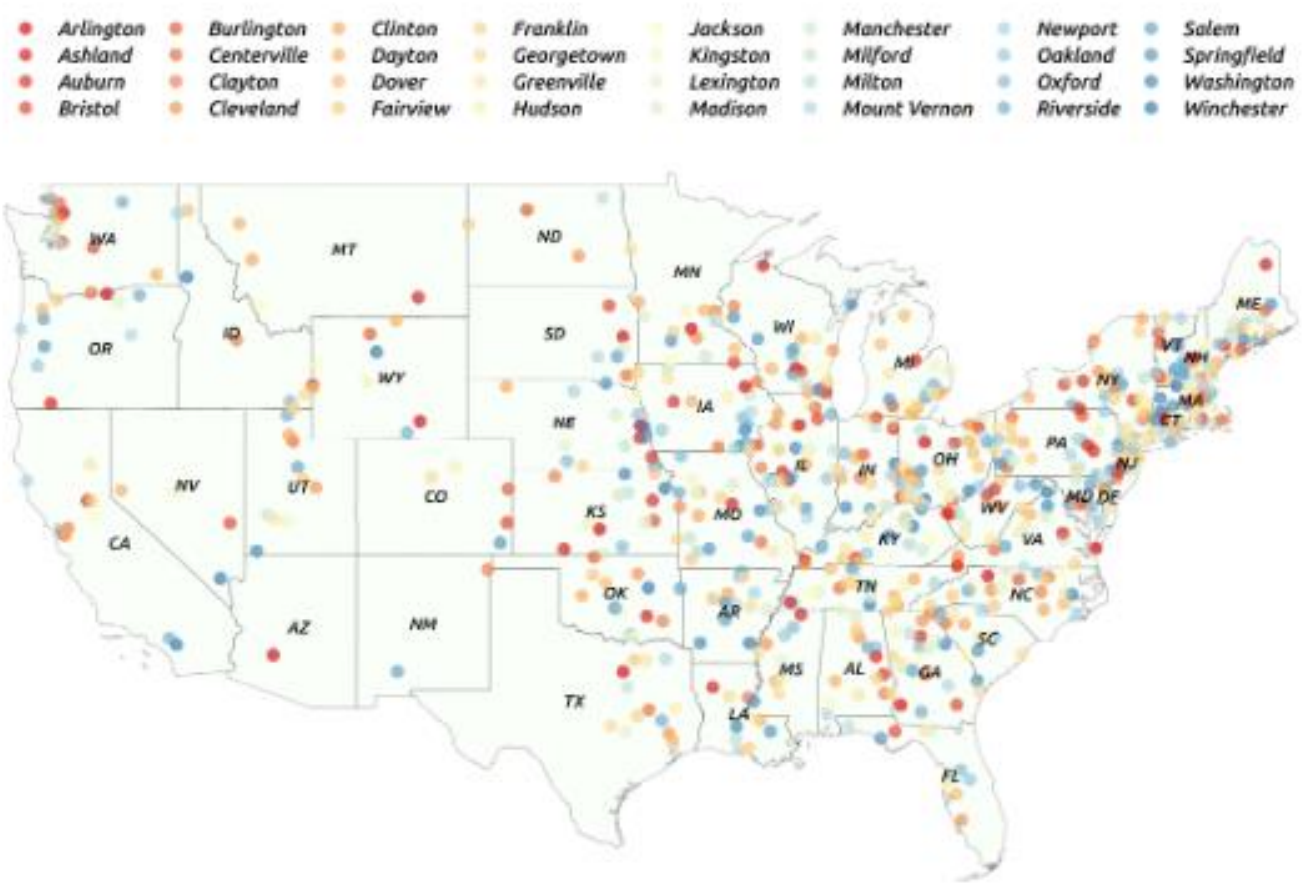
TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

Identifikation und Disambiguierung von Ortsbezeichnungen in bibliografischen Metadaten auf Grundlage von maschinellem Lernen zur Anreicherung mit Koordinaten




Arlington
Ashland
Auburn
Bristol
Burlington
Centerville
Clayton
Cleveland
Clinton
Dayton
Dover
Fairview
Franklin
Georgetown
Greenville
Hudson
Jackson
Kingston
Lexington
Madison
Manchester
Milford
Milton
Mount Vernon
Newport
Oakland
Oxford
Riverside
Salem
Springfield
Washington
Winchester
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
DE
NJ
CT
RI
MA
VT
NH
ME

[1]

