

NLP in «Impresso – Media Monitoring of the Past»

Simon Clematide

Institut für Computerlinguistik
Universität Zürich

NLP Research Workshop Bern

Danke an M. Ehrmann, D. Sandoz, Ph. Ströbel für Material.

Was ist impresso?

EPFL C²DH University of Zurich

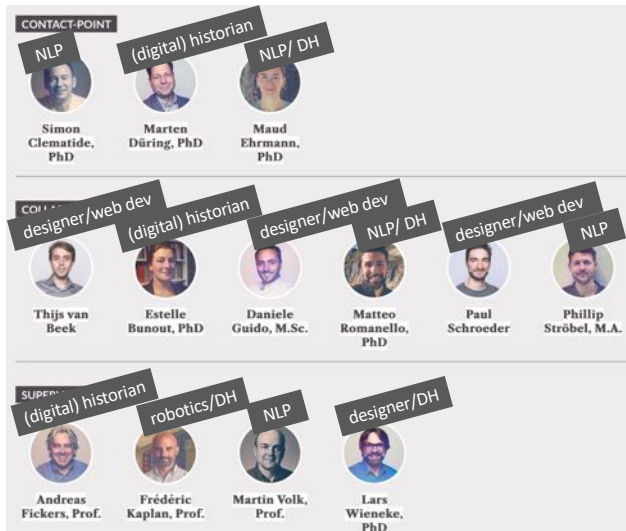
FNSNF Swiss Government Luxembourg Government infoctio.ch Unil University of Basel a NZZ LE TEMPS

impresso Media Monitoring of the Past

Critical content mining of 200 years of historical newspapers

Ein SNF-Sinergia-Projekt mit Beteiligung von Forschenden aus der Schweiz und Luxemburg

Wer macht mit? Eine transdisziplinäre Community



the team

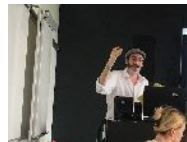
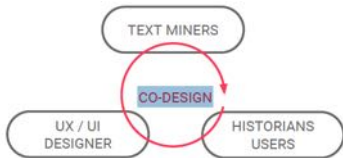
*+ a team of
historical advisors*

*+ a team of
associated historians*

Der “beschwerliche” transdisziplinäre Weg

5 challenges

How best accommodate text analysis research tools and their usage by humanities scholars?



Quelle: Maud Ehrmann

Eine Karikatur des Co-Designs: Q & A

NLP Researcher's question

What fancy NLP application do you want?

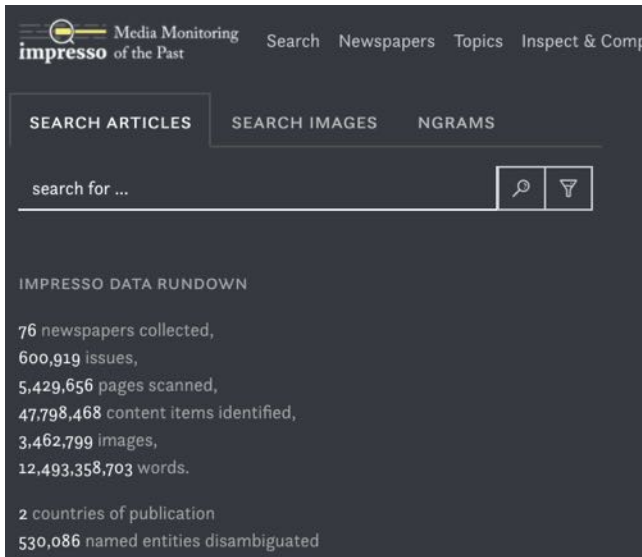
Historian's answer

Keyword search

NLP Researcher's answer

Ooooh, ok, I'll give you keyword search
with NLP-powered search facets, recommenders
and exploration capabilities

Document Retrieval



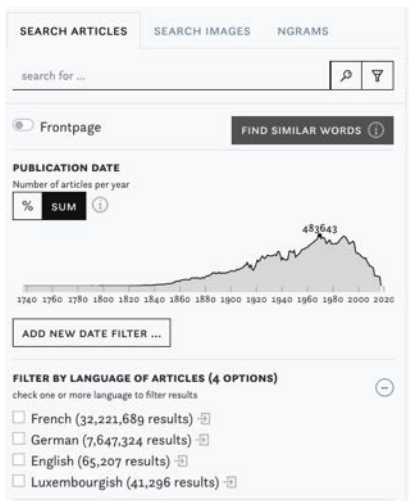
The screenshot shows the Impresso web application interface. At the top, there is a navigation bar with the logo "impresso" and the text "Media Monitoring of the Past". To the right of the logo are links for "Search", "Newspapers", "Topics", and "Inspect & Compare". Below the navigation bar, there are three tabs: "SEARCH ARTICLES" (which is selected), "SEARCH IMAGES", and "NGRAMS". Below the tabs is a search bar with the placeholder text "search for ...". To the right of the search bar are two icons: a magnifying glass and a funnel. Below the search bar, there is a section titled "IMPRESSO DATA RUNDOWN" which lists the following statistics:

- 76 newspapers collected,
- 600,919 issues,
- 5,429,656 pages scanned,
- 47,798,468 content items identified,
- 3,462,799 images,
- 12,493,358,703 words.
- 2 countries of publication
- 530,086 named entities disambiguated

Eine Web-Applikation für die Suche, Verwaltung, Kategorisierung und Vergleich von Zeitungsartikeln und Bildern.

