

Named Entity Recognition (NER)

Katharina Stein

Inhalt

- Named Entity Recognition
 - Was ist Named Entity Recognition?
 - Bedeutung für Natural Language Processing
 - Herausforderungen
 - Entity Resolution
- Tool für Named Entity Recognition: Stanford Named Entity Recognizer
 - Funktionsweise
 - Verwendung
- Übung

Was ist Named Entity Recognition?

- Teilaufgabe der Informationsextraktion aus Texten
- NER ist auch bekannt unter den Namen:
 - entity identification
 - entity extraction
- Named Entity:
 - Alles, worauf sich ein Eigenname bezieht
 - Im weiteren Sinne z.B. auch Zeiten, numerische Daten

The wedding of Prince William, Duke of Cambridge, and Catherine Middleton took place on 29 April 2011 at Westminster Abbey in London, United Kingdom.

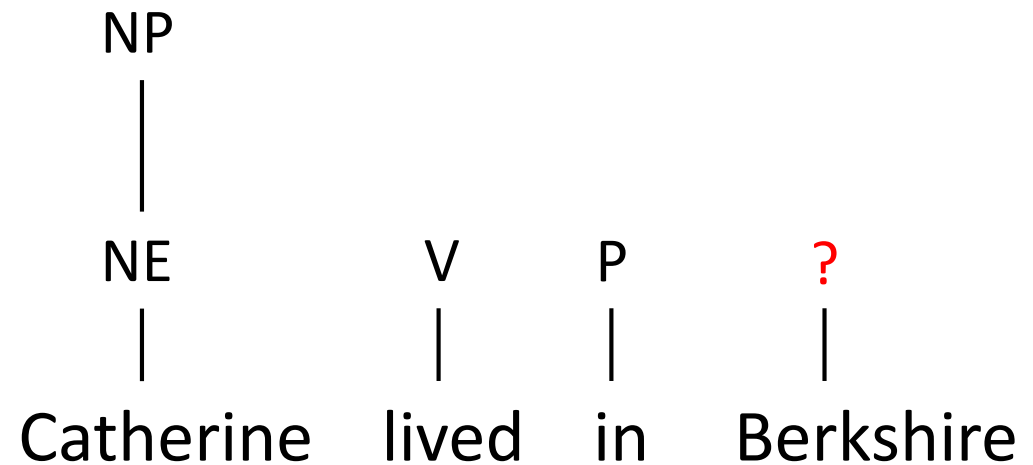
The groom, Prince William, Duke of Cambridge, is the eldest son of Charles, Prince of Wales and Diana, Princess of Wales, and second, after his father, in line to succeed his paternal grandmother, Queen Elizabeth II.

The wedding of Prince William, Duke of Cambridge, and Catherine Middleton took place on 29 April 2011 at Westminster Abbey in London, United Kingdom.

The groom, Prince William, Duke of Cambridge, is the eldest son of Charles, Prince of Wales and Diana, Princess of Wales, and second, after his father, in line to succeed his paternal grandmother, Queen Elizabeth II.

