Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Entity linking historical document OCR by combining Wikidata and Wikipedia

Kai Labusch, Clemens Neudecker

SWIB23 – 15th Semantic Web in Libraries Conference
11 – 13 September 2023, Berlin

Overview

- Introduction
- Named Entity Recognition
- Named Entity Linking / Disambiguation:
  - Why is it important?
  - Construction of knowledge base (KB)
  - Entity candidate lookup
  - Candidate evaluation
  - Ranking of candidates
- EL-based topic modelling
- Resources

Berlin State Library (SBB)

- Established 1661 in Berlin
- >33 million media objects
- >2.5 PetaBytes digital data
- Part of the Prussian Cultural Heritage Foundation
  https://staatsbibliothek-berlin.de/

- In-house Digitization Center since 2007
- Digital collections provide access to
  >212,000 digitized documents
  https://digital.staatsbibliothek-berlin.de/

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Named Entity Recognition

- BERT-based NER tagger
- Pre-trained for “Masked-LM” and “Next Sentence Prediction” tasks on historical text material of SBB digitized collections
- Supports PER, LOC and ORG entities
- German model: Trained on historical and contemporary German NER ground-truth
- French, English model: Trained on combined German, French, Dutch, and English NER ground-truth

Details → Kai Labusch, Clemens Neudecker and David Zellhöfer: BERT for Named Entity Recognition in Contemporary and Historic German

Named Entity Linking and Disambiguation

Why Entity Linking?

Dieselben Versuche , bei denen der elektrische Drache die Hauptrolle spielte , wiederholte der berühmte Lichtenberg in Göttingen ; in noch größerer Vollkom - menheit und mit aller Vorsicht aber ein Franzose , Namens de Romas zu Nerac .

Wikipedia	Wikidata	Confidence
Georg_Christoph_Lichtenberg	Q57554	0.50
Ludwig_Christian_Lichtenberg	Q1874282	0.18

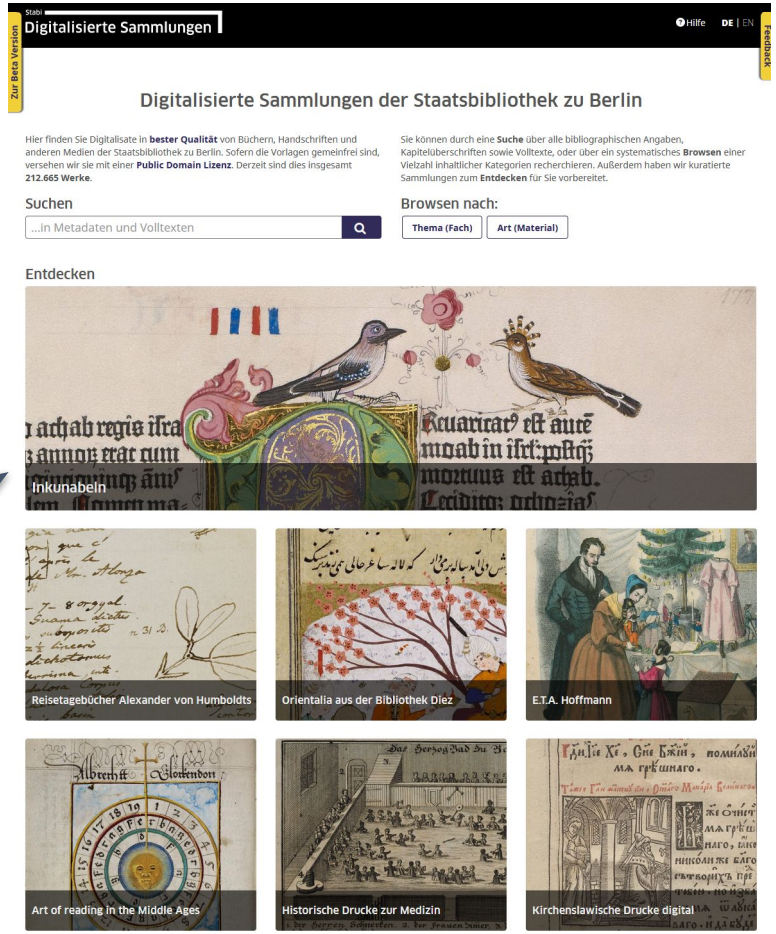
- Alexander von Lichtenberg (1880–1949), ungarischer Urologe
- Anna von Lichtenberg (1442–1474), Erbtochter der Herrschaft Lichtenberg
- Bernd Lichtenberg (* 1966), deutscher Drehbuchautor
- Bernhard Lichtenberg (1875–1943), deutscher Seliger und Gerechter unter den Völkern
- Betz von Lichtenberg († 1480), Großbailli
- Byron Kurt Lichtenberg (* 1948), US-amerikanischer Astronaut
- Carl Lichtenberg (1816–1883), deutscher Politiker
- Claudia Lichtenberg (* 1985), deutsche Radrennfahrerin
- Eleonora Jakowlewna Lichtenberg (* 1925), sowjetisch-russische Architektin
- Erik Lichtenberg, US-amerikanischer Agrarökonom
- Ernst Lichtenberg (* 1939), deutscher Politiker (CDU)
- Friedrich Lichtenberg (1801–1871), deutscher Politiker, MdL Hessen
- Friedrich David Lichtenberg (1774–1847), deutscher Apotheker
- Friedrich August von Lichtenberg (1755–1819), deutscher Politiker
- Georg Lichtenberg (Politiker) (1852–1908), deutscher Politiker, Bürgermeister von Linden
- Georg Lichtenberg (Landrat) (1886–1973), deutscher Verwaltungsjurist, Landrat von Neustadt
- Georg Christoph Lichtenberg (1742–1799), deutscher Physiker und Aphoristiker
- Gustav Wilhelm Lichtenberg (1811–1879), deutscher Jurist und Politiker, MdL Hessen