Einiger Aufwand notwendig, um Bilder und Texte synchronisiert zu zeigen
(HistorikerInnen wollen die Quellen sehen...)

Language Identification (LID)



- ▶ Hohe Multilingualität in Luxemburgischen Zeitungen
- ▶ Mehrere Sprachen in einem Artikel
- ▶ Sprach-Information in Originaldaten nur teilweise vorhanden, und nicht immer zuverlässig
- ▶ Sprache beeinflusst die Indizierung (Stoppwörter, Normalisierung) durch SOLR
- ▶ Artikel mit vielen Namen und Ortschaften sind “schwierig”
- ▶ Welche Tools können zuverlässig (historisches) Luxemburgisch unterscheiden von Deutsch mit vielen OCR-Fehlern?

Et en avant les gars du Luxembourg!
The Luxembourg boys go forward . . . Down with the Heinies!



CRIVEMACHER

Lëtzeburger Zaldaten traine'eren an England
Soldats luxembourgeois à l'entraînement en Angleterre
Luxembourg soldiers training in England

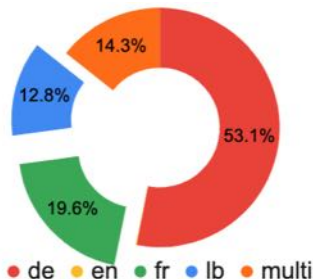
Quelle: ► Onsjongen 1945-03-03

Multilingualität innerhalb eines Dokuments

Fallstudie: Qualität von ursprünglicher Sprachidentifikation

Genauigkeit des Originals

Sampling von 1000 "offiziell" Luxemburgischen Artikeln, für die `langid` [LUI und BALDWIN 2012] eine andere Sprache vorhersagt.

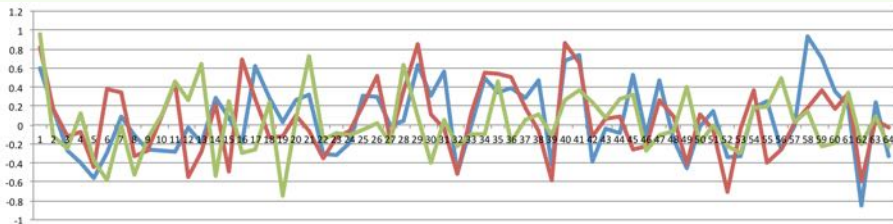


- ▶ `langid` Werkzeug "kennt" Luxemburgisch (lb) und wurde trainiert auf unterschiedlichen modernen Texttypen: JRC-Acquis, ClueWeb 09, Wikipedia, Reuters RCV2, Debian i18n
- ▶ [LUI und BALDWIN 2012] reportieren zwischen 88 und 99% Genauigkeit
- ▶ Genauigkeit auf unserem Sample: 79% (wenn 14% multilingual ignoriert werden)
- ▶ Einfaches Training auf eigenen Daten mit der fastText Textklassifikation^a erreicht mehr als 98% Genauigkeit

^a<https://fasttext.cc/docs/en/language-identification.html>

Word Embeddings (WE): Dense Vectors of Real Numbers

Which line represents which word? (“war”, “peace”, “tomato”)



Word Embeddings

- ▶ Continuous dense numeric word representations learned from raw text.
- ▶ Similar vectors mean similar words (cosine vector distance).
- ▶ Embeddings are a *perfect input* for numeric, neural machine learning methods.

Categorical Word Representations: One-Hot Encoding

The standard word representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: *hotel*, *conference*, *walk*

In vector space terms, this is a vector with one 1 and a lot of zeroes

$[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

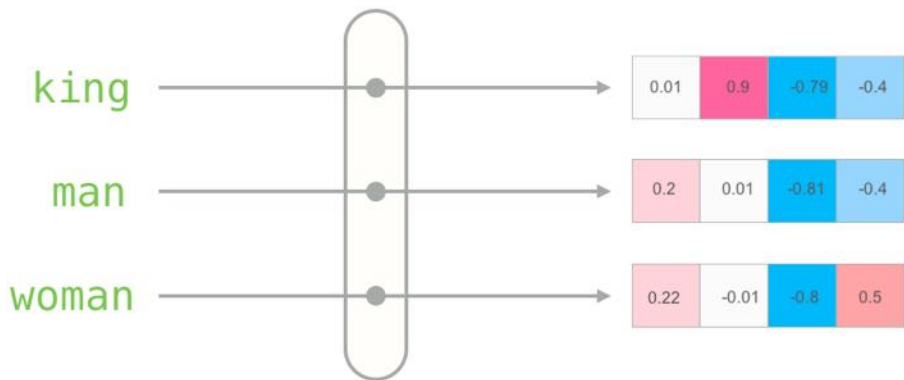
We call this a “one-hot” representation. Its problem:

motel $[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$ AND
hotel $[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$ = 0

Source: [SOCHER und MANNING 2013, 35]

Question: How ►similar are they distributionally?

