Eine NLP-Infrastruktur für KI-skeptische User

Research Lounge – Netzwerken zum Thema KI, HU Berlin, 18.6.2024

Dr. Andrea Beyer & Florian Kotschka (Humboldt-Universität zu Berlin)

daidalos-projekt.de & Zenodo-Community Daidalos

dAIdalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Das interdisziplinäre DFG-Projekt Daidalos …

… entwickelt eine webbasierte NLP-Infrastruktur,

… die es Forschenden aus der Klassischen Philologie (Latein, Altgriechisch) und angrenzenden Disziplinen ermöglicht,

… verschiedene NLP-Methoden auf selbst zusammengestellte Korpora im Kontext von sprach- oder literaturwissenschaftlichen Forschungsvorhaben anzuwenden.

dAIdalos — Digital Research for All —
Gefördert durch DFG Deutsche Forschungsgemeinschaft

Prototyp

Herausforderungen

Kleine, asynchrone, literarische Korpora

Aufwendige Modellierung vs. Mehrwert

Fachtradition: *close* vs. *distant reading*

Forschungskompetenz kaum digital

Kaum Lehr-/Lernmaterialien

dAIdalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

Lösungsansätze

Forschungstandems

User-Centered Design:
zielgruppenspezifische GUI &
Bedarfe, z. B. NER oder NLP-Pipelines
(NER + Sentiment-Analysis)

Lern- und Supportangebote

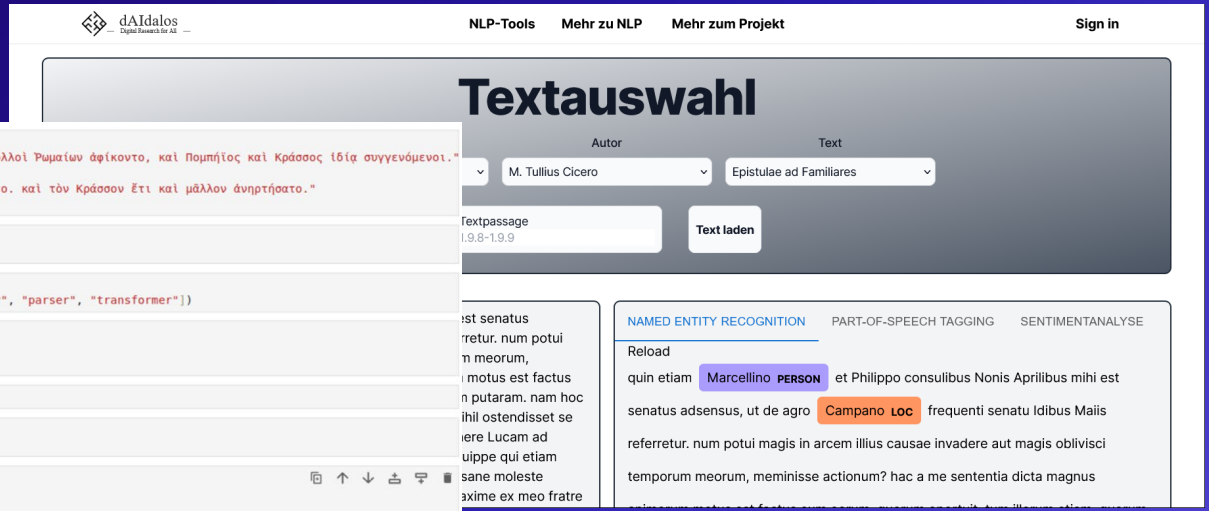
Kuratierte Workflows in Jupyter
Notebooks, Übungen mit H5P,
Literatur- und Tooldatenbank,
Community-Workshops & Beratung

Interpretable AI

Transparenz & Nachnutzbarkeit
durch Model Cards, Data Sheets,
dokumentierte Evaluation der
Methoden

dAIdalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft



```
[1]: # Auszug aus Plut. Crass. 14,5
text_with_luca: str = "Καίσαρος γὰρ εἰς Λοῦκαν πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων ἀφίκοντο, καὶ Πομπήϊος καὶ Κράσσος ἰδίᾳ συγγενόμενοι."
# Auszug aus Cass. Dio 26,3
text_no_luca: str = "τοιούτοις λογισμοῖς ὁ Πομπήϊος ἐπὶ τὸν Καίσαρα ὠπλίζετο. καὶ τὸν Κράσσον ἔτι καὶ μᾶλλον ἀνηρτήσατο."
all_texts: list = [text_with_luca, text_no_luca]

[ ]: !pip install spacy
!python -m grecy install grc_proiel_trf

[ ]: import spacy
lemmatizer = spacy.load("grc_proiel_trf", exclude=["morphologizer", "tagger", "parser", "transformer"])

[ ]: lemmatized_texts: list = [lemmatizer(text) for text in all_texts]
print(" ".join([token.lemma_ for token in lemmatized_texts[0]]))
print(" ".join([token.lemma_ for token in lemmatized_texts[1]]))

[ ]: !pip install flair

[ ]: from flair.models import SequenceTagger
tagger: SequenceTagger = SequenceTagger.load("UGARIT/flair_grc_bert_ner")

[ ]: from flair.data import Sentence
sentences: list = [Sentence(text) for text in all_texts]
for sentence in sentences:
    print(sentence)
    tagger.predict(sentence)
    for entity in sentence.get_spans('ner'):
        print(entity)

[8]: from flair.visual.ner_html import render_ner_html
from IPython.display import display, HTML
for sentence in sentences:
    html: str = render_ner_html(sentence)
    display(HTML(html))
```

Καίσαρος PER γὰρ εἰς Λοῦκαν LOC πόλιν καταβάντος ἄλλοι τε πολλοὶ Ῥωμαίων MISC ἀφίκοντο, καὶ Πομπήϊος PER καὶ Κράσσος PER ἰδίᾳ συγγενόμενοι.

