Text Mining - Praktikum

Satz-Reduktion mit Hilfe von NER-Tools

Text Mining Praktikum - Zwischenbericht - 14.12.2018

Marie Windhorst, Pia Sülzle, Vera Piontkowitz, Nils Wenzlitschke, Michael Brand

Satz-Reduktion

- Reduktion / Vereinfachung von Sätzen, z.B. Ersetzung von
  - Zeitangaben durch irgendwann
  - Ortsangaben durch irgendwo / da / dort etc.
  - Personenangaben durch irgendwer / jemand

- Beispiel:
  - Eingabe: Angela Merkel isst gerne bei McDonalds
  - Ausgabe: Irgendwer isst gerne irgendwo

Named Entity Recognition

- Eigennamenerkennung
- Tool zur automatischen Klassifizierung / Erkennung von Namen, Organisationen, Ortsangaben, Zeitangaben, Prozentangaben etc.
- Beispiel (IOB-chunk-representation):
  - Eingabe: Angela Merkel isst gerne bei McDonalds
  - Ausgabe:

    Angela B-PER

    Merkel I-PER

    isst O

    gerne O

    bei O

    McDonalds B-ORG

Verfügbare Named Entity Recognition Tools

spaCy

- in Python
- Corpus: TIGER und WikiNer

German NER

- in Java
- fordert IOB-Chunking
- Frankfurter Rundschau - 206.931 Tokens in 12.705 Sätzen

Stanford Named Entity Recognizer

- in Java
- eigene GUI
- Transkripte des Europäischen Parlaments

Stanford Named Entity Recognizer

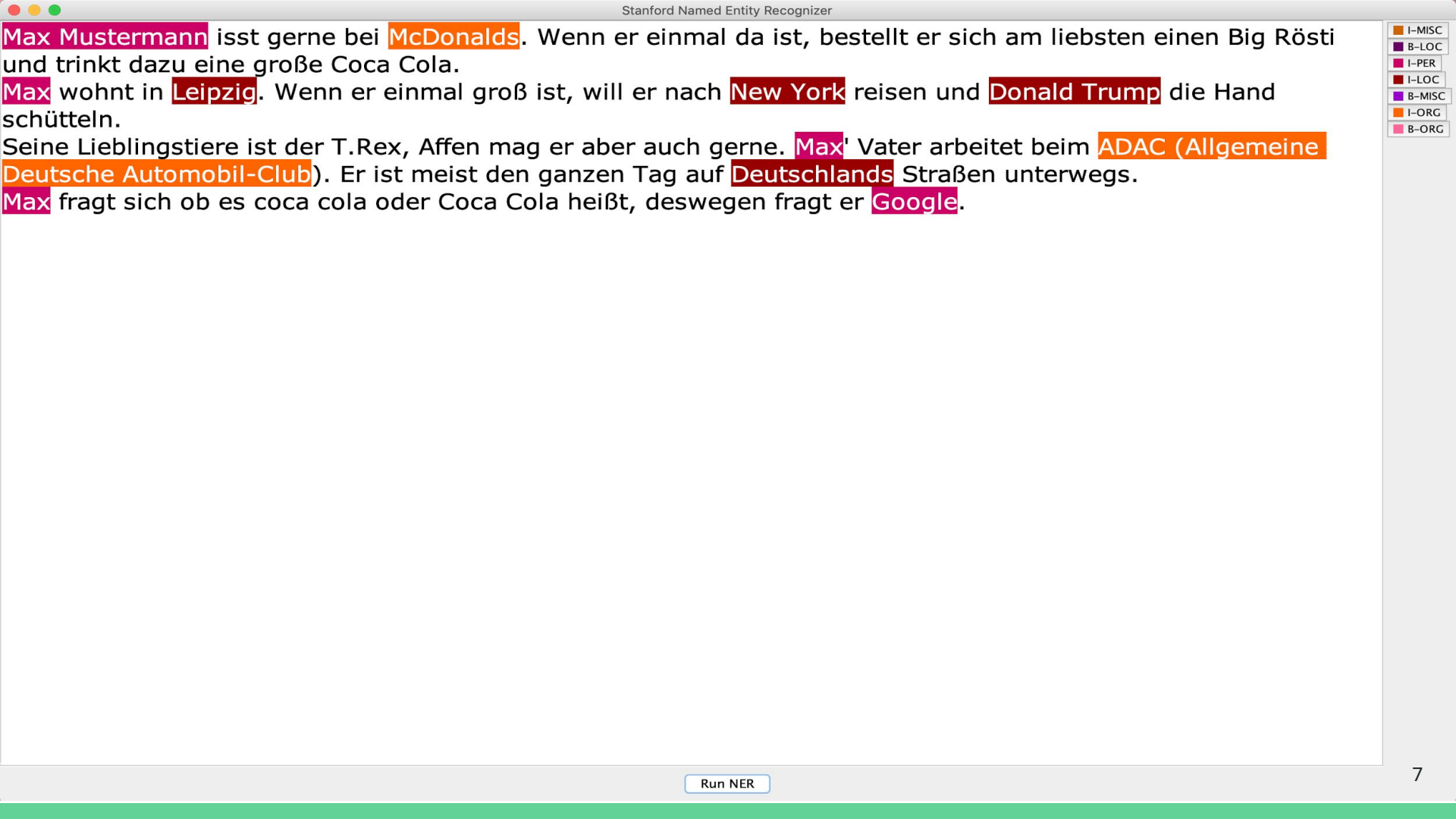
Max Mustermann isst gerne bei McDonalds. Wenn er einmal da ist, bestellt er sich am liebsten einen Big Rösti und trinkt dazu eine große Coca Cola.
Max wohnt in Leipzig. Wenn er einmal groß ist, will er nach New York reisen und Donald Trump die Hand schütteln. Seine Lieblingstiere ist der T.Rex, Affen mag er aber auch gerne. Max' Vater arbeitet beim ADAC (Allgemeine Deutsche Automobil-Club). Er ist meist den ganzen Tag auf Deutschlands Straßen unterwegs.
Max fragt sich ob es coca cola oder Coca Cola heißt, deswegen fragt er Google.