Why Entity Linking?

Musicians born in Rotterdam (the Netherlands)

- Items used: Rotterdam (Q34370) musician (Q639669)
- Properties used: instance of (P31) occupation (P106) subclass of (P279) place of birth (P19)

```
SELECT DISTINCT ?item ?itemLabel ?itemDescription where {
    ?item wdt:P106/wdt:P279* wd:Q639669 .
    ?item wdt:P19/wdt:P131* wd:Q34370 .
    SERVICE wikibase:label { bd:serviceParam wikibase:language "en,nl" }
}
```



Construction of KB for German, French and English

- EL system performs text comparisons that require sentences that refer to the entities in the KB

- Use SPARQL queries on Wikidata to find relevant entities that have a corresponding Wikipedia page
- Extract all sentences of Wikipedia where those entities have been referenced by Wikipedia authors

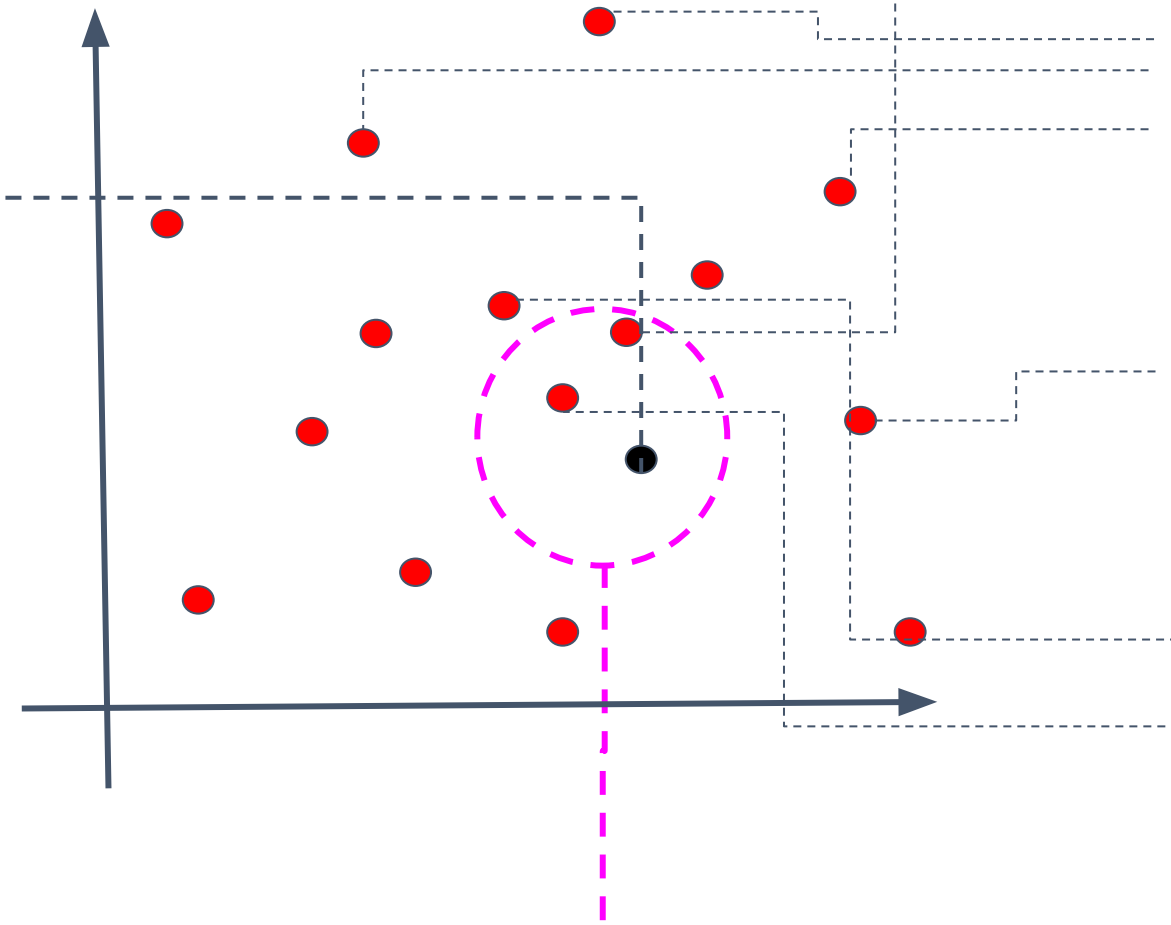
SPARQL - Example

```
SELECT ?organisation ?label ?sitelink ?gndid
WHERE
{
  # instance of any subclass of  fictional organisation
  ?organisation wdt:P31/wdt:P279* wd:Q14623646;rdfs:label ?label.
  FILTER(LANG(?label) = "fr").
  ?sitelink schema:about ?organisation.
  FILTER (CONTAINS(str(?sitelink), "fr.wikipedia.org")).
  OPTIONAL {
       ?organisation wdt:P227 ?gndid
  }
}
```

Entity Candidate Lookup

