Digitalisierung am *Dictionnaire de l'occitan médiéval (DOM) (work in progress)*

Matthias Schöffel

Bayerische Akademie der Wissenschaften
Matthias.Schoeffel@dom.badw.de

München, den 13.12.2021

Inhaltsverzeichnis

1. Allgemeine Einführung
2. Arbeitsschritte - Aufbau bei der Digitalisierung
   - Vorsortierung der Zettel
   - OCR-Analyse
3. Zusammenfassung/Ausblick

Ausgangssituation

Abbildung: Kasten

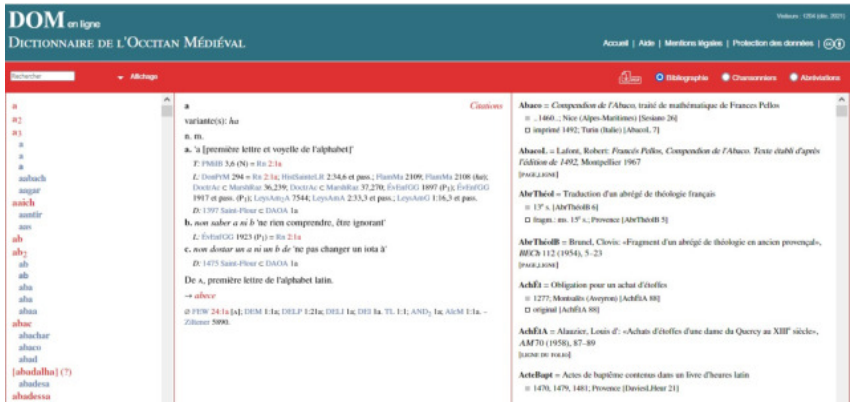

Abbildung: www.dom-en-ligne.de

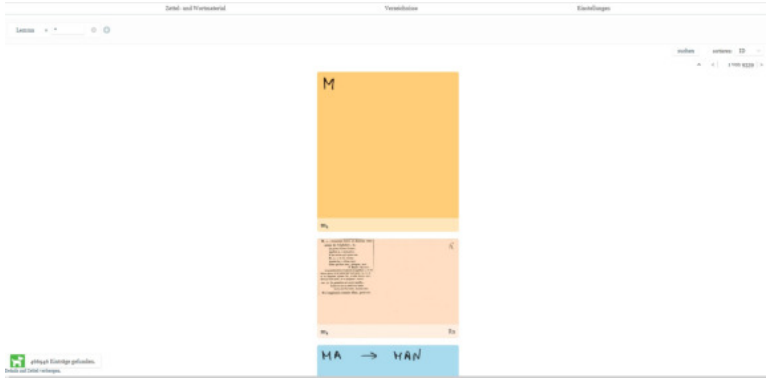

Abbildung: Datenbank: dDOM

Abbildung: Lemma

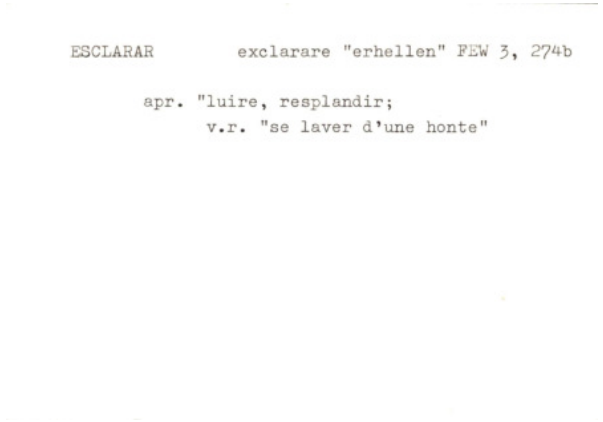

Abbildung: FEW-Karte

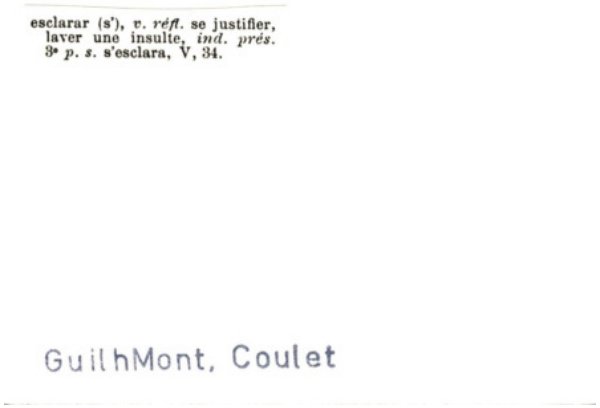

Abbildung: Beleg

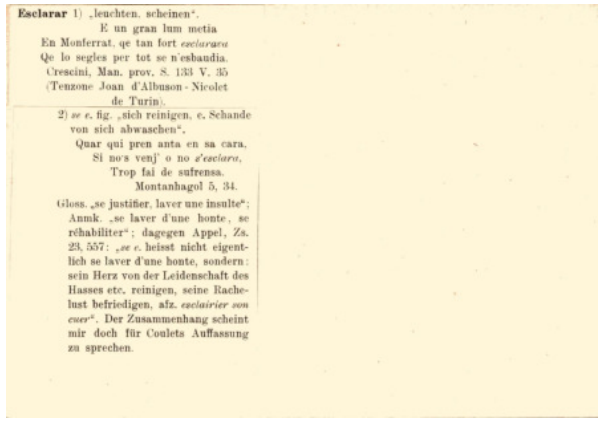

Abbildung: Rn/Lv

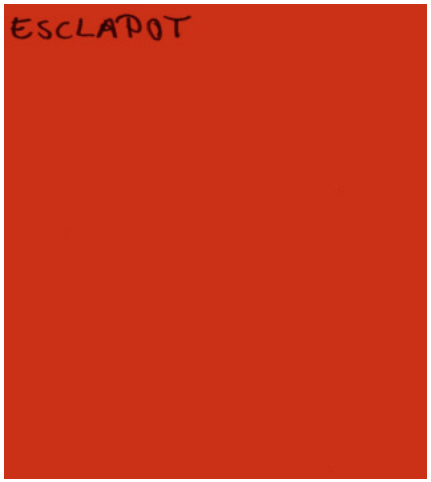

Abbildung: Lemma

Abbildung: “Scharnier”

Abbildung: Mot nouveau

Abbildung: Verweis: Schreibvariante - Lemma

Arbeitsschritte - allgemein

Ablauf

- Scan der ca. 600.000 Zettel (Arbeitsmaterial)
- Zeitvorgabe: 1 Jahr
- Einarbeitung in eine angepasste Datenbank
  (Basis: DB des Mittellateinischen Wörterbuchs (A. Häberlin))
