OCR-BW

# Tesseract – Volltexte für die Danziger Volksstimme


Danziger Volksstimme

Die „Danziger Volksstimme" erscheint täglich mit Ausnahme der Sonn- und Feiertage. – Bezugspreise: In Danzig bei freier Zustellung ins Haus monatlich 2,60 Mk., vierteljährlich 7,80 Mk. – Postbezug außerdem monatlich 30 Pfg Zustellungsgebühr.

Organ der Sozialistischen Partei der Freien Stadt Danzig

Stefan Weil
Universitätsbibliothek Mannheim

09.06.2021

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

# Volltexte für die Danziger Volksstimme

… als Anwendungsbeispiel für nachträgliche OCR.

Ausgangssituation:

Ein Archiv / eine Bibliothek hat eine Zeitung /
ein Druckwerk digitalisiert als PDF ohne Volltext.

Die Nutzer wünschen durchsuchbaren Volltext.

# Volltexte für die Danziger Volksstimme

Hier:

Anfrage auf der Mailingliste InetBib, weil die Digitalausgabe der Zeitung Danziger Volksstimme (1920–1936) keine durchsuchbaren Texte enthält.

Hintergrund: die von der Friedrich-Ebert-Stiftung veröffentlichten PDF-Dateien sind schon ältere Digitalisierungen aus einer Mikrofilmausgabe.

# Volltexte für die Danziger Volksstimme

Herausforderungen:

Bildqualität und Auflösung sind grenzwertig für OCR. Frakturschrift ist viel schwerer lesbar als moderne Schrift.

Wie weit kommt man trotzdem mit relativ einfachen Werkzeugen (Standardsoftware)?


Danziger Volksstimme

Organ der Sozialistischen Partei der Freien Stadt Danzig

Nr. 1 Freitag, den 2. Januar 1920 11. Jahrgang

Unser Ziel und unser Weg!

Zum Geleit!

Die Unterzeichnung des Friedensprotokolls.

# Volltexte für die Danziger Volksstimme

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Werkzeugkasten:

- PC mit Linux, MacOS oder Windows 10
- OCRmyPDF (verwendet Tesseract OCR)
- Tesseract-Modelle für die Erkennung von Fraktur

# Volltexte für die Danziger Volksstimme

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

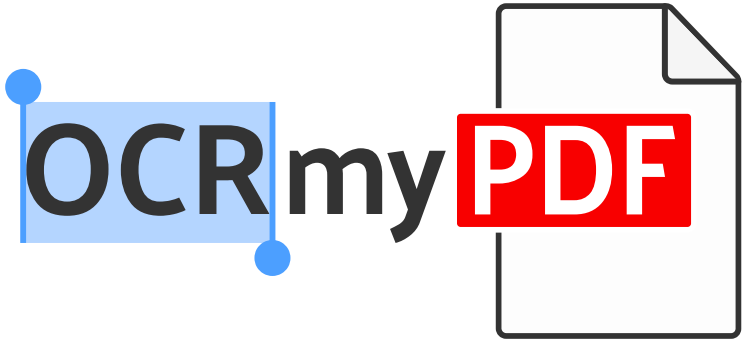
Ablaufplan:

Vorbereitung

- Installation: OCRmyPDF und Tesseract-Modelle

OCR

- Download der PDF-Dateien
- Aufruf von OCRmyPDF



# Volltexte für die Danziger Volksstimme

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Installation OCRmyPDF:

- Installationspaket ocrmypdf für Linux
  (alle gängigen Distributionen)
- Windows 10 mit Windows Subsystem for Linux
- MacOS mit Homebrew
- https://github.com/jbarlow83/OCRmyPDF#installation

# Volltexte für die Danziger Volksstimme

Installation Tesseract-Modelle für Fraktur:

- Standardmodelle frk und script/Fraktur von https://github.com/tesseract-ocr/tessdata_fast
- Frakturmodelle der UB Mannheim von https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/ (Fraktur_5000000 oder frak2021)

# Volltexte für die Danziger Volksstimme

Planänderung:

Statt der schon bearbeiteten Danziger Volksstimme (1920–1936) machen wir Texterkennung für deren Vorgängerzeitung Volkswacht Danzig (1912–1919).


Volkswacht

Beilage: Die Neue Welt, illustriertes Unterhaltungsblatt

Organ für die werktätige Bevölkerung der Provinz Westpreußen

Publikationsorgan der Freien Gewerkschaften

Nr. 1. Danzig, den 3. Januar 1912. 3. Jahrgang.

Warum ich schon in der Hauptwahl sozialdemokratisch wähle?

Politische Übersicht.

Rüstungswahnsinn.

# Volltexte für die Volkswacht Danzig

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Download der PDF-Dateien:

- Einheitliches Muster für URL: http://library.fes.de/danzig/pdf/1919/1919-099.pdf
- Download mit folgendem Kommando (dauert rund 5 Minuten)

```
time for jahr in $(seq 1912 1919); do (mkdir -p
$jahr; cd $jahr; for ausgabe in $(seq -w 1 330); do
wget http://library.fes.de/danzig/pdf/$jahr/$jahr-
$ausgabe.pdf; done); done
```

# Volltexte für die Volkswacht Danzig

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

## Erzeugung der Volltexte (txt und pdf):

## Die Erzeugung erfolgt mit diesen Kommandos:

```
mkdir -p ocr

for pdf in 19??/*.pdf; do
  name=$(basename $pdf .pdf);
  echo $pdf &&
  time ocrmypdf -l frak2021 --sidecar ocr/$name.txt -q $pdf ocr/$name.pdf;
done
```

# Volltexte für die Volkswacht Danzig

## Ergebnis (Auszug Seite 1 vom 3. Januar 1912):

Nr. 1.

Wrum ich ſchon in der Hauptwahl

ſozinldemokratiſch wähle?

Sechzehn Antworten.

ů 1. 5

Im Verlage Fortſchritt iſt eine Flugſchrift heraus≠

gegeben worden, die unter dem Titel: Warum ich in der

Hauptwahl nicht ſozialdemohratiſch wähle, ſechzehn

freigewählte Gründe, die Arbeiter vor der Sozialdbemokentie

# Volltexte für die Danziger Volksstimme


Optionale Anzeige als Flip Book mit Suchfunktion

https://github.com/iberan/pdfjs-flipbook

# Volltexte für die Danziger Volksstimme

Alternative für einzelne PDF-Dateien:

Zotero Plugin für OCR

https://github.com/UB-Mannheim/zotero-ocr

Erweiterung für die Literaturverwaltungssoftware Zotero, die mit Hilfe von Tesseract PDF-Dateien um fehlende Volltexte erweitert

# Volltexte für die Danziger Volksstimme

Abbildungen / Links:

Digitale Bibliothek der Friedrich-Ebert-Stiftung

Volkswacht Danzig – http://library.fes.de/inhalt/digital/danziger-volkswacht.htm

Danziger Volksstimme – http://library.fes.de/inhalt/digital/danziger-volksstimme.htm