East London, Südafrika

Berlin, 5660, Südafrika

Reiseziel hinzufügen

Jetzt starten OPTIONEN

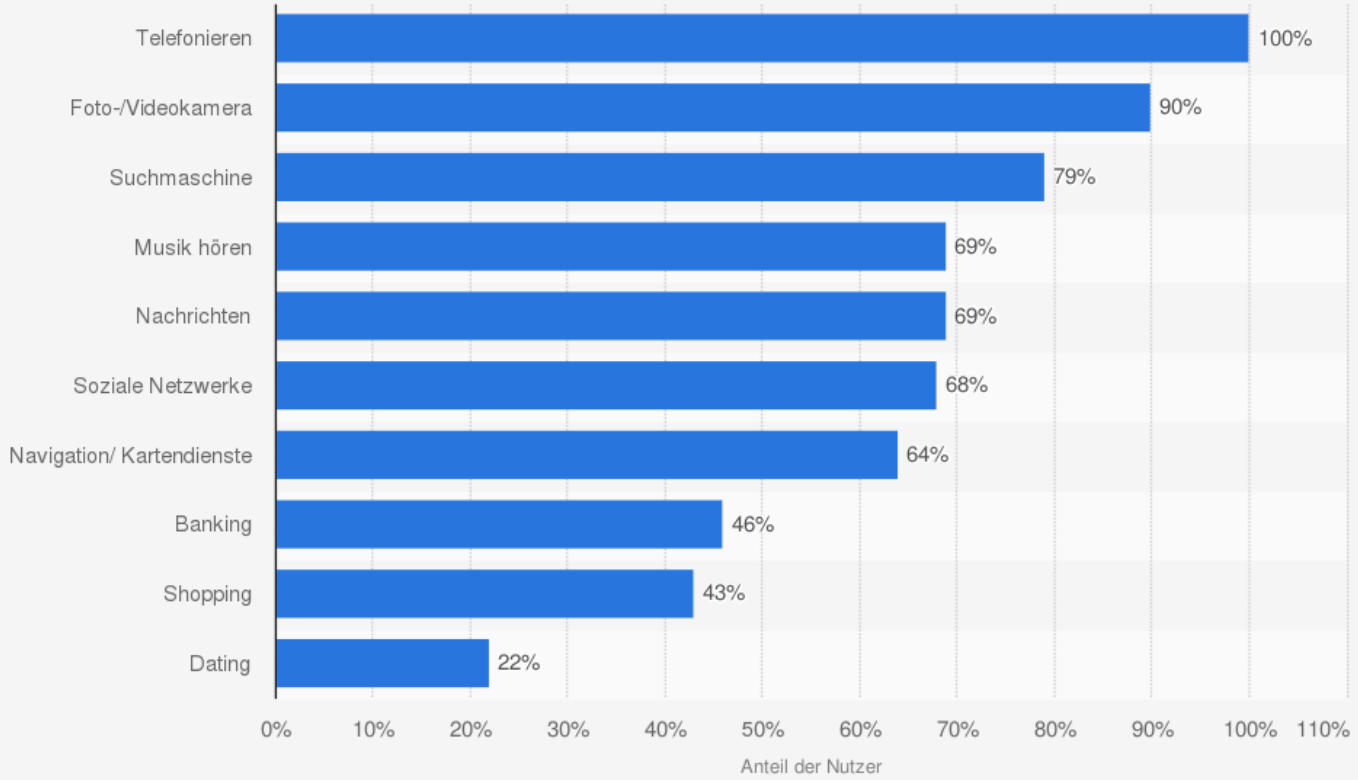
Wegbeschreibung an mein Smartphone senden

über N2 46 min

Schnellste Route; übliche Verkehrslage 53,2 km

DETAILS

Anteil der befragten Smartphone-Nutzer, die die folgenden Funktionen mit ihrem Smartphone nutzen