  1. Zettelvorsortierung durch machinelles Lernen anhand der Funktion der Zettel (aktuell: Verweiszettel)
  2. OCR-Scan der Verweiszettel
     (Umfang: ca. 25% aller Zettel)
  3. Überprüfung der OCR-erkannten (Ziel-)Lemmata mit DOM-interner-Liste
- Ziel:
  1. Arbeitsmaterial zu systematisieren
  2. Konsistenzprüfung des (Arbeits-)Materials ($\rightarrow$ Hierarisierung der Daten)

Maschinelles Lernen - Vorsortierung (Schritt 1)

Warum maschinelles Lernen? - Vorteile

- Binäre Angabe notwendig: Verweis vs. Nicht-Verweis
- Kasten: Klassifizierung der Zettel (one-hot-encoding)
- Hervorragende Ergebnisse (100%)
- Keine weitere Informationen notwendig

Aufbau

1. Auslesen der RGB-Farbwerte mit OpenCV-Bibliothek → Clustering (K-Means)
2. GradientBoostingClassifier → GridSearchCV für Hyperparameter → Bestes Model für die Vorhersage
3. Ergebnis (Anwendung, Vorhersage): 55s für 1800 Zettel (8 logische Prozessoren, multiprocessing)

↪ Genauigkeit: 100% (seit 4 Monaten im Einsatz)

Ausgangspunkt - OCR (Schritt 2)

Verwendete Umgebung

- Verwendung von tesseract (5.0.0-alpha.20210506)
- Sprache: oci (neu-)okzitanisch