Dieselben Versuche , bei denen der elektrische Drache die Hauptrolle spielte , wiederholte der berühmte Lichtenberg in Göttingen ; in noch größerer Vollkom - menheit und mit aller Vorsicht aber ein Franzose , Namens de Romas zu Nerac .

- BERT embeddings stored in an approximate nearest neighbour index
- 100 random projection search trees
- Lookup of up to max_{cand} candidates with a distance less than Δ_l
- Angular distance measure



Approximate nearest neighbour search

Siegfried Schürenberg
Ludwig Christian Lichtenberg
Paul Isenberg
Manfred Lichtenfeld
Victoria Eugénie von Battenberg
Burkhard I. von Woldenberg
Julius Rodenberg
Bettina Blumenberg
Karl Pfeiffenberger
Werner Wollenberger
Hayim Greenberg
Adolph Scharenberg
Fritz Lichtenhahn
Johann Reinhard II. von Hanau-Lichtenberg
Georg Christoph Lichtenberg
Alexander Stuckenberg
Gertrud von Plettenberg
Michael Goldenberg
Anke Feuchtenberger
Peter Krukenberg
Reinhard von Hanau-Lichtenberg

Entity Candidate Evaluation

Dieselben Versuche , bei denen der elektrische Drache die Hauptrolle spielte , wiederholte der berühmte Lichtenberg in Göttingen ; in noch größerer Vollkom - menheit und mit aller Vorsicht aber ein Franzose , Namens de Romas zu Nerac .

?