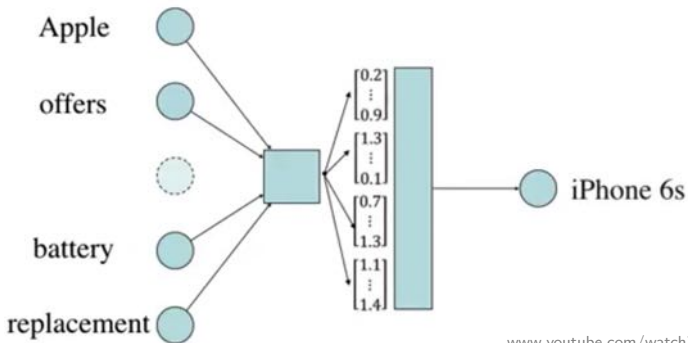
word2vec: Representation Learning



For a gentle introduction, see ► [The Illustrated Word2vec](#)

CBOW: Word Prediction Task

Given a context of n surrounding words, what is the probability of seeing the center word w ? [MIKOLOV et al. 2013]



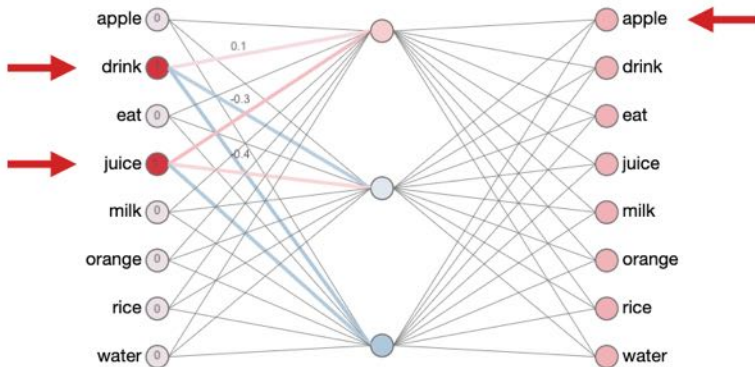
www.youtube.com/watch?v=aZarigloqXc

Pre-training: Neuronales Netzwerk lernt über grossen Textmengen gute Wahrscheinlichkeiten.

► wevi: Visualization of a CBOW Training Step

Context words: drink, juice

Training center word: apple

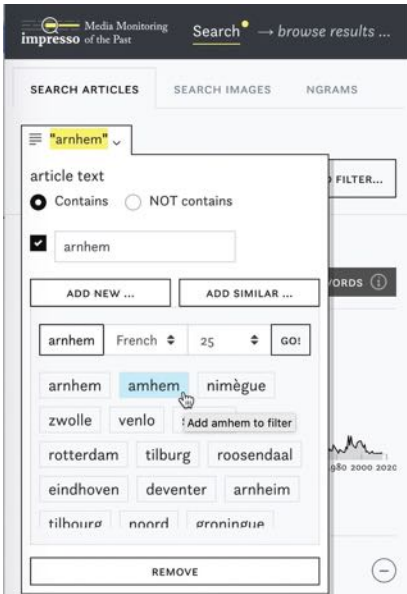


Source: ► wevi: word embedding visual inspector

Input/Output hat Dimensionalität von One-Hot-Vektoren: drink=[0,1,0,0,0,0,0,0]

Embeddings sind gelernte Gewichte zu "hidden" Layer: drink=[0.1,-0.3,-0.4]

Einsatz von Word Embeddings in impresso



- ▶ Gelernt mit fastText auf impresso Korpora
- ▶ Schlägt alternative relevante Suchterme vor
- ▶ Daten-getrieben aus den impresso-Korpora
- ▶ Semantische Ähnlichkeit
- ▶ Varianten aufgrund von OCR-Problemen
- ▶ Varianten aufgrund von Multilingualität
- ▶ Coming soon: Bilinguale Embeddings ("état/Staat") für sprachübergreifende Ähnlichkeit

Grenzen von word2vec-artigen Embeddings und Ausblick

- ▶ Jedes Wort hat genau 1 Vektor. Egal in welchem Kontext es benutzt wird.
- ▶ Der Vektor von mehrdeutigen Wörtern (**Bank**) repräsentiert die Mischung von Bedeutungen.
- ▶ Neuer mächtiger Trend: kontextualisierte Word Embeddings (ELMo, BERT, etc.)
- ▶ Jedes Wortvorkommen bekommt in abhängig vom Kontext einen anderen Vektor.
- ▶ Distributionalismus pur:
"You shall know a word by the company it keeps!" (J. R. Firth 1957)
- ▶ Aber auch mit "statischen" word2vec Embeddings lassen sich NLP-Probleme gut lösen...

Keyphrase-Extraktion (KPE) mit Embeddings: EmbedRank [BENNANI-SMIREs et al. 2018]

2 Ziele der Keyphrase-Extraktion

- ▶ Informativeness: Keyphrase ist repräsentativ/informativ für Dokument
- ▶ Diversity: Wenn mehr als 1 Keyphrase extrahiert wird, sollten sie unterschiedlich sein voneinander.