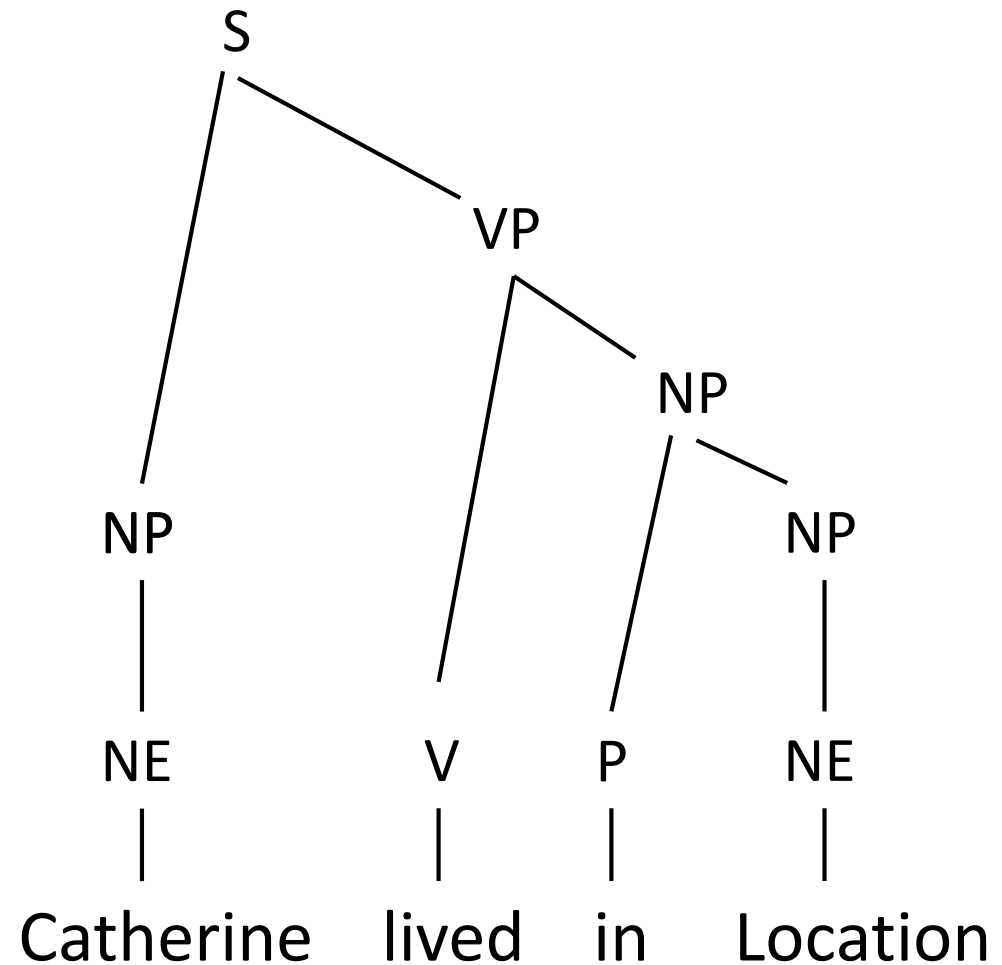
Parsing

- Catherine lived in Berkshire.
- Grammatik:
 - $S \rightarrow NP VP$
 - $NP \rightarrow NE$
 - $VP \rightarrow V PP$
 - $PP \rightarrow P NP$
- Lexikon oder Tagging:
 - Catherine: NE
 - lived: V
 - in: P