Bei Georg Christoph Lichtenberg hörte er im Sommersemester 1796 Experimentalphysik und sehr wahrs Wintersemester Astronomie. In Göt

Im deutschsprachigen Raum wurde die Bezeichnungsweise von Franklin vermutlich vor allem durch Leonhard Euler bzw. Georg Christoph Lichtenberg verbreitet.[2]

Zu den namhaftesten Satirikern der Spätaufklärung zählen Georg Christoph Lichtenberg, der den kurzen, geschliffenen *Aphorismus* populär machte, und Jean Paul, dessen gesamtes Werk eine Neigung zur Satire zeigt. In

Jakob von Lichtenberg wurde 1416 als Sohn von Ludwig IV. von Lichtenberg und Markgräfin *Anna von Baden* (1399–1421), einer Tochter des Markgrafen Bernhard I. von Baden, geboren.

Um 1330 kam es zu einer ersten Landesteilung zwischen Johann II. von Lichtenberg, aus der älteren Linie des Hauses, und Ludwig III. von Lichtenberg. Dabei fiel

Das Pluszeichen als Symbol für eine positive elektrische Ladung oder den Pluspol ist abgeleitet von mathematischen Zeichen und geht auf den Mathematiker und Physiker Georg Christoph Lichtenberg zurück.

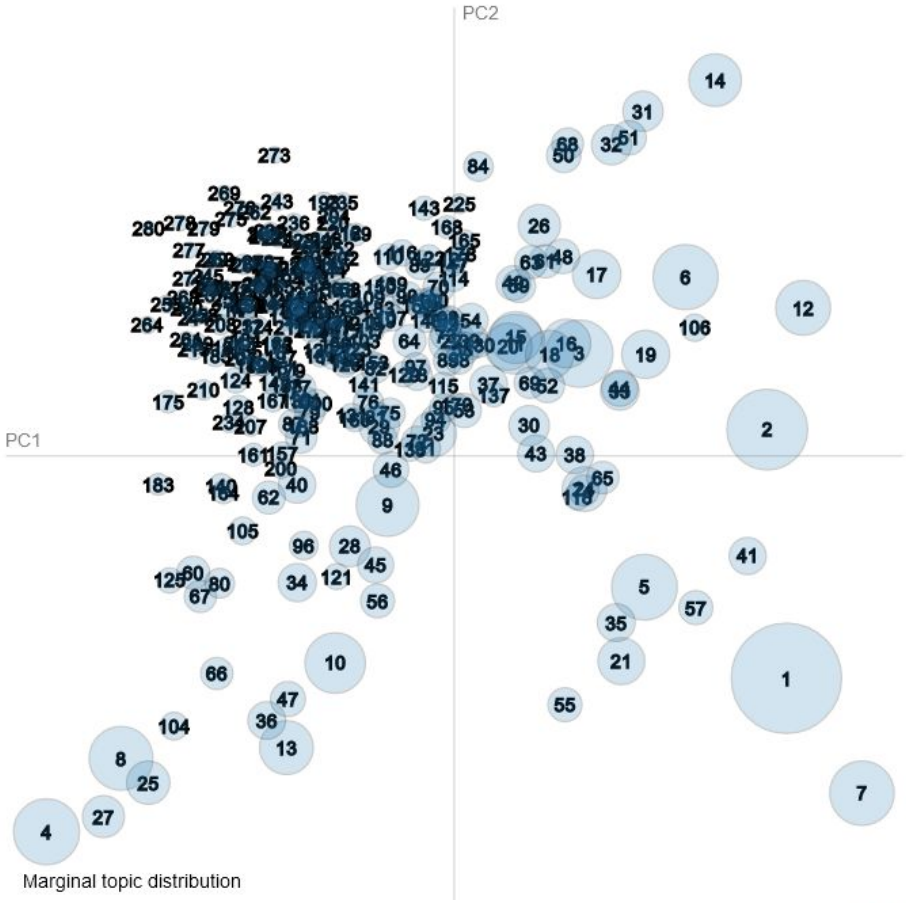
Entity Candidate Evaluation

- For each candidate consider up to 50 sentence pairs (A,B):
- Sentence A is part of text being subject to NEL
- Sentence B is part of Wikipedia and contains explicit link to candidate
- Purpose trained BERT-model determines probability of sentence (A,B) referring to the same entity
- Outcome is a set of matching probabilities per candidate
- Final ranking of candidates on the basis of matching probabilities by ranking model