Unsupervisierte EmbedRank-Methode

- ▶ Keyphrase-Kandidaten mittels POS-Tag Sequenzen über Inhaltswörtern (ADJ* NOUN+) identifizieren (z.B. mit SpaCy)
- ▶ Einbettung von Kandidaten und Dokument (nur Adjektive und Nomen) in denselben niedrig-dimensionalen Vektor-Raum unter Benutzung von word2vec Word Embeddings
- ▶ Ranking der Kandidaten nach Informativität (Ähnlichkeit von Kandidat und Dokument) und Diversität

KPE für Französisch: Verschiedene Ansätze

EXP-1950-12-27-a-i0004

1'910 characters, 1950



Le gouvernement français engage aujourd'hui la bataille du budget militaire

e M. Plevén a toutes les chances de l'emporter Notre correspondant de Paris nous téléphone : Alors que le président du Conseil pressait très vivement l'Assemblée nationale de commencer l'examen du budget spécial de réarmement, celle-ci, par 302 voix contre 2 S 4, a refusé de suivre M. René Plevén et a fixé à aujourd'hui seulement l'examen des comptes de la défense nationale. Cette manifestation gratuite de mauvaise humeur, sans grande importance politique, soulignons-le, marque cependant avec beaucoup de clarté l'état d'esprit du parlement actuel auquel le cabinet demande, à quelque neuf mois des élections générales, de voter 105 milliards d'impôts nouveaux destinés à financer les dépenses extraordinaires résultant du programme de réarmement. Depuis plus d'une semaine déjà, la discussion a commencé en privé au sein de la commission des finances, mais alors que le gouvernement réclame ces fameux 165 milliards, les commissaires ont fait savoir qu'à leur sens, les charges fiscales ne devraient pas excéder 111 milliards, le reste, c'est-à-dire 54 milliards, devant être trouvé par des économies massives réa- Usées à la fois sur le budget et sur, l'ensemble des autres dépenses militaires. Le débat public qui va s'ouvrir aujourd'hui durera au maximum deux jours. Vingt-sept orateurs sont inscrits auxquels M. Jules Moch, ministre de la défense nationale, et M. René Plevén, chef du gouvernement ! donneront la réplique en appuyant leurs propos de l'arme atomique de la question de confiance. Le cabinet l'emportera, chacun en demeure persuadé dans les couloirs, mais uniquement en raison de la gravité de la situation internationale. S'il n'y avait pas, en effet, les lourdes menaces qui s'accumulent à l'horizon, on pourrait prédire sans grand risque au ministère Plevén la plus funeste des aventures.

ERFULL

budget militaire

gouvernement

milliards

ministère plevén

réarmement

ERWP

budget

défense nationale

gouvernement

milliards

réarmement

SR

autres dépenses militaires

auxquels m. jules moch

budget militaire m. plevén

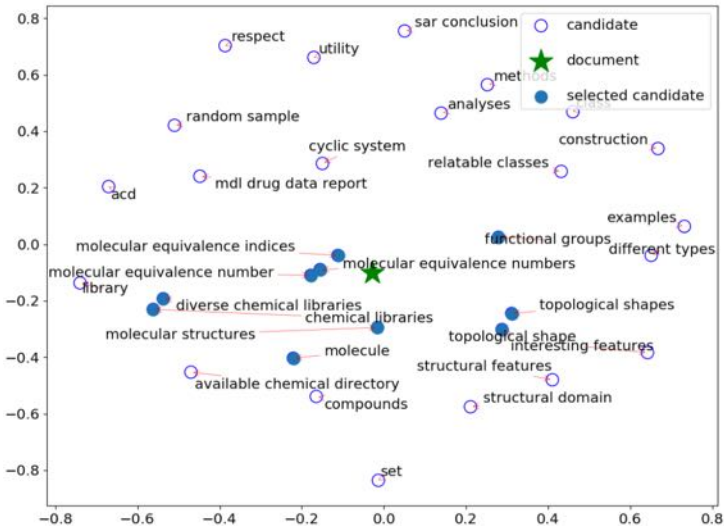
m. \ rené plevén

m. rené plevén

SR = Single Rank graphbasiert ; ERFULL = Embed Rank mit Embeddings von impresso Korpus; ERWP = Embed Rank mit Embeddings von Wikipedia

Quelle: BA-Arbeit von Dominique Sandoz

Ähnlichkeitsraum von Kandidaten und Dokument (in 2D)

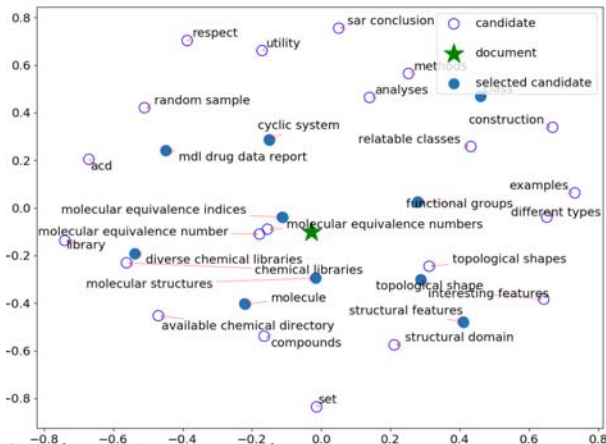


Textbasis:
Abstract von
Artikel mit Titel
“Using molecular
equivalence
numbers to
visually explore
structural
features that
distinguish
chemical
libraries”
Hohe Ähnlichkeit
von Keyphrases
(“topological
shape(s)”)

Quelle: [BENNANI-SMIREs et al. 2018]

Ähnlichkeitsraum mit MMR-Diversität

MMR (Maximal Margin Relevance)



Quelle: [BENNANI-SMIREs et al. 2018]

Füge iterativ denjenigen Kandidaten j mit dem grössten Wert der Differenz $A-B$ hinzu.