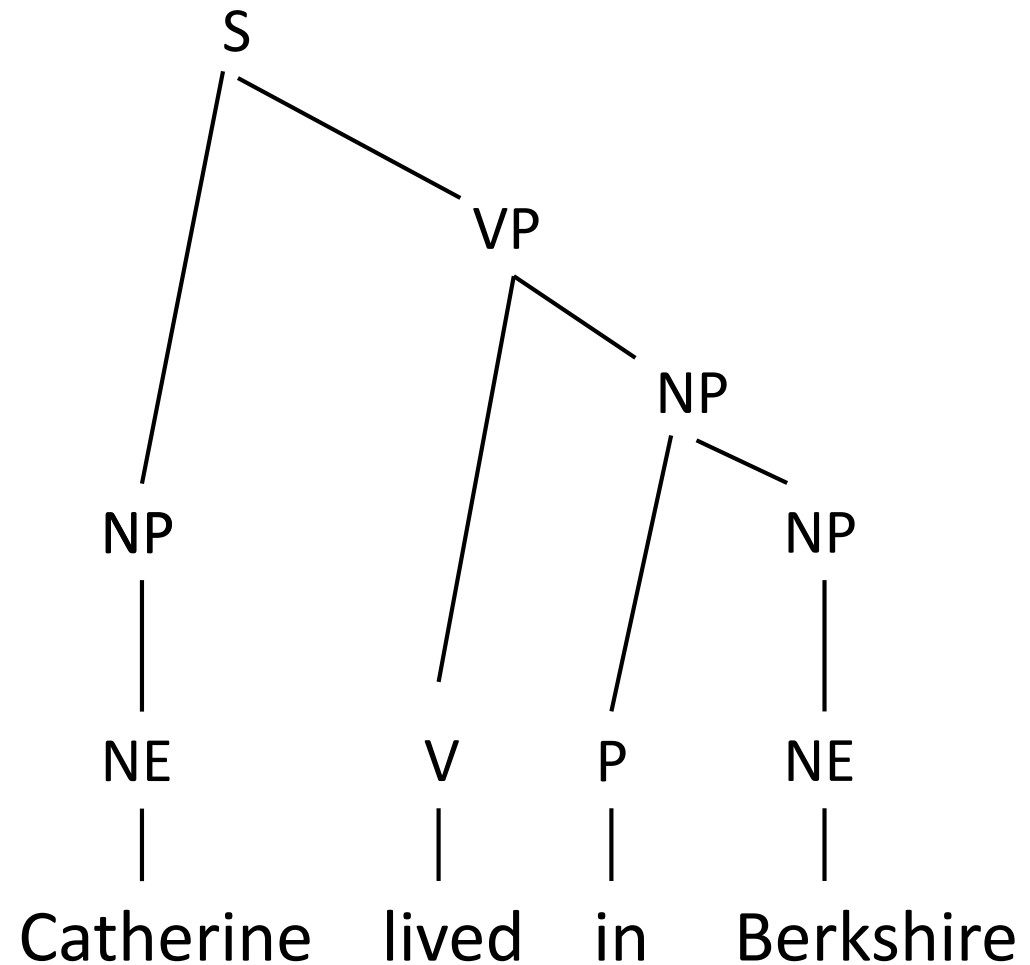
Parsing

- NER:
Catherine lived in **Berkshire**.
- Catherine lived in Location.
- Grammatik:
 - $S \rightarrow NP VP$
 - $NP \rightarrow NE$
 - $VP \rightarrow V PP$
 - $PP \rightarrow P NP$
- Lexikon oder Tagging:
 - $NE \rightarrow$ Catherine
 - $V \rightarrow$ lived
 - $P \rightarrow$ in
 - $NE \rightarrow$ **Location**



Parsing

- NER:
Catherine lived in **Berkshire**.
- Catherine lived in Location.
- Grammatik:
 - $S \rightarrow NP VP$
 - $NP \rightarrow NE$
 - $VP \rightarrow V PP$
 - $PP \rightarrow P NP$
- Lexikon oder Tagging:
 - $NE \rightarrow$ Catherine
 - $V \rightarrow$ lived
 - $P \rightarrow$ in
 - $NE \rightarrow$ **Location**



Bedeutung für Natural Language Processing

- NER meist einer der ersten Schritte für Informationsextraktion
- Wichtig für weitere Verarbeitung des Textes
 - Maschinelle Übersetzung
 - POS-Tagging
 - Semantic Role Labeling
 - Semantic Relations

Herausforderung Ambiguitäten (1/4)

Bedeutungstragende Namen:

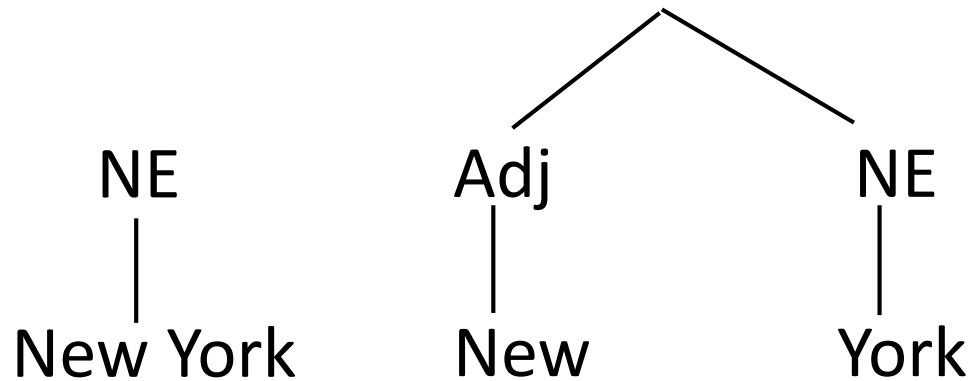
- Stein hält heute ihren Vortrag.
- Brown meets his friends.



Herausforderung Ambiguitäten (2/4)

Bedeutungstragende Namen:

- New York



- Los Angeles International Airport



Herausforderung Ambiguitäten (3/4)

- Er besuchte **JFK**: Person



- Er flog ab **JFK**: Flughafen bei New York



Herausforderung Ambiguitäten (4/4)

- Er besuchte **JFK**: Person



- Er besuchte **JFK**: Person



- Er flog ab **JFK**: Flughafen bei New York



- Er besuchte **JFK**: Person



Entity Resolution (ER)

- NER nicht immer ausreichend
- zusätzlich zu Klasse auch genaue Entität benötigt
- ER auch bezeichnet als Named Entity Linking
- Weist der Entität einen Link zu einem Datenbankeintrag zu