Entity Candidate Ranking

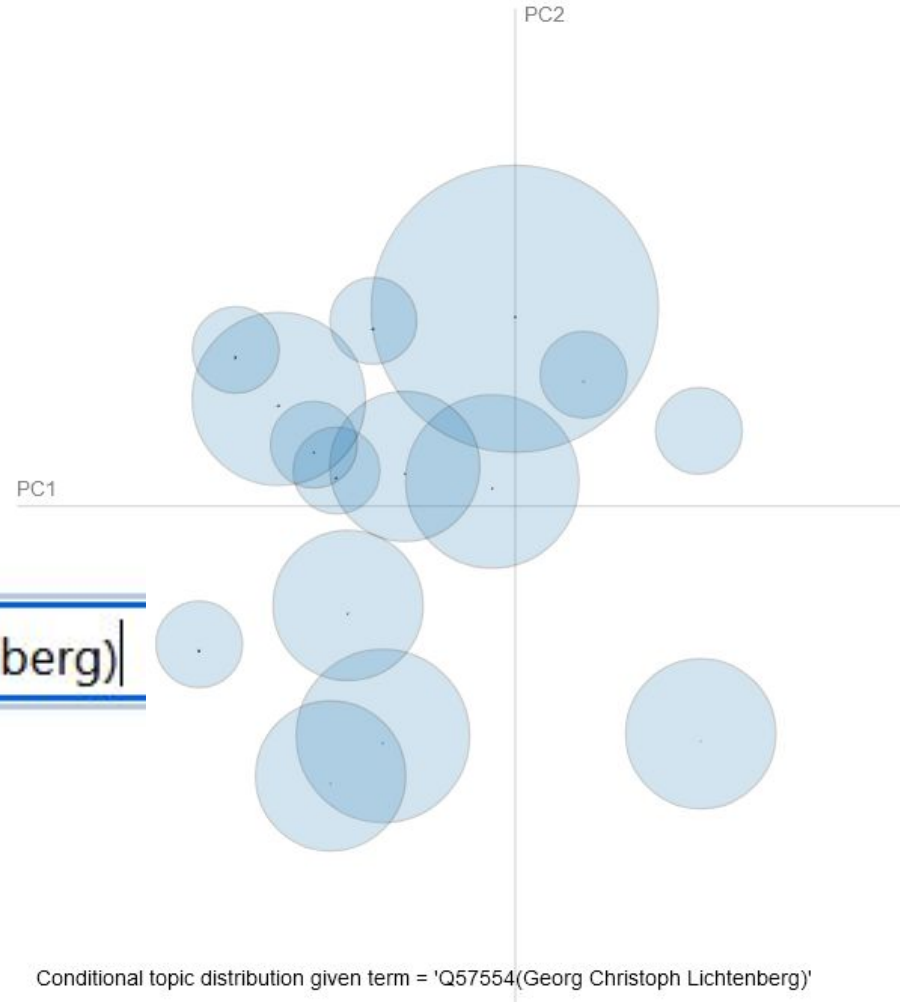
- Outcome previous steps: Set of matching probabilities per candidate
- Compute statistical features of sets of matching probabilities:
  - Mean, median, min, max, standard deviation, various quantiles
  - Ranking statistics over all candidates
- Add same statistical features for embedding similarities of lookup step
- Random forest model estimates overall matching probability per candidate
- Final output:
  - Sorted list of candidates that have matching probability > Δ_r
  - NIL: not implemented. Either list of sorted candidates or "-" if there is not any candidate with matching probability above Δ_r.