τοιούτοις λογισμοῖς ὁ Πομπήϊος PER ἐπὶ τὸν Καίσαρα ὠπλίζετο. καὶ τὸν Κράσσον PER ἔτι καὶ μᾶλλον ἀνηρτήσατο.

Lösungsansatz: Forschungstandems

dAIdalos — Digital Research for All —

Gefördert durch DFG Deutsche Forschungsgemeinschaft

dAIdalos
Digital Research for All

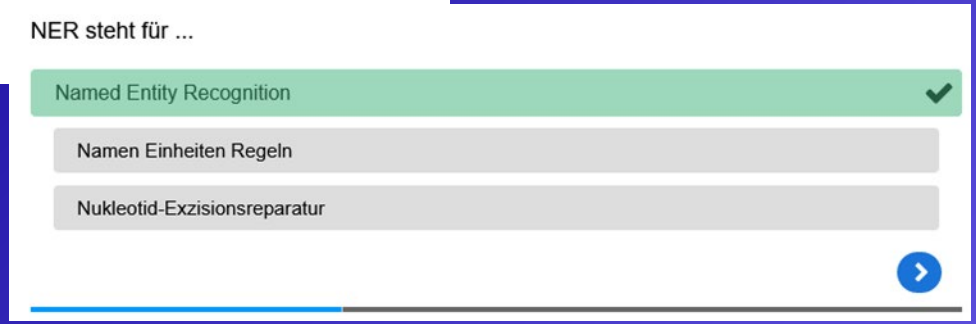
03 Word Embeddings (Demo-Workflow)

Daidalos 2024 (https://daidalos-projekt.de)

Lernziel

- NLP-Methode: Word Embeddings = Überführung von Wörtern in Vektoren
- Die Texte müssen vorher aufbereitet werden, z.B. Entfernung der Satzzeichen, Lemmatisierung des Textes.
- Die Berechnung der Kosinusähnlichkeit zweier Vektoren ermöglicht Aussagen zur Beziehung der durch die Vektoren dargestellten Wörter. In diesem Beispiel liegen die Ergebniswerte zwischen 0 und 1: Je näher der Wert an der 1 ist, desto häufiger tauchen die Worte im gleichen Kontext auf.

Einführung



Lösungsansatz: Lern- und Supportangebote

dAIdalos — Digital Research for All
Gefördert durch DFG Deutsche Forschungsgemeinschaft

la_core_web_lg

- **Person or organization developing model**: Patrick J. Burns; with Nora Bernhardt [ner], Tim Geelhaar [tagger, morphologizer, parser, ner], Vincent Koch [ner]
- **Model date**: May 2023
- **Model version: 3.7.4**
- Model type: spaCy
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: For information on the training workflow see p.4-5 of LatinCy: Synthetic Trained Pipelines for Latin NLP (https://arxiv.org/pdf/2305.04365v1)
- Paper or other resource for more information: **Burns, P.J. 2023. "LatinCy: Synthetic Trained Pipelines for Latin NLP." arXiv:2305.04365 [cs.CL]. http://arxiv.org/abs/2305.04365.
- License: *MIT*
- Where to send questions or comments about the model: https://diyclassics.github.io/

Intended Use

- Primary intended uses: Morphological analysis, POS-Tagging, Lemmatizing, Parsing
- Primary intended users: Classical Scholars
- Out-of-scope use cases: unknown

Data, Limitations, and Recommendations

- Data selection for training: Training data consists of latin UD-Treebanks, Wikipedia CC-100 Latin dataset and the Herodotos Project NER dataset
- Data selection for evaluation: Evaluation was done according to the spaCy workflow meta.json file found in the repository (https://huggingface.co/latincy/la_core_web_
- Limitations: unknown

Datasheet: Herodotos Project Dataset

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- created for Herodotos Project to train NER-Tagger (BiLSTM CRF; see: Alexander Erdmann, David Joseph Wrisley, Benjamin Allen, Christopher Brown, Sophie Cohen Bodénès, Micha Elsner, Yukun Feng, Brian Joseph, Béatrice Joyeaux-Prunel and Marie-Catherine de Marneffe. 2019. "Practical, Efficient, and Customizable Active Learning for Named Entity Recognition in the Digital Humanities." In Proceedings of North American Association of Computational Linguistics (NAACL 2019). Minneapolis, Minnesota.)
- Goal of Herodotos Project: catalogue and compendium of ancient ethnic groups
- For more info on the corpus see: https://aclanthology.org/W16-4012.pdf

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

- from the documentation: „The data files in the **Annotation** directory were annotated for named entities by a team of Classics experts at Ohio State University. Texts presently included are excerpts from Caesar's Wars, both Gallic (GW) and Civil (CW), the Plinies' writings, both Elder and Younger, and Ovid's Ars Amatoria. "

Lösungsansatz: Interpretierbare KI