- Viele Touristen besuchen Westminster Abbey in London.
 - NER: London → LOCATION
 - ER: London → <https://de.wikipedia.org/wiki/London>
London → [https://de.wikipedia.org/wiki/London_\(Ontario\)](https://de.wikipedia.org/wiki/London_(Ontario))

Tool für NER: Stanford Named Entity Recognizer

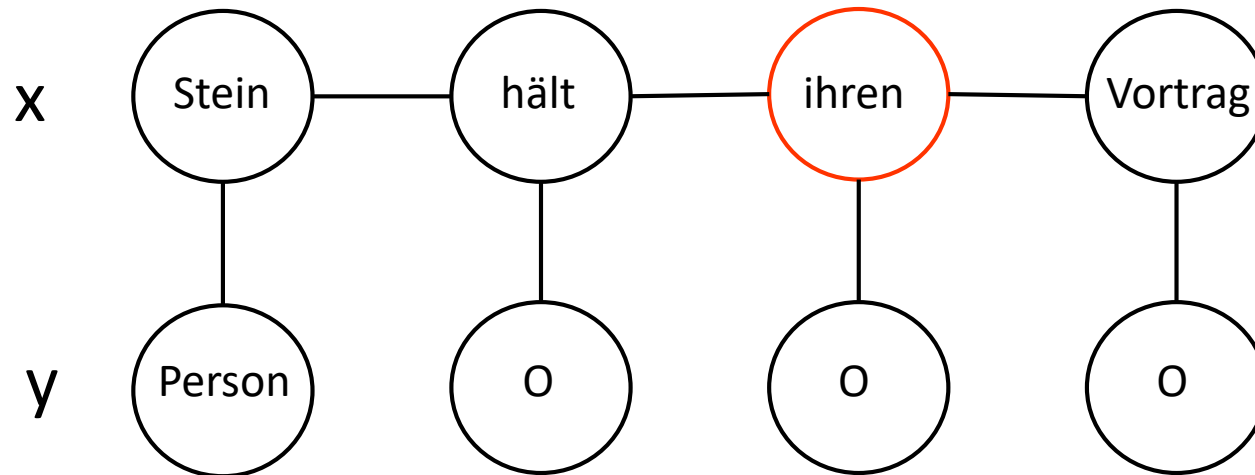
Stanford NER

- Stanford NLP Tools: Verschiedene Natural Language Processing Software zum Lösen computerlinguistischer Probleme
- Java Implementierung
- Training auf annotierten Daten, überwachtetes Verfahren:

John	PERSON
Kerry	PERSON
will	O
fly	O
to	O
Paris	LOCATION
this	O
weekend	O
.	O

Conditional Random Field Classifiers

- Sequence-to-sequence problem
- $\operatorname{argmax}_y P(y|x)$



Features

- Wörter: aktuelles, vorheriges und folgendes Wort, Wörter in festgelegtem Fenster
- Rechtschreibung:
 - Groß- und Kleinschreibung: Hans → Xxxx, eBay → xXxx
 - Buchstaben und Zahlen: MG3 → XX#
- Präfixe und Suffixe
- Distributionelle Ähnlichkeit

Featurefunktionen

- Featurefunktionen definieren
- Gewichte aus Trainingsdaten lernen
- Vorteil CRF:
 - Viele Kombinationsmöglichkeiten
 - Keine Unabhängigkeitsannahme

Gazette features

- Stanford NER verwendet keine
- Können beim Training den annotierten Daten hinzugefügt werden
- Gazetteer: eigentlich ein Ortslexikon, das alle Orte oder Regionen einer Gemeinde auflistet
- Gazette file:
 - Location Berkshire
 - Location Algermissen

Klassifizierer

- 3 Klassen:
 - PERSON, ORGANIZATION, LOCATION
 - Trainingsdaten: CoNLL, MUC6, MUC7, ACE, OntoNotes, Wikipedia
 - Robust bezüglich amerikanischem und britischem Englisch
- 4 Klassen:
 - PERSON, ORGANIZATION, LOCATION, MISC
 - Trainingsdaten: CoNLL 2003 Shared Task training data
- 7 Klassen:
 - PERSON, ORGANIZATION, LOCATION, DATE, MONEY, PERCENT, TIME
 - Trainingsdaten: MUC