Quelle
Bitkom
© Statista 2021

Weitere Informationen:
Deutschland; Bitkom Research; Januar 2017; Smartphone-Nutzer; ab 14 Jahre

vufind

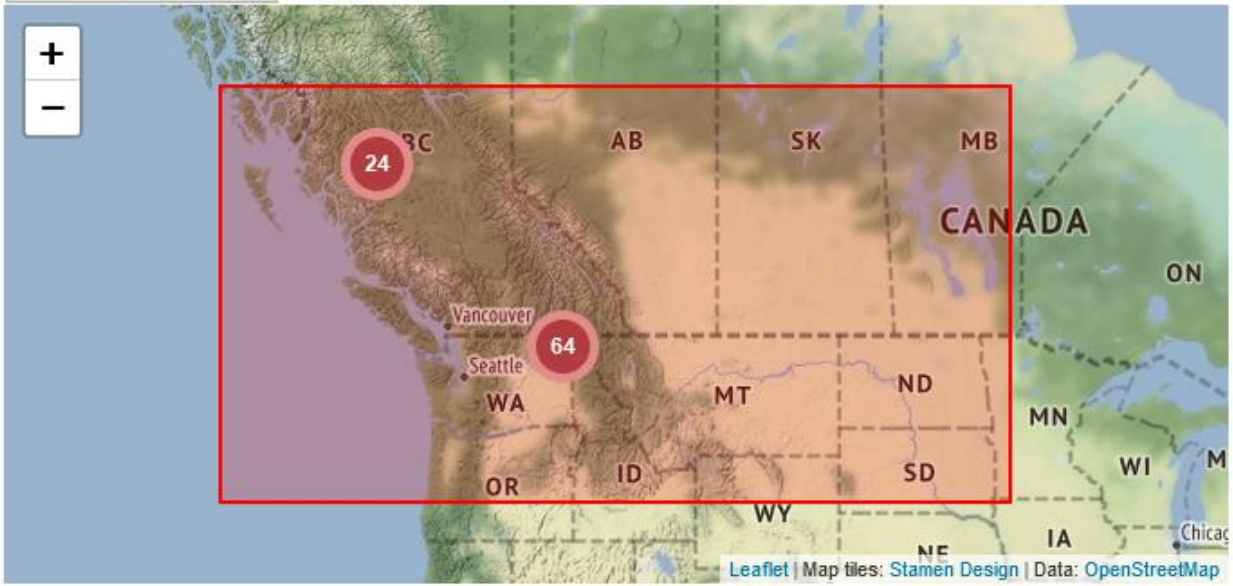
Search. Discover. Share.

Management | Alle Felder | Suchen | Erweitert

Filter zurücksetzen | Geografische Suche: Intersects(ENVELOPE(-134.208984375, -95.712890625, 56.2954074618

Suchergebnisse

Suchbereich definieren | Hilfe



Gesamt 314.492
Mit Koordinaten 58.758

```
=LDR  03747nz  a2200709nc 4500
=001  040182665
=003  DE-101
=005  20181113150134.0
=008  880701n||azznnaabn\\\\\\\\\\|\ana\\\\|c
=024  7\$a4018266-6$0http://d-nb.info/gnd/4018266-6$2gnd
=034  \\$dE 013 20 19$eE 013 20 19$fN 050 54 39$gN 050 54 39$2geonames$0http://sws.geonames.org/2925192$9A:agx
=034  \\$dE013.338611$eE013.338611$fN050.910833$gN050.910833$2geonames$0http://sws.geonames.org/2925192$9A:dgx
=035  \\$a(DE-101)040182665
=035  \\$a(DE-588)4018266-6
=035  \\$z(DE-588)2010860-6
=035  \\$z(DE-588b)6519955-8
=035  \\$z(DE-588b)2010860-6$9v:zg
=035  \\$z(DE-588c)4018266-6$9v:zg
=040  \\$aDE-101$cDE-101$9r:DE-101$bger$d1210$erda
=042  \\$agnd1
=043  \\$cXA-DE-SN
=075  \\$bg$2gndgen
=075  \\$bgik$2gndspec
=079  \\$ag$qs$qf$qh$qg$uw$uz$uv$uo
=083  04$z2$a432161$9d:3$9t:2010-06-07$222/ger
=089  04$z2$a4321613$9t:2007-01-01$9g:2010-06-07$222/ger
=151  \\$aFreiberg
=410  2\$aFreiberg$bOberbürgermeister$4spio$4https://d-nb.info/standards/elementset/gnd#variantName$wr$iSpitzenorgan$eSpitzenorgan
=410  2\$aOberbürgermeister$gFreiberg$4spio$4https://d-nb.info/standards/elementset/gnd#variantName$wr$iSpitzenorgan$eSpitzenorgan
=410  2\$aFreiberg$bAmt des Oberbürgermeisters, Städtepartnerschaften$4spio$4https://d-nb.info/standards/elementset/gnd#variantName$wr$iSpitzenorgan$eSpitzenorgan
=410  2\$aAmt des Oberbürgermeisters, Städtepartnerschaften$gFreiberg$4spio$4https://d-nb.info/standards/elementset/gnd#variantName$wr$iSpitzenorgan$eSpitzenorgan
=410  2\$aFreiberg$bAbteilung Jugend, Sport und Naherholung$4spio$4https://d-nb.info/standards/elementset/gnd#variantName$wr$iSpitzenorgan$eSpitzenorgan
```