A ist die Ähnlichkeit zwischen j und dem Dokument.

B ist die höchste Ähnlichkeit zwischen j und einem bereits akzeptierten Kandidaten.

Diversität von Keyphrases mit Embed Rank

En l'honneur de M. Garbani

• De Berne, le 23 juin :

Jeudi soir, les sociétés tessinoises ont nisé, au Buirgerimis, une manifestation en l'honneur du nouveau juge fédéral, M. Gacbani. La nombreuse colonie tessinoise assista presque complète à cette manifestation, au cours de laquelle des discours furent cés par le Dr Gianini, au nom de la colonie, par le conseiller national Vigizzi, au nom de la députation * tessinoise aux Chambres, et par M. le conseiller fédéral Motta. Très ému, M. Garbani-Nerini répondit, déclarant avoir donné son nom pour la réalisation d'une des plus chères revendications du peuple tessinois. Tous les orateurs furent très applaudis.

ERFULL

conseiller national manifestation
nombreuse colonie tessinoise peuple tessinois vigizzi

ERWP

conseiller fédéral manifestation nom
nombreuse colonie tessinoise peuple tessinois

SR

m. garbani m. garbani • nombreuse colonie tessinoise
peuple tessinois sociétés tessinoises

Quelle: BA-Arbeit von Dominique Sandoz

Topic Modeling (TM): Soft-Clustering von Texten

Text-Clustering

- ▶ Unsupervisiertes Gruppieren von Text-Dokumenten
- ▶ Keine vom Menschen vorgegebenen Klassen

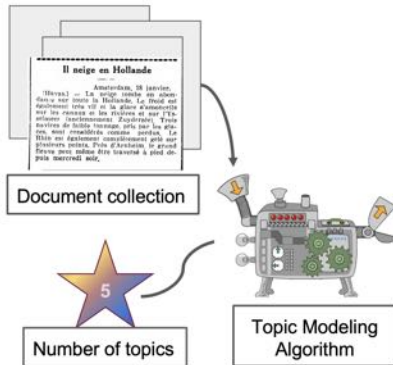
Clustering-Prinzipien

- ▶ Maximiere Ähnlichkeit von Dokumenten innerhalb eines Clusters.
- ▶ Minimiere Ähnlichkeit von Dokumenten in unterschiedlichen Clustern.

Topics interpretieren

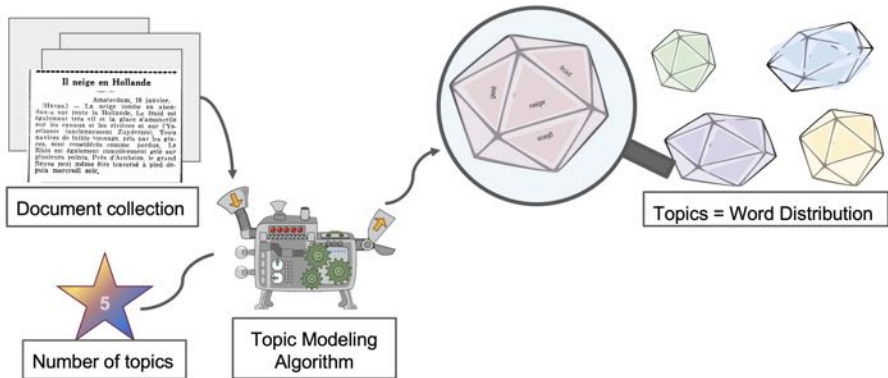
neige>montagne>ski>hiver>vallée>région>tourisme>station>saison>sport>route>fond>chalet>lac>
temp>altitude>pied>glace>sommet>camp>village>année>été>chalet>col>avalanch
ascension>forêt>ber>intemps>chemin>glacier>cabane>descente>ser</p></div>
<div data-bbox="885 337 938 354" data-label="Text">(MORE ...)</div>
<div data-bbox="541 445 936 648" data-label="List-Group">▶ Die häufigsten Wörter eines Topics auflisten.▶ Texte mit 1 dominanten Topic "verstehen".</div>
<div data-bbox="519 709 727 755" data-label="Section-Header"><h2>Soft-Clustering</h2></div>
<div data-bbox="519 772 934 864" data-label="Text"><p>Ein Dokument kann zu mehr als 1 Topic gehören.</p></div>
<div data-bbox="17 963 253 990" data-label="Page-Footer">NLP Research Workshop Bern</div>
<div data-bbox="396 963 477 989" data-label="Page-Footer">14.8.2020</div>
<div data-bbox="517 963 977 992" data-label="Page-Footer">NLP in «Impresso – Media Monitoring of the Past» 23 / 43</div>

