbeitet werden.
- Die Bilder sollten eine Auflösung von mindestens 200ppi haben. Alternativ – wenn die Bilder mit einer Kamera aufgenommen wurden – gilt die Faustregel, dass ein einzelnes Zeichen mindestens 15-20 Pixel hoch sein soll,
- Die Genauigkeit der HTR ist natürlich auch von der Bildqualität abhängig. Aber auch schwer zu verarbeitende Daten von Mikrofilmen und Schwarzweißbildern können mit genug Trainingsdaten verarbeitet werden.

Transkriptionsdateien

- Alle Transkripte sollten als einfache Textdateien (TXT) gespeichert werden.

- Wenn Sie Transkriptionen im TEI (Text Encoding Initiative), Word, XML oder HTML-Format haben, sollten Sie diese in TXT-Dateien umwandeln, indem Sie zum Beispiel die Transkripte in den Editor kopieren.
- Transkriptionen sollten auf Seitenlevel gespeichert werden, das heißt eine txt-Datei pro Seite.

Transkriptionen

- Transkriptionen sollten keine unnötigen Formatierungen enthalten.
- Wenn Ihre Transkription Zeilenumbrüche enthält, können diese beibehalten werden. Es ist jedoch nicht nötig am Ende jeder Zeile einen Zeilenumbruch zu setzen.
- Das t2i-Tool meistert auch Fälle, in denen das Wort ohne Bindestrich auf zwei Zeilen aufgeteilt wurde.
- Wenn ein Wort in Ihrem Transkript unleserlich ist, ist es am besten einfach die ganze Zeile zu löschen in der das Wort vorkommt. Somit wird diese Zeile nicht für das HTR-Training verwendet.
- Transkriptionen müssen nicht komplett sein. Wenn Wörter im Transkript fehlen, werden sie nicht für das HTR-Training verwendet.
- Das Arbeiten mit verschiedenen Unicode-Zeichen, unter anderem Arabisch und Hebräisch, ist möglich
- In manchen Fällen können Transkriptionen in denen Abkürzungen ausgeschrieben wurden auch für t2i und HTR-Training verwendet werden (die Abkürzungen werden automatisch ausgeschrieben).

Dateien benennen

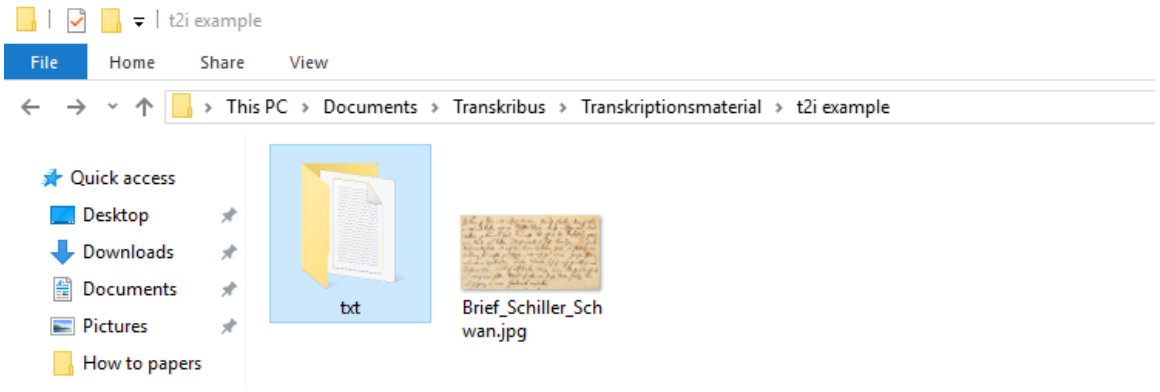
- Die Dateien die Ihre Bilder und Transkriptionen beinhalten sollten klar zuordenbar sein
- Dafür speichern Sie am besten jede Bilddatei mit dem exakt selben Namen wie die zugehörige TXT-Datei.