Topic Modeling

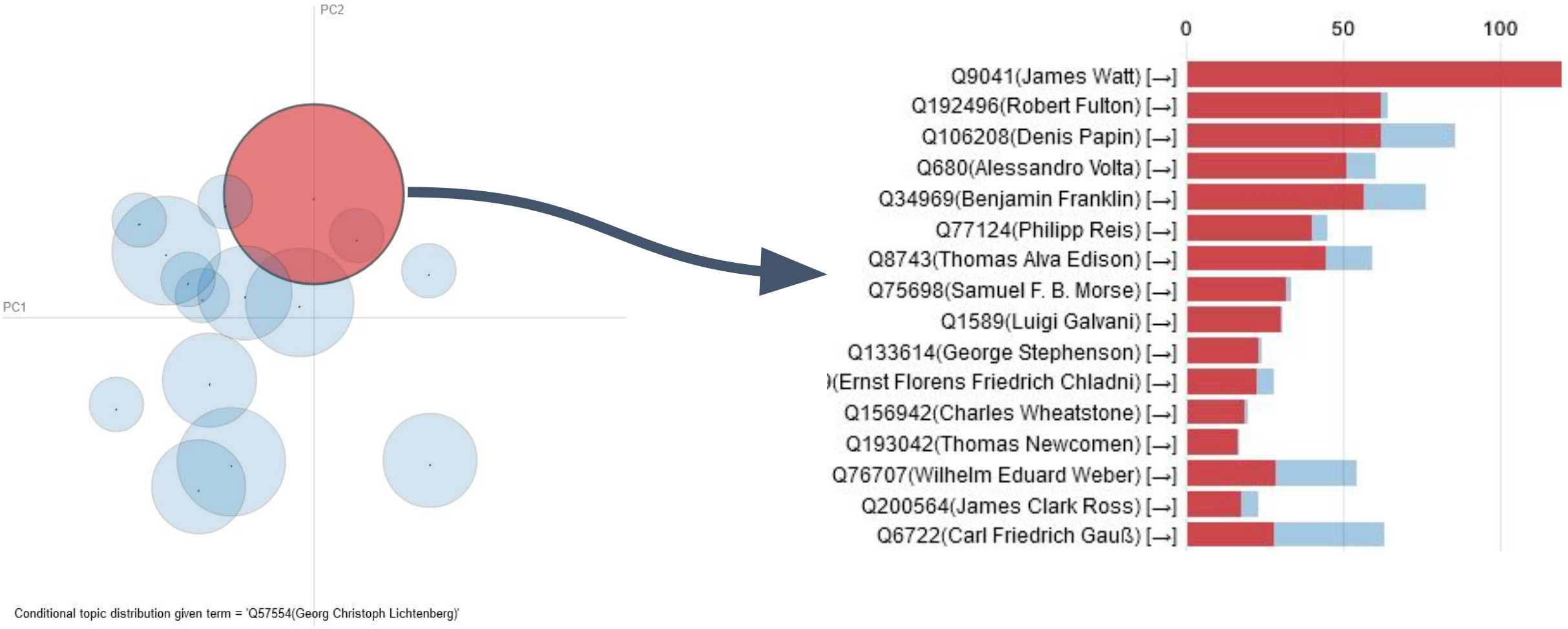


Search:

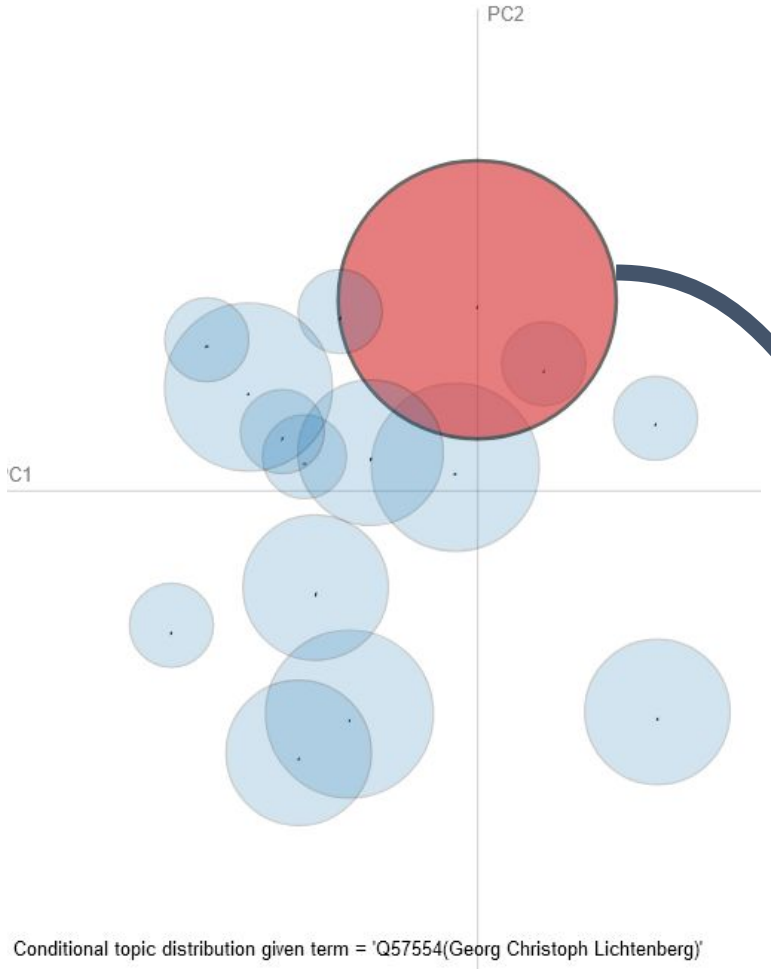
Q57554(Georg Christoph Lichtenberg)