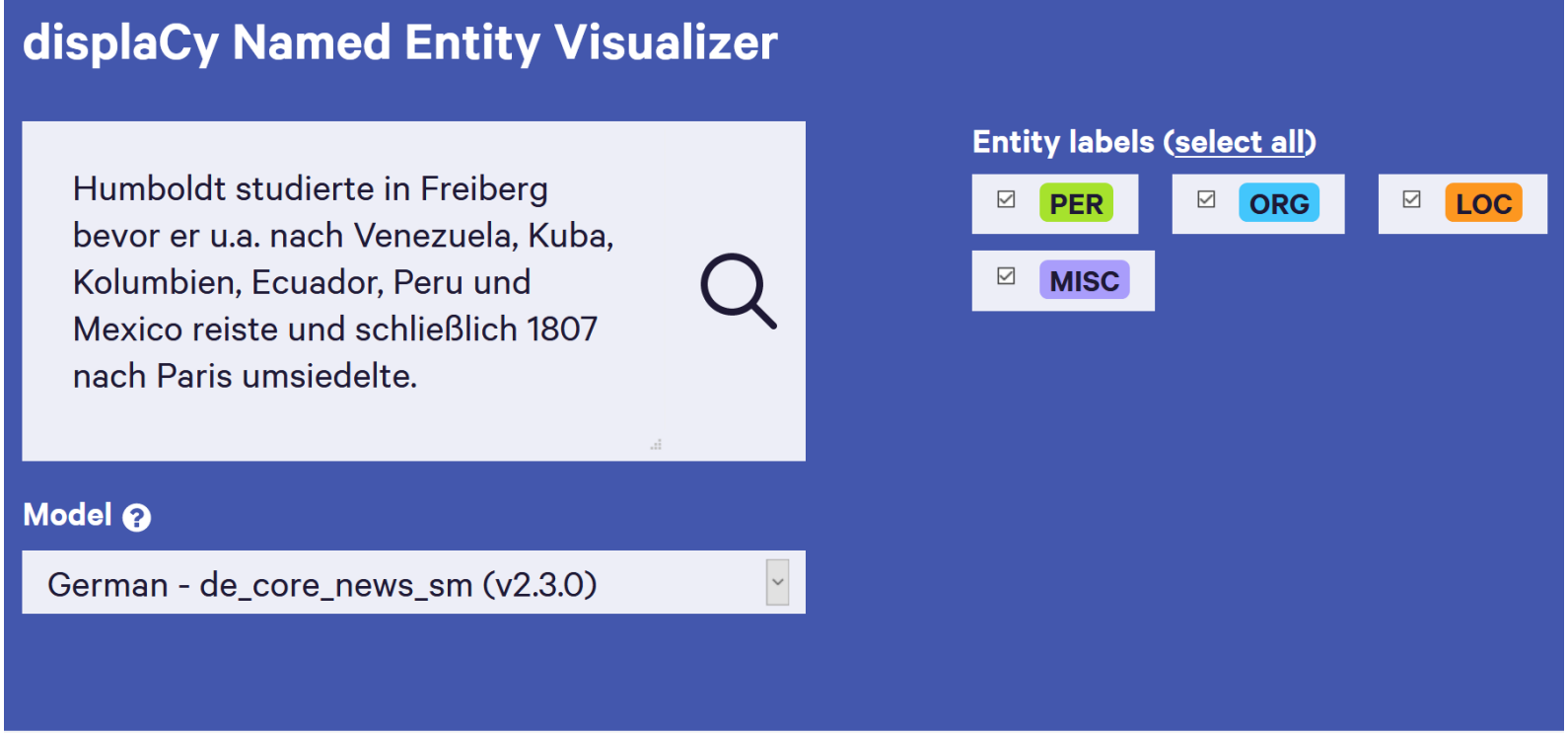
NER

- *Named Entity Recognition (ner) is the information extraction task of identifying and classifying mentions of people, organisations, locations and other named entities (nes) within text.* [2]

= ein Teilgebiet von NLP (Natural Language Processing)