Topic Modeling in a Nutshell



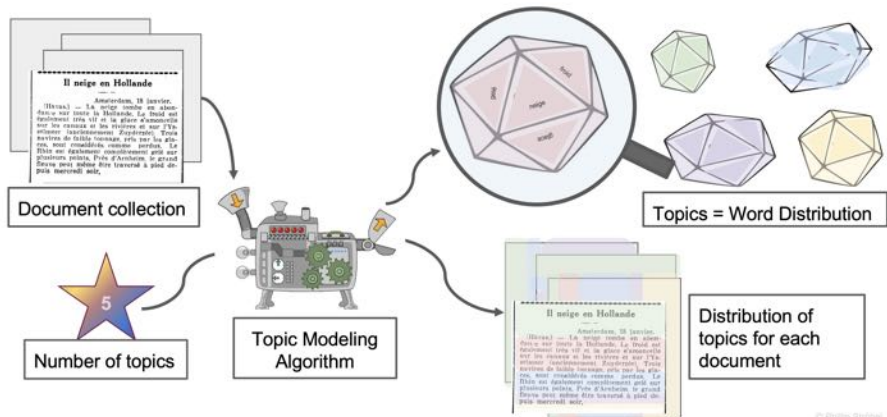
© Philippe Stroh

Topic Modeling in a Nutshell



© Philip Schrödel

Topic Modeling in a Nutshell



© Philip Brödel

Topic Modeling: Intuitiv erklärt

Dokumente als Bag-Of-Word (Multimenge von Wörtern)

Welche Wörter kommen wie oft vor?

Was sind Dokumente? Eine Wahrscheinlichkeitsverteilung von Topics

Typischerweise ein Haupt- und mehrere Neben-Themenbereiche, welche anteilmässig für den Inhalt charakterisieren.

Was sind Topics? Eine Wahrscheinlichkeitsverteilung von Wörtern

Jedes Topic ist ein grosser unfairer Würfel mit vielen Seiten, wo jeweils ein Wort drauf steht.

Generatives LDA Topic Modeling intuitiv

Würfle für jedes Dokument anteilmässig mit jedem Topic-Modelingwürfel die Wörter, die letztlich den Bag-Of-Word des Dokuments ausmachen.

Dokumentähnlichkeit bezüglich Topic-Verteilung

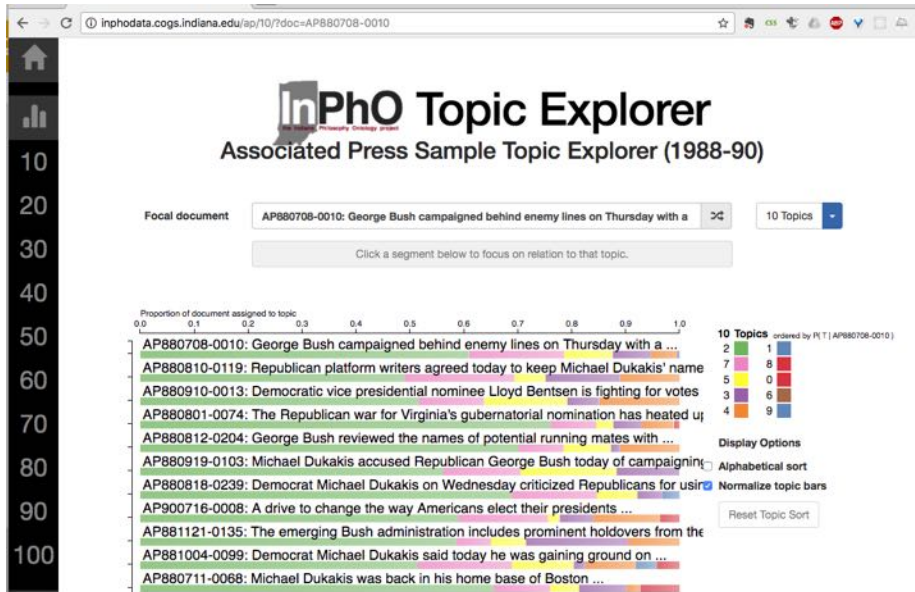


Illustration zu Topic Modeling

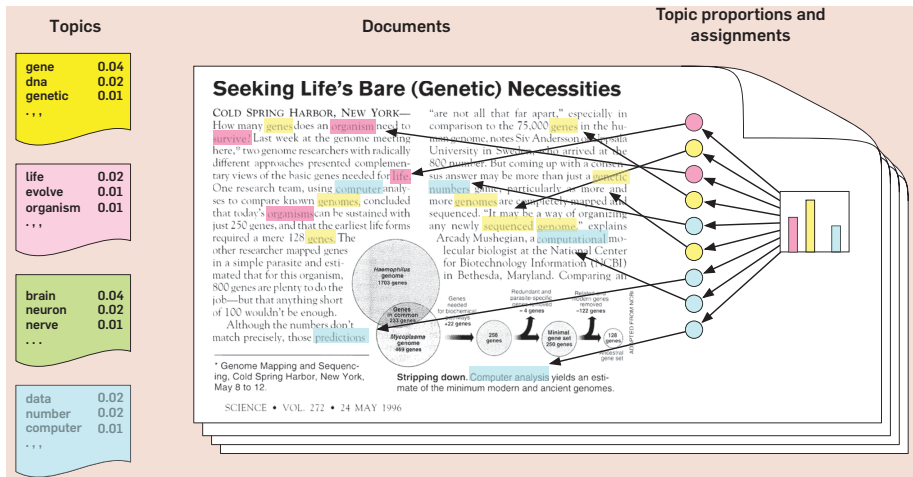


Abbildung: Zutaten für probabilistisches *Topic Modeling* nach [BLEI 2012, 78]

Topic Modeling in impresso

- ▶ Texte werden auf Nomen reduziert und lemmatisiert.
- ▶ Konkret: Part-of-Speech Tagging mit SpaCy und Verwendung von zusätzlichen morphologischen Ressourcen zur Reduktion auf Grundformen
- ▶ Motivation 1: Interpretierbarkeit: Nomen lassen sich in Isolation gut verstehen.
- ▶ Motivation 2: Datenreduktion: Effizientere Berechnung (Rechenzeit und Speicher) unserer “big impresso data”
- ▶ 100 Topics pro Sprache
- ▶ Tooling: Mallet-Werkzeug
- ▶ Millionen von Artikeln können auch Mallet überfordern. Lösung: Topic Modeling auf Sample mit nachfolgender Topic Inference auf allen Dokumenten

Topics als Suchfacetten in impresso

SEARCH ARTICLES SEARCH IMAGES NGRAMS

SEARCH FOR: "arnhem OR arnhem(2 more)"

search for ... ADD FILTER...