Hilfe bei der Dateivorbereitung

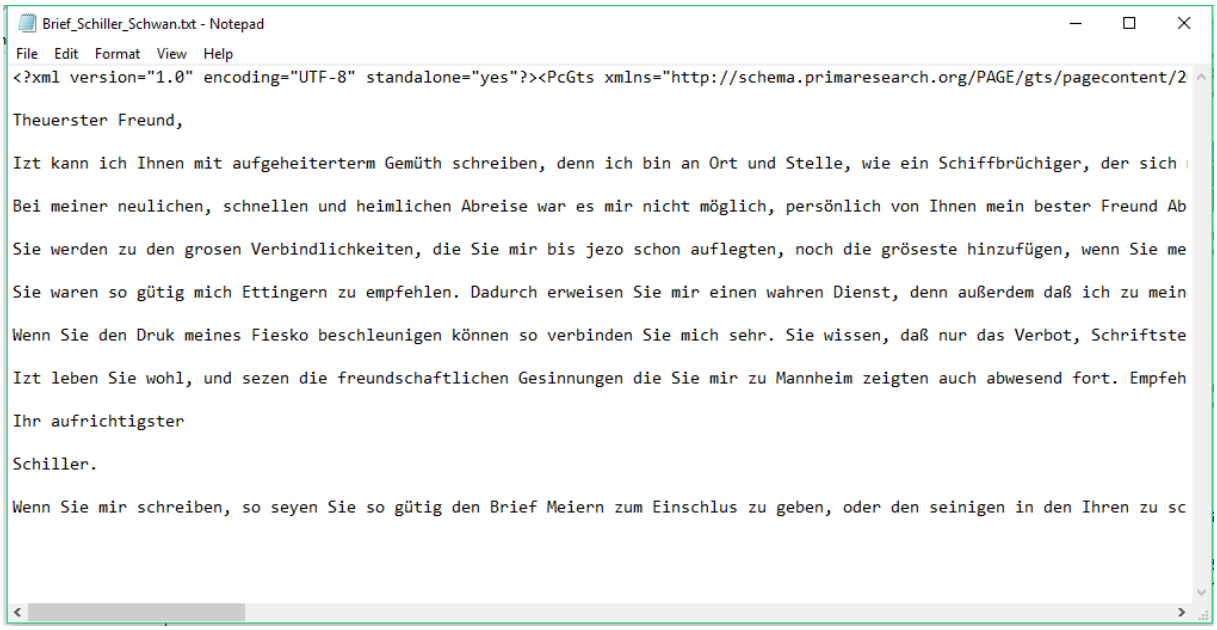
- Wenn Sie nicht die Resourcen haben um Ihre Dateien wie beschrieben vorzubereiten, kann Ihnen das Transkribus Team helfen. Schreiben Sie uns einfach eine Email (email@transkribus.eu).

Lieferung der Dateien

- Wenn Sie die Bilder und Transkripte vorbereitet haben, müssen Sie sie richtig strukturieren:
    - Name des Dokuments
        - TXT
- Sie können Ihre Dateien direkt auf Transkribus hochladen. Für den Upload der TXT-Dateien sollte ein separater Ordner “Text” im Ordner mit den Bilddateien vorhanden sein.



Darstellung 1 Aufteilung der Ordner



Darstellung 2 TXT Dateil

- Alternativ können Sie Ihre Dateien auch an das Transkribus Team (email@transkribus.eu) übermitteln, indem Sie ein File-Sharing-System wie zum Beispiel WeTransfer verwenden.
- Senden Sie uns in beiden Fällen eine Email (email@transkribus.eu) damit wir wissen, dass Ihre Dateien bereit sind.
- Wenn Sie schon mit uns in Kontakt waren und mehr Informationen darüber benötigen, wie Sie Modelle trainieren können, dann finden Sie diese in Modell Training in Transkribus.

Danksagung

Wir möchten den vielen Nutzern danken, die mit ihrem Feedback zur Verbesserung der Transkribussoftware beigetragen haben.

Transkribus wird der Öffentlichkeit im Rahmen des H2020e Infrastrukturprojekts READ (Recognition and Enrichment of Archival Documents) zugänglich gemacht. Dieses Projekt wird von der Europäischen Kommission im Rahmen des Fördervertrags Nr. 674943 finanziert