- Grundlage: trainierte Daten
    - Supervised Learning: Annotation von Objekten

```
144 Altea|NE|I-LOC hatte|VAFIN|O bei|APPR|O einer|ART|O Fläche|NN|O von|APPR|O 34,4|CARD|O km²|NN|O am|APPRART|O 1.|ADJA|O Januar|NN|O
    2009|CARD|O 23.780|CARD|O Einwohner|NN|O .|$.|O
145 In|APPR|O Altea|NE|I-LOC befindet|VVFIN|O sich|PRF|O die|ART|O Fakultät|NN|O für|APPR|O Schöne|ADJA|O Künste|NN|O der|ART|O
    Universität|NN|I-LOC Miguel|NE|I-LOC Hernández|NE|I-LOC Elche|NN|I-LOC .|$.|O
146 Die|ART|O Gründung|NN|O erfolgte|VVFIN|O durch|APPR|O Iberer|NN|O und|KON|O Römer|NE|I-LOC .|$.|O
147
148
149 Hermann|NE|I-PER Jüngken|NN|I-PER war|VAFIN|O ein|ART|O Rittergutsbesitzer|NN|O und|KON|O Reichstagsabgeordneter|NN|O .|$.|O
150 Jüngken|NN|I-PER war|VAFIN|O 1859|CARD|O bis|KON|O 1876|CARD|O Rittergutsbesitzer|NN|O in|APPR|O Reinsdorf|NN|I-LOC bei|APPR|O
    Artern|NN|I-LOC .|$.|O
151
152 Auch|ADV|O Graf|NE|O Ludwig|NE|I-PER Heinrich|NE|I-PER von|APPR|I-PER Nassau-Dillenburg|NN|I-PER hatte|VAFIN|O sich|PRF|O um|APPR|O
    ihre|PPOSAT|O Hand|NN|O bemüht|VVPP|O ,|$,|O was|PRELS|O zu|APPR|O Streitigkeiten|NN|O führte|VVFIN|O .|$.|O
153
```



[3]

Humboldt PER studierte in Freiberg LOC bevor er u.a. nach Venezuela LOC , Kuba LOC , Kolumbien LOC , Ecuador LOC , Peru LOC und Mexico LOC reiste und schließlich 1807 nach Paris LOC umsiedelte.

Accuracy Evaluation

TAG_ACC	Part-of-speech tags (fine grained tags, Token.tag)	0.98
ENTS_P	Named entities (precision)	0.86
ENTS_R	Named entities (recall)	0.85
ENTS_F	Named entities (F-score)	0.85
SENTS_P	Sentence segmentation (precision)	0.95
SENTS_R	Sentence segmentation (recall)	0.96
SENTS_F	Sentence segmentation (F-score)	0.95
TOKEN_ACC	Tokenization	1.00
POS_ACC	Part-of-speech tags (coarse grained tags, Token.pos)	0.98
MORPH_ACC		0.92
LEMMA_ACC		0.73
DEP_UAS	Unlabelled dependencies	0.93
DEP_LAS	Labelled dependencies	0.91

[4]

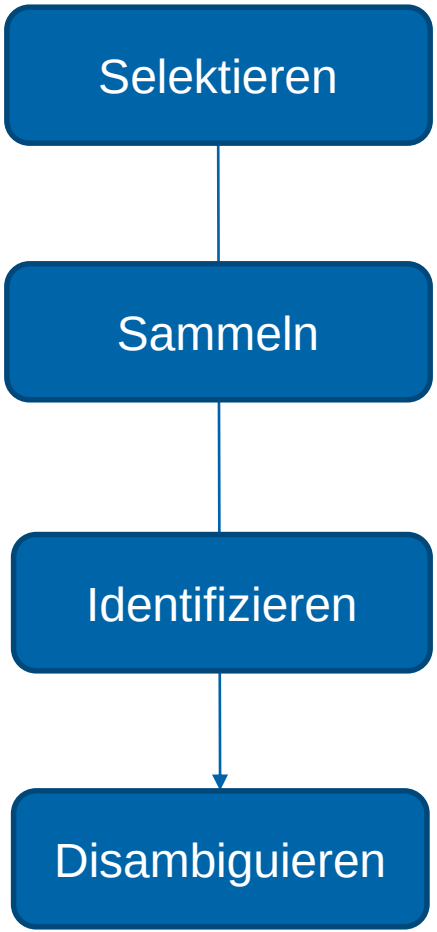
Akt. Highscore: LUKE = 94,3% (https://github.com/studio-ousia/luke)