France (984 results) ☐
 Eindhoven (910 results) ☐
 Basel (804 results) ☐
 Lage (776 results) ☐
 MORE... (5385 MORE OPTIONS)

FILTER BY TOPIC (227 OPTIONS) ☐
 filter results if one of 3 selected topic applies

☒ fr match - équipe - ligue - club - saison (1,333 results) ☐
☐ de teil - zeitung - frage - negierung - weise (715 results) ☐
☐ fr front - armée - guerre - ennemi - nord (706 results) ☐
☐ de berlin - münchen - wien - mark - frankfurt (587 results) ☐
☐ de schweiz - frankreich - deutschland - italien - schwede (583 results) ☐
☒ fr match - jeu - but - équipe - minute (515 results) ☐
☒ fr bat - finale - tour - tournoi - match (493 results) ☐
☐ de stadt - schweiz - bern - genf - teil (480 results) ☐
☐ de london - armee - angriff - luftwaffe - general (412 results) ☐
☐ de don - and - min - sage - uli (385 results) ☐
 MORE... (217 MORE OPTIONS)

GROUP BY ARTICLE ORDER BY RELEVANCE DISPLAY AS LIST TILES

6,840 articles found containing **arnhem** or **arnhem** or **arnheim** or **arnheim** COMPARE ... SAVE / EXPORT ...

Algemene Kunstzijde Unie N. V.
Journal de Genève MONDAY, APRIL 29, 1968 – p.2
 Algemene Kunstzijde Unie N. V. (AKU), Arnhem (Pays-Bas) L'assemblée générale des actionnaires se tiendra à Amsterdam, le mercredi 25 mai 1968, à 11 h.
 LOCATIONS: Suisse, Moselle
 Algemene Kunstzijde Unie N. V. (AKU), Arnhem (Pays-Bas) L'assemblée générale des actionnaires
 être consultés par les intéressés aux bureaux de la société, Velperweg 76, Arnhem. Les actions peuvent
 Le conseil de surveillance Arnhem, le 29 avril 1968 ASSA 679 / 2
 VIEW ADD TO COLLECTION ...

AIV 11 BH1 11 V
Gazette de Lausanne WEDNESDAY, MAY 21, 1862 – p.4
 AIV 11 BH1 11 V 11 B pour la mi-août, dans un pensionnat de UI WJH&E UE / J demoiselles, près d'Arnhem (Hollande), une sous-maitresse de la Suisse
 LOCATIONS: Suisse, Moselle
 d'Arnhem (Hollande), une sous-maitresse de la Suisse française, religion réformée, capable d'enseigner
 à Mlle Wildrick, à Osterbeck, près d'Arnhem, Hollande.
 VIEW ADD TO COLLECTION ...

Retraite alliée à Arnhem...
Le Peuple, La Sentinelle 1 2 3 4 ... p.4

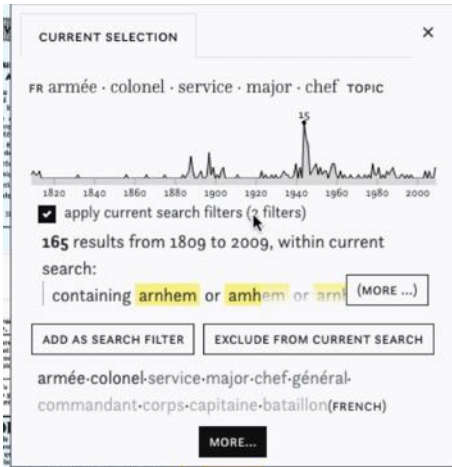
Topics aus dem Bereich Sport, Krieg, Städte sind prominent

Topics ausschliessen

The screenshot shows the 'impresso' Media Monitoring of the Past interface. At the top, there's a search bar with the text '"arnhem OR amhem(2 more)"'. Below the search bar, there are tabs for 'SEARCH ARTICLES', 'SEARCH IMAGES', and 'NGRAMS'. A dropdown menu is open, showing 'filter by topic' with two options: 'Containing' (radio button) and 'NOT containing' (radio button). Under 'NOT containing', there are three checked topics: 'FR match · équipe · ligue · club · saison', 'FR match · jeu · but · équipe · minute', and 'FR bat · finale · tour · tournoi · match'. At the bottom of the dropdown, there are buttons for 'APPLY CHANGES' and 'REMOVE'. A link 'MORE... (2,527 MORE OPTIONS)' is also visible.

Suchverfeinerung über positive oder negative Filter

Suchanfrage und Topics auf Zeitachse überblicken



Wie verteilen sich die Treffer?
Wo gibt es Peaks?