NER für Englisch

- Zuerst in den Ordner wechseln, in dem sich der Stanford NER befindet

Starten unter Windows

- `java -mx600m -cp "*" ; lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] -textFile [inputfile]`

Starten unter Linux / MacOSX

- `java -mx600m -cp "*" :lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] -textFile [inputfile]`

NER für Deutsch

- `java -cp "*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier edu/stanford/nlp/models/ner/german.conll.hgc_175m_600.crf.ser.gz -tokenizerOptions latexQuotes=false -textFile [inputfile]`

Argumente

- Klassifizierer (nur für Englisch):
 - classifiers/english.all.3class.distsim.crf.ser.gz
 - classifiers/english.conll.4class.distsim.crf.ser.gz
 - classifiers/english.muc.7class.distsim.crf.ser.gz
- Ausgabedatei:
 - `java -mx600m -cp "*;lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] -outputFormat [option] -textFile [inputfile] > [outputfile]`
 - OutputFormat: slashTags (default), tsv, tabbedEntities, xml, inlineXML

Ausgabe für sample.txt

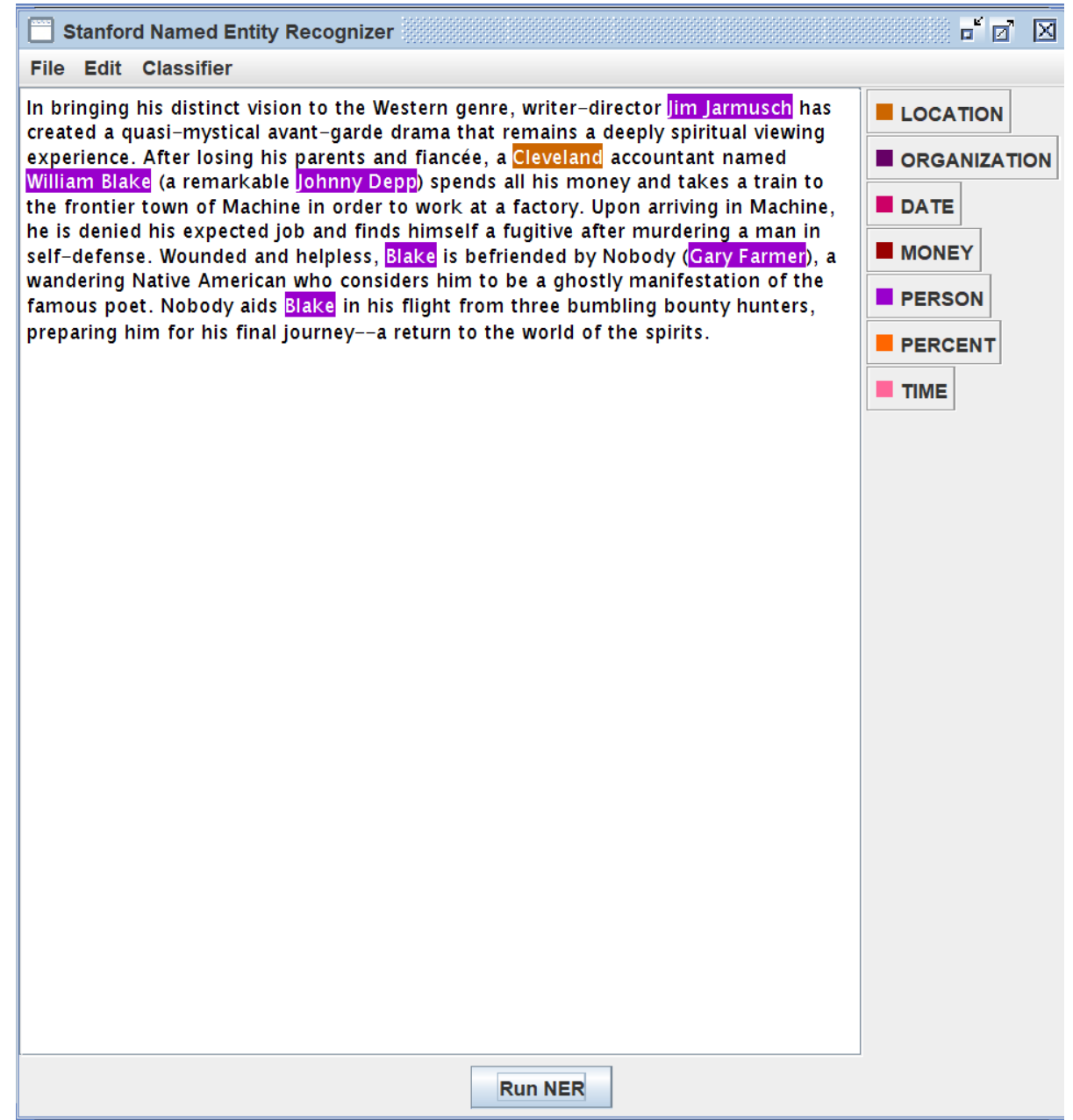
```
C:\Users\Katharina\Documents\Studium\ProSeminar\NER\stanford-ner-2017-06-09\stanford-ner-2017-06-09>java -mx600m -cp "*/lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier classifiers/english.all.3class.distsim.crf.ser.gz -textFile sample.txt
Invoked on Mon Nov 27 18:25:38 CET 2017 with arguments: -loadClassifier classifiers/english.all.3class.distsim.crf.ser.gz -textFile sample.txt
loadClassifier=classifiers/english.all.3class.distsim.crf.ser.gz
textFile=sample.txt
Loading classifier from classifiers/english.all.3class.distsim.crf.ser.gz ... done [1.6 sec].
The/O fate/O of/O Lehman/ORGANIZATION Brothers/ORGANIZATION ,/O the/O beleaguered/O investment/O bank/O ,/O hung/O in/O
the/O balance/O on/O Sunday/O as/O Federal/ORGANIZATION Reserve/ORGANIZATION officials/O and/O the/O leaders/O of/O majo
r/O financial/O institutions/O continued/O to/O gather/O in/O emergency/O meetings/O trying/O to/O complete/O a/O plan/O
to/O rescue/O the/O stricken/O bank/O ./O
Several/O possible/O plans/O emerged/O from/O the/O talks/O ,/O held/O at/O the/O Federal/ORGANIZATION Reserve/ORGANIZAT
ION Bank/ORGANIZATION of/ORGANIZATION New/ORGANIZATION York/ORGANIZATION and/O led/O by/O Timothy/PERSON R./PERSON Geith
ner/PERSON ,/O the/O president/O of/O the/O New/ORGANIZATION York/ORGANIZATION Fed/ORGANIZATION ,/O and/O Treasury/ORGAN
IZATION Secretary/O Henry/PERSON M./PERSON Paulson/PERSON Jr./PERSON ./O
CRFClassifier tagged 85 words in 2 documents at 416,67 words per second.

C:\Users\Katharina\Documents\Studium\ProSeminar\NER\stanford-ner-2017-06-09\stanford-ner-2017-06-09>
```