Mordecai

- Ermittelt aus unstrukturiertem Text Geo-Entities
- Nutzt spaCy (NER)
  - Textkorpus Trainingset auf Basis von Wikipedia-Artikeln [5]
  - .. und TIGER Korpus (Uni Stuttgart, Institut für Maschinelle Sprachverarbeitung, [6]
  - Annotation mittels prodigy [7]
  - KNN implementiert in Keras (Open Source DL-Bibliothek)
- Ermittelte Geo-Entities werden an Geonames-Gazetteer geschickt und Koordinaten ausgegeben
  - The GeoNames gazetteer was chosen for GeoTxt because of its extensive coverage, quality, inclusion of metadata (such as alternate names and geographic hierarchical information), and frequent updates (Acheson, Sabbata, & Purves, 2017).

Toponym Resolution

- Über Heuristiken: population, area, or geographic-level prominence
- Co-occurences
  - spatial minimality: spatial proximity (i.e. assuming that toponyms in a document are likely to constitute a “spatial cluster”) and therefore toponym candidates for co-occurring place names that minimize the average distance between all toponym predictions are prioritized over distant candidates
  - spatial hierarchy (country, state, county, township, populated place) or on names that belong to the same subtree of such hierarchy (counties that are both located within a particular state) [2]



Alle GND-IDs ohne 034
DNB-IDs aus 024
Ländercode 043
Ortsnamen und –varianten 151, 451?

Alle im k10+ verknüpften Objekte plus Katalogisate

Named Entity Recognition (NER)

Toponym Resolution

Geoparsing
(zB mordecai)

Vorgehen

- Geoparser-Software evaluieren [8]
- Korrektheit der vorgefertigten Modelle überprüfen
  - EN/DE-cores
  - Ggf. Anpassungen vornehmen

=> Intellektuell vergebenen GnD-Geo-Entities mit mordecai-Output vergleichen

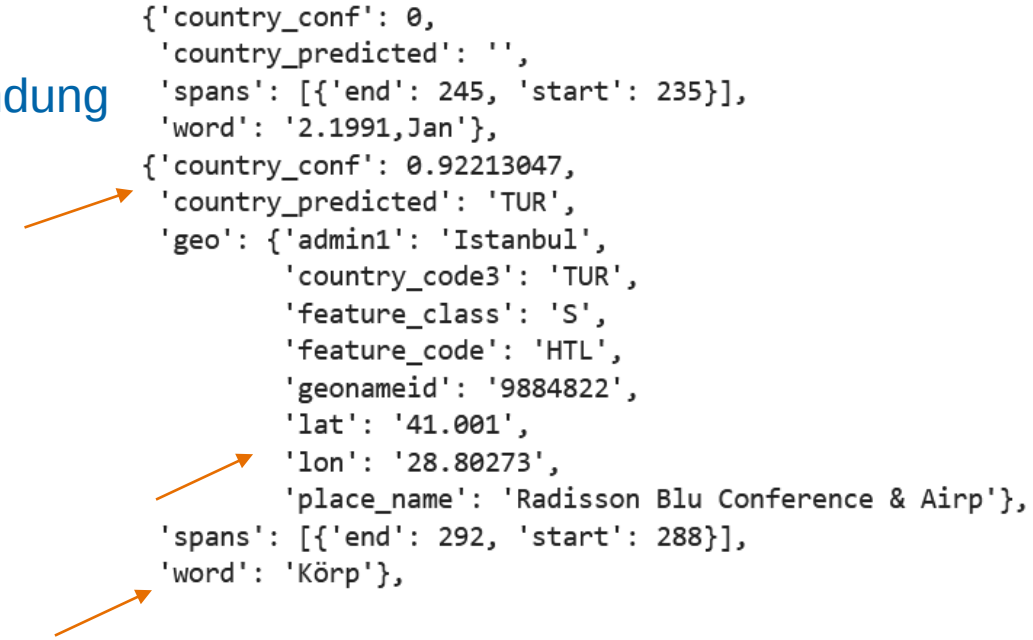
- Verarbeitung mit unterschiedlich komplexen Heuristiken auswerten
  - Katalogisat
  - TOC, Abstract, Volltext
  - z.B. über SRU-K10+
    - etwa 70.000 Zeichen Zusatzinformationen