Tesseract - Image processing

Tessaract: "Tesseract does various image processing operations internally (using the Leptonica library) before doing the actual OCR. It generally does a very good job of this, but there will inevitably be cases where it isn't good enough, which can result in a significant reduction in accuracy." (Dokumentation, Github, Aufruf: 1.12.2021)

Image preprocesing (Tesseract)

- Skalierung
- Binarisierung (Tests mit Otsu (veraltet))
- Rauschentfernung
- Dilation/Erosion
- Transparenz
- Helligkeit

→ Ziel: Optimale Kombination aus diesen Bildbearbeitungsmethoden

Lösungsansatz

1. Maschinelles Lernen
   ↓
   Aktueller Ansatz
1. Transformation des Histogramms
   ↓
   Optimale Farbverteilung (offen)

OCR-Erkennung ohne Bearbeitung

OCR_Text	dist	Text
mate ats	7	ma -> mais
h a_s mal	8	ma -> mal
mar > " mhnan	8	maa -> man
haarh ar	7	maar -> mar
haas->3 mhais	7	maas -> mais
habable -> novable	4	mabable -> movable
habedor => movedoir	5	mabedor -> movedor
maqer -> hover	4	maber -> mover
mabre -> hôharme	5	mabre -> marme
mabrin => marbrin	3	mabrin -> marbrin
maça => wuasa	6	maca -> masa
macel -> mwhazel	4	macel -> mazel
macelier > hazelier	4	macelier -> mazelier
macellier => mazelier	3	macellier -> mazelier
maccceria -> maze€eria	6	maceria -> mazeria
hach => mag	4	mach -> mag
*hacha => whercatlar. 222 hercha	24	macha -> mercat
machacolladura ->3machacoladura	4	machacolladura -> machacoladura
machacoulis =>machacoladipr	8	machacoulis -> machacolis
machadura => macadupa	4	machadura -> macadura

Konstante Helligkeit - Faktor 150 - Ergebnisse

OCR_Text	dist	Text
ma = =3 has	8	ma -> mais
halha	8	ma -> mal
mar => man	4	maa -> man
haar=>h rr	11	maar -> mar
hmaas -3 mais	4	maas -> mais
habable -> movarle	4	mabable -> movable
hmabedoir => movedoir	6	mabedor -> movedor
magbeter -> mover	5	maber -> mover
mabre -> vwarme	4	mabre -> marme
marrin => marbrin	4	mabrin -> marbrin
maça = => hasa	7	maca -> masa
macel -> whazel	4	macel -> mazel
macelier => hazelier	4	macelier -> mazelier
macellier => mazelier	3	macellier -> mazelier
maccceriéa -> maz€rila	7	maceria -> mazeria
h ach-3	9	mach -> mag
*"hacua -=> whercat	9	macha -> mercat
machacolladura =>machacoladura	4	machacolladura -> machacoladura
hachacoulis =>machacoladifz	9	machacoulis -> machacolis
machadura => macadubpa	5	machadura -> macadura

Weiß-Schwarz: Graustufe und Threshold-Entfernung (otsu)

Text (schwarz)- Hintergrund: weiß

OCR_Text	dist	Text
endecrepitat =>decrepitat	4	endecrepitat -> decrepitat
endeclrepitut >decrep i/itut	8	endecrepitut -> decrepitut
endecs > endeç	4	endecs -> endec
bebndedvdvde -35 endedia	10	endedie -> endedia
endeficameint-)`edificamen	6	endeficament -> edificamen
ende ficansaiedificans a	8	endeficansa -> edificansa
endetficar ->tedificaq	6	endeficar -> edificar
endegesno >> indigestion	6	endegestio -> indigestio
endega 3 endenhar	5	endegna -> endenhar
endegnanza ->endenhans a	4	endegnanza -> endenhansa
endegoï > e nde jotz	7	endegot -> endejotz
e [ndeguar > e ndeg ar	7	endeguar -> endegar
endem a > endeman	4	endema -> endeman
e ndemain > endeman	4	endemain -> endeman
endemang -> endehman	3	endemang -> endeman
endematiuu >d ndematin	7	endematii -> endematin
e /vdemayn e nde man	9	endemayn -> endeman
endementres ->dementre	3	endementres -> dementre
depeis => despens	3	depeis -> despens
depeis -> depenher	2	depeis -> depenher