Topic Modeling



Topic Modeling



Physikalisches Spielbuch für die Jugend ; Donath, Bruno; 1907 NER+EL Graphical Objects (68)

Die denkwürdigsten Erfindungen im neunzehnten Jahrhundert ; Thomas, Louis; 1883 NER+EL Graphical Objects (73)

Das Buch wunderbarer Erfindungen ; Thomas, Louis; 1860 NER+EL Graphical Objects (73)

Physikalisches Spielbuch für die Jugend ; Donath, B.; 1902 NER+EL Graphical Objects (75)

Die physikalische Technik oder Anleitung zur Anstellung von physikalischen Versuchen und zur Herstellung von physikalischen Apparaten mit möglichst einfachen Mitteln ; Frick, J.; 1876 NER+EL Graphical Objects (257)

Kantiana Hungarica ; Meltzl, Hugo; 1881 NER+EL

Die denkwürdigsten Erfindungen im neunzehnten Jahrhundert ; Thomas, Louis; 1895 NER+EL Graphical Objects (80)

Des deutschen Knaben Experimentirbuch ; Emsmann, H.; 1874 NER+EL Graphical Objects (173)

ML@SBB

- The NER and EL methods developed were applied successfully for the enrichment of 5,000,000 pages from SBB's digital collections
- Next steps
  - NER and EL results become integrated into the Digital Collections presentation portal, search, and discovery
  - NER and EL for use cases in the Digital Humanities
    → Historical Social Network Analysis (e.g. SoNAR IDH DFG-project)
  - NER and EL for use in subject indexing and classification
- BKM-funded project "Mensch.Maschine.Kultur" allows us to continue the work on NER and EL for another 3 years

Resources

- Full-texts of digitized collections (SQLite database):
  https://zenodo.org/record/7716098

- NER system: https://github.com/qurator-spk/sbb_ner/
  - NER models: https://huggingface.co/SBB/sbb_ner

- EL system: https://github.com/qurator-spk/sbb_ned/
  - EL models:
    https://huggingface.co/SBB/sbb_ned-de
    https://huggingface.co/SBB/sbb_ned-fr
    https://huggingface.co/SBB/sbb_ned-en
- EL knowledge bases (source code + sqlite databases):
  https://github.com/qurator-spk/sbb_knowledge-base
  https://zenodo.org/record/7767404 (German)
  https://zenodo.org/record/7773987 (English)
  https://zenodo.org/record/7773746 (French)

- Topic-Modelling:
  https://github.com/qurator-spk/sbb_topic-modelling

- Publications:

  BERT for named entity recognition in contemporary and historical German

  Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT.

  Entity linking in multilingual newspapers and classical commentaries with BERT

  Named Entity Linking mit Wikidata und GND–Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten

- Web-Demonstrators (NER+EL+Topic-Modeling):

  => https://ravius.sbb.berlin <=

Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Thank you for your attention!
Questions?

Kai Labusch, Clemens Neudecker
@sbb.spk-berlin.de

SWIB23 – 15th Semantic Web in Libraries Conference
11 – 13 September 2023, Berlin