Aktuelles Setting

- Docker Desktop mit Elasticsearch 5.5.2
- Python 3.9.4
- Tensorflow: 2.5.0
- spaCy: 3.0.6
    - en- und de-cores: 3.0.0
- Mainboard: MSI B450M PRO-VDH Plus, AMD B450
- Prozessor: AMD Ryzen 5 2600X 6x 4.2GHz
- Arbeitsspeicher: 16GB DDR4-RAM PC-3000 (2x 8GB)
- Grafikkarte: Nvidia GeForce RTX2060 6GB, Palit Gaming Pro OC
    - 1920 CUDA-Cores
- Verarbeitung: knapp 7h

Beobachtungen

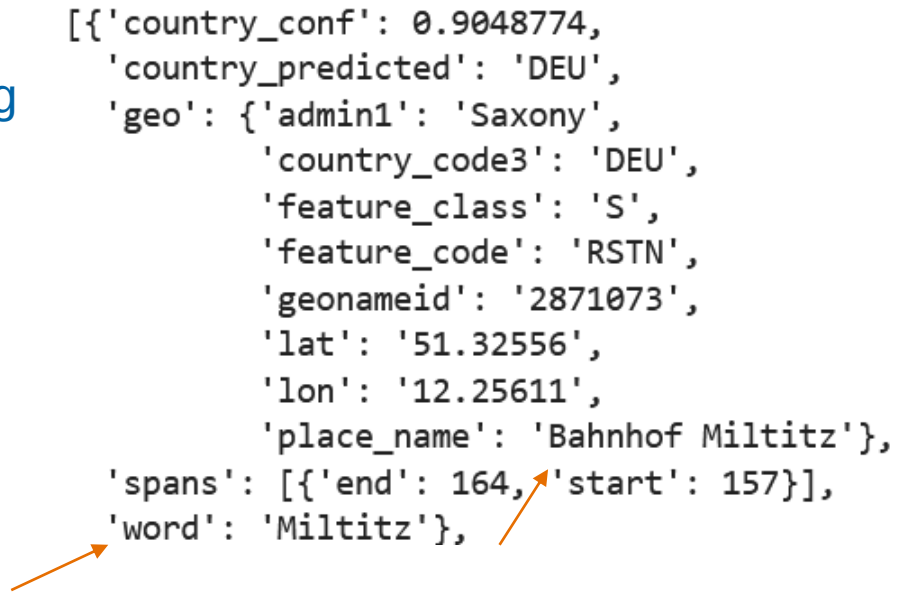
- Verarbeitung von Heuristiken mitunter erschwert durch Verwendung von Abkürzungen -> nur Titel



Beobachtungen

- Verarbeitung von Heuristiken mitunter erschwert durch Verwendung von Abkürzungen -> nur Titel

- Ausgabe von Ortsteilen / -bereichen



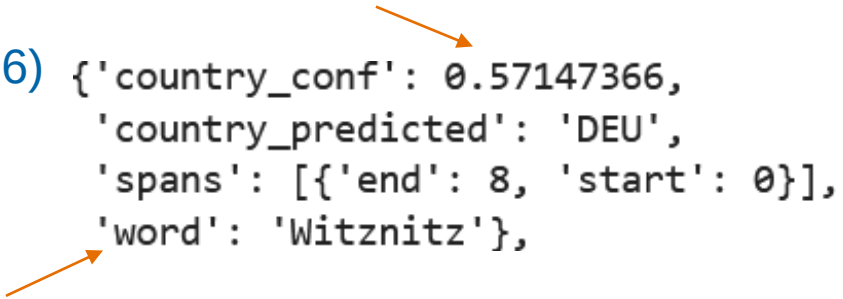
Beobachtungen

- Verarbeitung von Heuristiken

mitunter erschwert durch Verwendung von Abkürzungen -> nur Titel

- Ausgabe von Ortsteilen / -bereichen

- Schwellwert anpassen (default: 0,6)