Fokus: Optimale Helligkeit und Skalierung (Random Search)

brightness	scale	real	prediction	adist
150	350	ebetud -> ebetut	ebetud => ebetut	2
105	330	bacha -> bas	bacha => bas	2
150	255	datz -> dat	datz > dat	2
130	315	darres -> darre	darres > darre	2
125	165	eginiar -> engenhar	eginiar > engenhar	2
60	200	bariou -> berrion	bariou => berrion	2
120	230	baratrer -> baratier	baratrer => baratier	2
75	225	alcaot -> alcavot	alcaot => alcavot	2
80	275	alegragge -> alegratge	alegragge ->alegratge	2
150	310	elegit -> elegir	elegit -3 elegir	2
70	180	alet -> alen	alet -> alen	1
135	335	bailie -> bailia	bailie => bailia	2
170	340	barde -> barda	barde > barda	2
140	320	efrei -> esfre	efrei -3 esfre	2
95	115	davalar -> devalar	davalar -> devalar	1
60	135	alcia -> ausar	alcia => ausar	2
120	315	egance -> egansa	egance > egansa	2
110	180	alcaot -> alcavot	alcaot => alcavot	2
135	290	datial -> datil	datial +> datil	2
120	305	eisercir -> exercir	eisercir => exercir	2
110	255	balan -> balar	balan -> balar	1
150	205	alete -> aleta	alete => aleta	2
90	220	d'arap -> arap	d'arap > arap	2
100	190	baleste -> balesta	baleste => balesta	2

Methodische Überlegungen (Python-Keras/scikit-learn)

- Mittelwerte, RGB-Werte, Max-, Min-Werte → Wenig Varianz → Schwierige Klassifikation
- Deep-Learning als Feature-Extraktor mit PCA-Analyse/RandomForest
- Transfer Learning
- Support-Vector-Machine
- OneVsRest-Classifier
- (Dynamic-Time-Warping (DTW) → Transformation)
- Image Segmentation

Vorgehen - Maschinelles Lernen

Maschinelles Lernen - Vorgehen - Bsp: CNN (TensorFlow/Keras)

- Zettel mit Helligkeitslabel (8871 Zettel) → multiclass-Problem
- One hot encoding
- Laden der Bilder mit Normalisierung und Standardgröße
- train_test_split (vgl. scikit-learn-Bibliothek)
- (wahlweise) ImageDataGenerator (Grenze: Arbeitsspeicher)
- Softmax für multiclass-Problem
- Loss-Funktion: categorical_crossentropy → Minimierung
- Berücksichtigung: balanced classes
- Fit und Ausgabe der Loss und Accuracy Kurven

Ergebnisse (Compute Cloud)

Allgemeines

- Sättigungswert für Genauigkeit
- Hohe Einheitlichkeit der Zettel → Schwierigkeit: Feature zu extrahieren
- Reduzierung der aktuellen Klassenzahl durch Helligkeitsbereiche

Übersicht

	Genauigkeit
SVC()	41,2%
OneVsRestClassifier(SVC())	(33-42)%
Transfer learning	41,1%
CNN	62%
CNN (Klassenreduzierung)	$> 89\%$

Ausblick

- Transformation (Histogramm) auf Normhelligkeit (vielversprechend)→ GANs (Generative Adversarial Networks) (vgl. Superresolution)
- Übergang von Multiclass Classification → Multilabel → Intervalle
- Größere Zettelzahl (aktuell bis max. 60.000) → mehr Arbeitsspeicher notwendig ($>$ 45GB)
- Ausbau: Integration in den Upload-Prozess zur Datenbank dDOM
- Training: rechen- und zeitintensiv
- Fazit: Kombination aus mehreren Methoden (Transformationsansatz und “Intervall”-Ansatz)

Verwendete Software

- Python-Bibliothek: Adist, Matplotlib, Multiprocessing, NumPy, OpenCV, Pandas, scikit-learn, SciPy
- TensorFlow/Keras
- Tesseract (5.0.0-alpha.20210506)