Run NER

Max Mustermann isst gerne bei McDonalds. Wenn er einmal da ist, bestellt er sich am liebsten einen Big Rösti und trinkt dazu eine große Coca Cola.
Max wohnt in Leipzig. Wenn er einmal groß ist, will er nach New York reisen und Donald Trump die Hand schütteln.
Seine Lieblingstiere ist der T.Rex, Affen mag er aber auch gerne. Max' Vater arbeitet beim ADAC (Allgemeine Deutsche Automobil-Club). Er ist meist den ganzen Tag auf Deutschlands Straßen unterwegs.
Max fragt sich ob es coca cola oder Coca Cola heißt, deswegen fragt er Google.

I-MISC
B-LOC
I-PER
I-LOC
B-MISC
I-ORG
B-ORG

Run NER

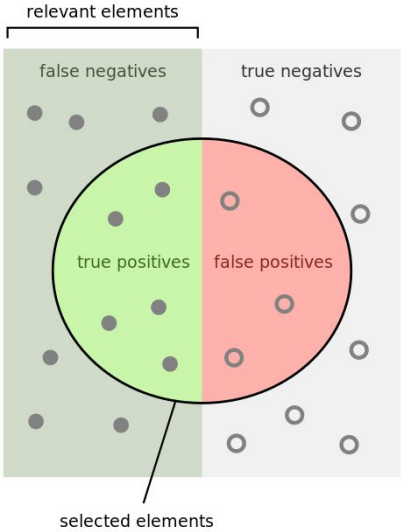

Stanford Named Entity Recognizer
Max Mustermann isst gerne bei McDonalds. Wenn er einmal da ist, bestellt er sich am liebsten einen Big Rösti und trinkt dazu eine große Coca Cola.
Max wohnt in Leipzig. Wenn er einmal groß ist, will er nach New York reisen und Donald Trump die Hand schütteln.
Seine Lieblingstiere ist der T.Rex, Affen mag er aber auch gerne. Max' Vater arbeitet beim ADAC (Allgemeine Deutsche Automobil-Club). Er ist meist den ganzen Tag auf Deutschlands Straßen unterwegs.
Max fragt sich ob es coca cola oder Coca Cola heißt, deswegen fragt er Google.
I-MISC
B-LOC
I-PER
I-LOC
B-MISC
I-ORG
B-ORG
Run NER
7