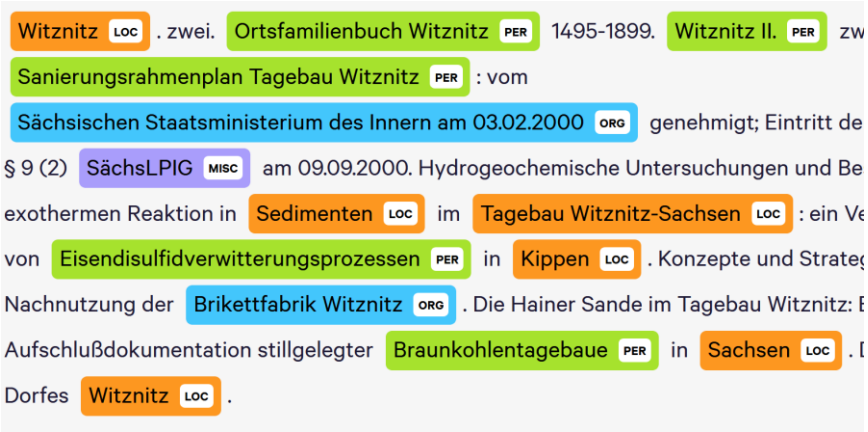
Beobachtungen

- 27 / 314 Matches

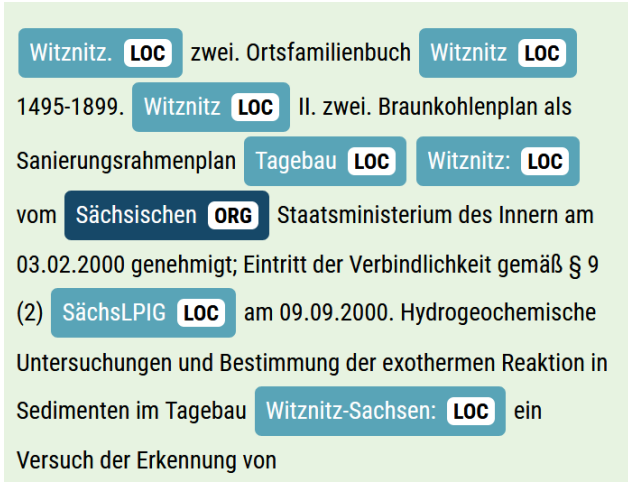
dnb_id	coordinates	placename	mordecai-conf	mordecai-placer	mordecai-latlon
http://d-nb.info/gnd/126317-1	E 011 36 01, E 011 36 01, N 050 55 57, N 050 55 57	Wenigenjena	0.9052457	Wenigenjena	https://www.latlong.net/c/?lat=50.93225&long=11.59993
http://d-nb.info/gnd/2019606-4	E 012 15 17, E 012 15 17, N 051 19 28, N 051 19 28	Miltitz	0.8359622	Bahnhof Miltitz	https://www.latlong.net/c/?lat=51.32556&long=12.25611
http://d-nb.info/gnd/3016944-6	E 011 19 09, E 011 19 09, N 051 08 37, N 051 08 37	Ellersleben	0.9052457	Ellersleben	https://www.latlong.net/c/?lat=51.15&long=11.31667
http://d-nb.info/gnd/4006424-4	E 009 47 00, E 009 47 00, N 048 06 00, N 048 06 00	Biberach an der Riß	0.8359622	Biberach an der Riß	https://www.latlong.net/c/?lat=48.08942&long=9.79942
http://d-nb.info/gnd/4050796-8	E 011 42 24, E 011 42 24, N 051 06 32, N 051 06 32	Rudelsburg	0.8359622	Rummelsburg	https://www.latlong.net/c/?lat=52.50146&long=13.4934
http://d-nb.info/gnd/4064419-4	E 008 16 09, E 008 16 09, N 047 37 53, N 047 37 53	Waldshut	0.8359622	Landkreis Waldshut	https://www.latlong.net/c/?lat=47.70556&long=8.255
http://d-nb.info/gnd/4104598-1	E 009 08 45, E 009 08 45, N 049 18 50, N 049 18 50	Burg Hornberg	0.8359622	Neckarzimmern Burg Hornberg	https://www.latlong.net/c/?lat=49.3139&long=9.14583

Ausblick

- Transfer Learning
- BERT (z.B. German Bert Model)
- Heuristiken (Ländercode)
- Fachdatenbank (Wikidata, etc.)



https://explosion.ai/demos/displacy-ent



https://demos.deepset.ai/ner

Vielen Dank für Ihre Aufmerksamkeit!

oliver.loewe@ub.tu-freiberg.de

Fachinformationsdienst Montan
https://montanportal.info

Universitätsbibliothek Freiberg
https://tu-freiberg.de/ub

Quellen

[1] https://grantmckenzie.com/academics/McKenzie_2016_EKAW.pdf
[2] https://onlinelibrary.wiley.com/doi/full/10.1111/tgis.12510
[3] https://explosion.ai/demos/displacy-ent
[4] https://spacy.io/models/de
[5] https://figshare.com/articles/dataset/Learning_multilingual_named_entity_recognition_from_Wikipedia/5462500
[6] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger
[7] https://prodigy.io
[8] https://github.com/geoai-lab/EUPEG