GUI (für Englisch)

- Windows: ner-gui.bat
- Linux / MacOSX: ner-gui.sh

```
C:\WINDOWS\system32\cmd.exe
C:\Users\Katharina\Documents\Studium\ProSeminar\NER\stanford-ner-2017-06-09\stanford-ner-2017-06-09>java -mx1500m -cp "stanford-ner.jar;lib/*" edu.stanford.nlp.ie.crf.NERGUI
Loading classifier from C:\Users\Katharina\Documents\Studium\ProSeminar\NER\stanford-ner-2017-06-09\stanford-ner-2017-06-09\classifiers\english.muc.7class.distsim.crf.ser.gz ... done [1.1 sec].
content type: text/rtf
PERSON: Jim Jarmusch
LOCATION: Cleveland
PERSON: William Blake
PERSON: Johnny Depp
PERSON: Blake
PERSON: Gary Farmer
PERSON: Blake
```



Übung

Übung

- Stanford NER über Kommandozeile starten
 - Englisch: sample.txt, sample-w-time.txt
 - Eigene Textdatei
 - Ausgabe auf der Kommandozeile und in Datei
 - Klassifizierer wechseln
 - OutputFormat wechseln

Übung

- Verschiedene Sätze und Texte ausprobieren
 - Wo sind die Problemstellen?
 - Was wird nicht erkannt oder falsch klassifiziert?
 - Was klappt gut?
 - Sätze mit Namen mit Bedeutung: Wann wird Name als Named Entity erkannt, wann nicht? Welche Rolle spielt der Kontext?
 - Ambige oder ungewöhnliche Eigennamen
 - Den gleichen Satz in unterschiedlichen Sprachen ausprobieren
- <http://corenlp.run/> oder GUI

Vielen Dank für eure Aufmerksamkeit