Named Entity Recognition für Geisteswissenschaftler:innen

7. August 2023, SUB Hamburg

**Downloads vorab**

Stanford Named Entity Recognizer Version 4.2.0

<https://nlp.stanford.edu/software/CRF-NER.html#Download>

Texteditor: BBEdit oder Notepad ++

Aktuell Java-Version: <https://www.java.com/de/download/manual.jsp>

**Classifier zum Download**

Stanforder Modell für deutschsprachige Texte:<https://www.dropbox.com/s/mfnj349ezc1y8x1/german.conll.germeval2014.hgc_175m_600.crf.ser.gz?dl=0>

Dehmel digital-Modell: <https://www.fdr.uni-hamburg.de/record/10830>

Gender-Classifier: <https://zenodo.org/record/5555952>

**Weitere Tutorials**

Lerneinheit NER: <https://fortext.net/routinen/methoden/named-entity-recognition-ner>

Lernvideo NER: https://fortext.net/ressourcen/videos/tutorials/named-entity-recognition-und-literaturanalyse

Weitere Informationen zum Modelltraining: <https://nlp.stanford.edu/software/crf-faq.html>

**StanfordNER Benutzeroberfläche öffnen:**

Ordner „stanford-ner-2020-11-17“ öffnen 🡪 Doppelklick auf Datei mit der Endung „.jar“ (Kaffeetassen-Symbol) 🡪 Benutzeroberfläche öffnet sich

**Bestehenden Classifier hochladen und nachnutzen (z.B. den Gender-Classifier):**

Auf Benutzeroberfläche des StanfordNER-Recognizers klicken 🡪 oben in der Bedienungsleiste auf „Classifier“ klicken -🡪 auf „Load CRF from File“ klicken -🡪 in der eigenen Ordnerstruktur den heruntergeladenen Classifier auswählen (zum Beispiel: „Gender\_Modell320000\_mitListe\_mitDrama18-20\_ner-model.ser“) und per Klick auf „Öffnen“ diese Auswahl bestätigen 🡪 rechts im Interface öffnen sich die Kategorien, die automatisch annotiert werden

**Einen Text hochladen und annotieren lassen:**

1. Beispieltext im Textfeld löschen 🡪 gewünschten Text (txt-Format) in das Textfeld kopieren
2. In der Bedienungsleiste auf „File“ klicken 🡪 „Open File“ auswählen 🡪 in der eignen Ordnerstruktur den gewünschten Text auswählen 🡪 per Klick auf „Öffnen“ die Auswahl bestätigen

🡪 Klick auf „Run NER“-Button, um die automatische Annotation zu starten

**Annotierten Text herunterladen:**

Auf Benutzeroberfläche des StanfordNER-Recognizers klicken 🡪 in der Bedienungsleiste „File“ anklicken 🡪 „Save Tagged File As“ auswählen 🡪 Datei benennen (evt. die Endung „.xml“ hinzufügen) und einen Speicherort festlegen 🡪 mit Klick auf „Speichern“ bestätigen

**Datengrundlage**

Testdaten

Trainingsdaten

Kernkorpus

Properties-Datei:

* Pfad zu den Trainingsdaten („train file =“): ner\_meine\_trainingsdaten.tsv
* Möglicher Name für Ihr Modell („serializeTo =“): mein\_classifier.ser.gz

**Eingabebefehle (über das Terminal)**

1. **Navigation via Command Line zum *Stanford Named Entity Recognizer***

cd desktop/stanford-ner-2020-11-17

1. **Tokenisierung und TSV-Datei erstellen (am Beispiel der Briefkorrespondenz Behrens/Dehmel)**

java -cp stanford-ner.jar edu.stanford.nlp.process.PTBTokenizer Behrens\_Dehmels.txt > Behrens\_Dehmels.tok

perl -ne 'chomp; print "$\_\tO\n"' Behrens\_Dehmels.tok > Behrens\_Dehmels.tsv

1. **Modellerstellung**

java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -prop mein\_classifier.prop

1. **Test des Modells**

java -cp stanford-ner.jar edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier mein\_classifier.ser.gz -testFile ner\_meine\_testtexte.tsv