stanford_output_max.txt

<I-PER>Max Mustermann</I-PER> isst gerne bei <I-ORG>McDonalds</I-ORG>. Wenn er einmal da ist, bestellt er sich am liebsten einen Big Rösti und trinkt dazu eine große Coca Cola.
<I-PER>Max</I-PER> wohnt in <I-LOC>Leipzig</I-LOC>. Wenn er einmal groß ist, will er nach <I-LOC>New York</I-LOC> reisen und <I-LOC>Donald Trump</I-LOC> die Hand schütteln.
Seine Lieblingstiere ist der T.Rex, Affen mag er aber auch gerne. <I-PER>Max</I-PER>' Vater arbeitet beim <I-ORG>ADAC (Allgemeine Deutsche Automobil-Club</I-ORG>). Er ist meist den ganzen Tag auf <I-LOC>Deutschlands</I-LOC> Straßen unterwegs.
<I-PER>Max</I-PER> fragt sich ob es coca cola oder Coca Cola heißt, deswegen fragt er <I-PER>Google</I-PER>.

Named Entity Recognition - Evaluation

- Vergleich von NER-Tools
- Statistische Evaluation - Evaluationsmaße:
  - Precision = Anzahl korrekt klassifizierter NEs / Anzahl NEs gefunden
    - Qualität
  - Recall = Anzahl korrekt klassifizierter NEs / Anzahl vorhandener NEs
    - Quantität
  - F1-Test = (2 * Precision * Recall) / (Precision + Recall)
    - Idealwert: 1 (perfect precision and recall)



Evaluation

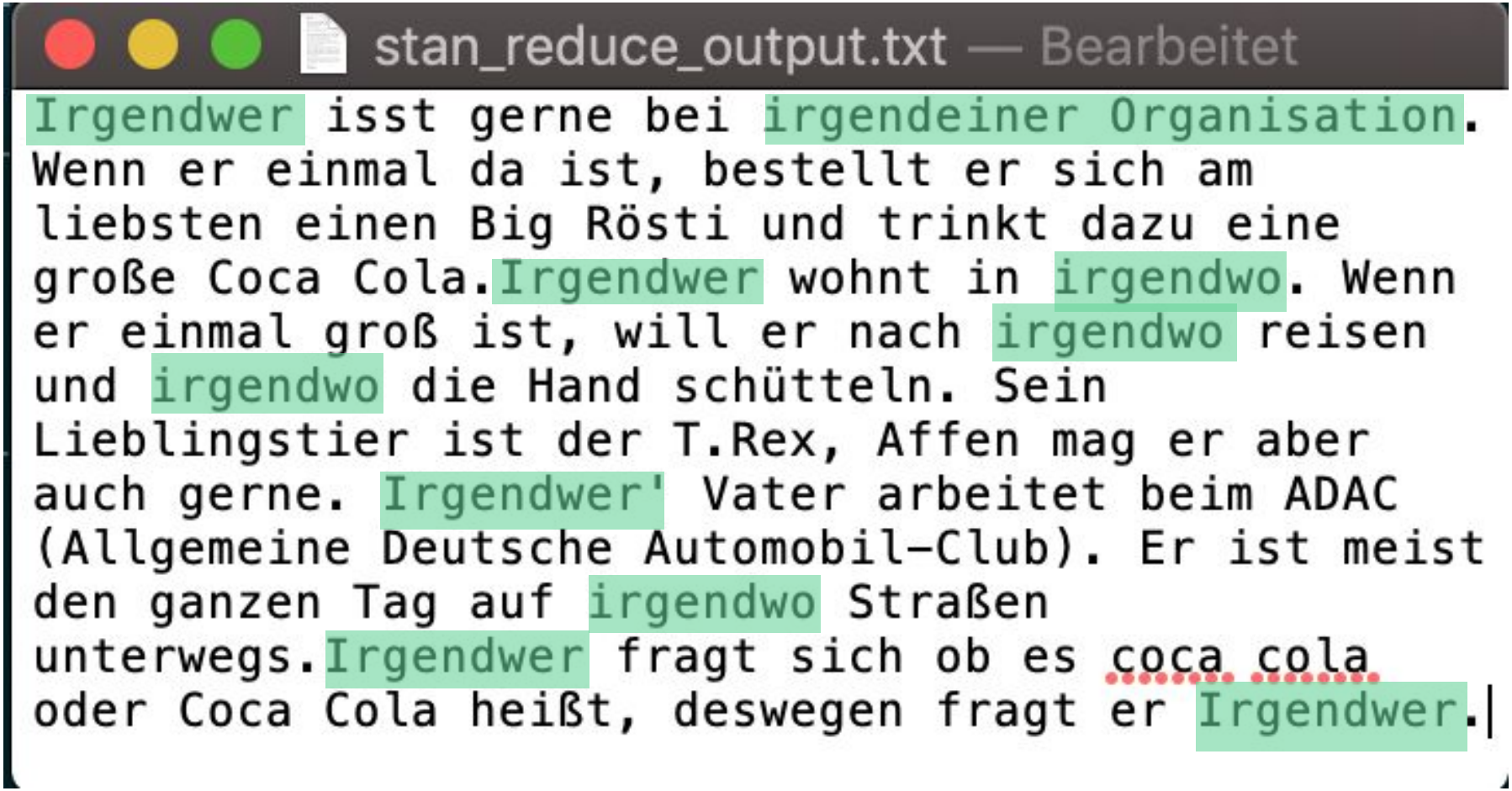
	Precision	Recall	F1-Wert
SpaCy	$\frac{78}{109} = 0,715$	$\frac{78}{116} = 0,672$	$\frac{(2 \cdot 0,715 \cdot 0,672)}{(0,715 + 0,672)} = 0,693$
Stanford NER	$\frac{54}{60} = 0,9$	$\frac{54}{116} = 0,466$	$\frac{(2 \cdot 0,9 \cdot 0,466)}{(0,9 + 0,466)} = 0,614$
GermaNER	$\frac{58}{70} = 0,829$	$\frac{58}{116} = 0,5$	$\frac{(2 \cdot 0,829 \cdot 0,5)}{(0,829 + 0,5)} = 0,624$

Problem-Evaluierung

- kein Tool perfekt
  - Kombination verschiedener Tools?
  - Eigenes Modell erstellen?
- Satzreduktion abhängig von Satzstruktur
  - besonders Personen (irgendwer/wem/was/wen)
  - Wie Organisationen ersetzen? (irgendwo/irgendwas)
  - Relativsätze

stanford_output_max.txt

<I-PER>Max Mustermann</I-PER> isst gerne bei <I-ORG>McDonalds</I-ORG>. Wenn er einmal da ist, bestellt er sich am liebsten einen Big Rösti und trinkt dazu eine große Coca Cola.
<I-PER>Max</I-PER> wohnt in <I-LOC>Leipzig</I-LOC>. Wenn er einmal groß ist, will er nach <I-LOC>New York</I-LOC> reisen und <I-LOC>Donald Trump</I-LOC> die Hand schütteln.
Seine Lieblingstiere ist der T.Rex, Affen mag er aber auch gerne. <I-PER>Max</I-PER>' Vater arbeitet beim <I-ORG>ADAC (Allgemeine Deutsche Automobil-Club</I-ORG>).
Er ist meist den ganzen Tag auf <I-LOC>Deutschlands</I-LOC> Straßen unterwegs.
<I-PER>Max</I-PER> fragt sich ob es coca cola oder Coca Cola heißt, deswegen fragt er <I-PER>Google</I-PER>.


stan_reduce_output.txt — Bearbeitet
Irgendwer isst gerne bei irgendeiner Organisation.
Wenn er einmal da ist, bestellt er sich am
liebsten einen Big Rösti und trinkt dazu eine
große Coca Cola.Irgendwer wohnt in irgendwo. Wenn
er einmal groß ist, will er nach irgendwo reisen
und irgendwo die Hand schütteln. Sein
Lieblingstier ist der T.Rex, Affen mag er aber
auch gerne. Irgendwer' Vater arbeitet beim ADAC
(Allgemeine Deutsche Automobil-Club). Er ist meist
den ganzen Tag auf irgendwo Straßen
unterwegs.Irgendwer fragt sich ob es coca cola
oder Coca Cola heißt, deswegen fragt er Irgendwer.