Topic-basierter Artikel-Recommender: Such-Treffer

The screenshot displays the Impresso Media Monitoring interface. The top navigation bar includes 'Media Monitoring of the Past', a search bar, and links for 'Newspapers', 'Topics', 'Inspect & Compare', 'Text reuse', and 'Collections'. The user is logged in as 'Simon Clemente'.

The main search area shows 'SEARCH ARTICLES' with filters for 'SEARCH IMAGES' and 'NGRAMS'. The search term is 'arnhem', and the topic filter is 'match - équipe - ligue - club - saison'. The results are grouped by 'ARTICLE' and ordered by 'RELEVANCE'. There are 1,182 articles found.

On the left, the 'FILTER BY LOCATION' section shows 1,415 options, including Eindhoven (675 results), Belgium (430 results), Enschede (411 results), Glasgow (402 results), Sparta (347 results), Vienne (318 results), Milan (302 results), France (288 results), R.S.C. Anderlecht (286 results), and Tilburg (260 results). The 'FILTER BY TOPIC' section shows 100 options, including 'match - équipe - ligue - club - saison' (1,182 results), 'fr match - jeu - but - équipe - minute' (324 results), 'fr bat - finale - tour - tournoi - match' (305 results), 'fr armée - colonel - service - major - chef' (190 results), 'fr monde - concours - record - sport - suisse' (144 results), and 'fr course - monde - place - temps - étape' (125 results).

The search results list includes:

- FOOTBALL: Bellinzzone promu en LNA** (Gazette de Lausanne) - MONDAY, MAY 12, 1980 - p.5. The article discusses the promotion of Bellinzzone to the LNA and mentions the FC Sion.
- Bellinzzone promu en LNA Désormais** (Journal de Genève) - MONDAY, MAY 12, 1980 - p.7. The article discusses the promotion of Bellinzzone to the LNA and mentions the FC Sion.
- 25.10.1980** (La Liberté) - MONDAY, OCTOBER 30, 1980 - p.25. The article discusses the promotion of Bellinzzone to the LNA and mentions the FC Sion.

Facettierte Suche nach “Arnhem” mit Topic Fussball

Topic-basierter Artikel-Recommender I

Gazette de Lausanne — Monday, May 12, 1980 ⏮ p.5 ⏭

[Article](#) [FR](#) [p.5](#) [FACSIMILE](#) [TRANSCRIPT](#) [FULLSCREEN: OFF](#)

Alkmaar, PSV Eindhoven, FC Utrecht et Twente Enschede seront engagés en Coupe de l'UEFA. Enfin, Vitesse Arnhem et le FC Haarlem ont été relégués en deuxième division.

Similar Articles

Béroche et Châtelard en verve
L'Express [FR](#) [FRIDAY, JUNE 2, 1972](#) — p.19

Béroche et Châtelard en verve IV LIGUE NEUCHÂTOISE | LES FBNALES ONT DÉBUTÉ Le championnat de IVe ligue est terminé, seuls les champions de groupes [article](#) | [short text](#) p.19

LOCATIONS: La Chaux-de-Fonds [FR](#), Vevey [FR](#), Zürich [FR](#), Sion, Switzerland [FR](#)

PEOPLE: Floria Lasky [FR](#)

TOPICS IN COMMON ⓘ

[FR match · équipe · ligue · club · saison](#)

Le retour de trois anciens
L'Impartial [FR](#) [THURSDAY, SEPTEMBER 2, 1982](#) — p.15

Le retour de trois anciens Pour la rencontre Suisse-Bulgarie Karl Grob : un retour en équipe nationale après six ans d'absence. (Photo asl) Mardi proc [article](#) | [short text](#) p.15

LOCATIONS: Saint Gall [FR](#), Switzerland [FR](#), Brigitta Bulgari [FR](#), Zürich [FR](#), Sweden [FR](#), FC Sion [FR](#), 2009 Finnish Cup [FR](#), Norway [FR](#), Nottingham [FR](#), SC Bastia [FR](#), Denmark [FR](#), Nuremberg [FR](#)

PEOPLE: Lucien Favre [FR](#), Karl Engel [FR](#), Raimondo Ponte [FR](#), Charles George Herbermann [FR](#), André Egli [FR](#), Heinz Hermann [FR](#), Umberto Barberis [FR](#), René Botteron [FR](#), Erni Maissen [FR](#), Claudio Sulser [FR](#), Roger Wehrli [FR](#), Martin Andermatt [FR](#), Markus Tanner [FR](#), Beat Sutter [FR](#)

TOPICS IN COMMON ⓘ

[FR match · équipe · ligue · club · saison](#)

Cinq formules en 14 ans
L'Express [FR](#) [THURSDAY, SEPTEMBER 7, 1989](#) — p.41

Cinq formules en 14 ans Depuis 1976, soit durant les 14 dernières saisons, le championnat de Suisse de ligue A a connu pas moins de cinq formules diff [article](#) | [short text](#) p.41

LOCATIONS: Suisse, Moselle [FR](#)

TOPICS IN COMMON ⓘ

[FR match · équipe · ligue · club · saison](#)

Ähnliche Dokumente (aufgrund der Topic-Distribution)

Topic-basierter Artikel-Recommender II

Journal de Genève — Wednesday, April 18, 1945

p.2

FR | p.2

FACSIMILE

TRANSCRIPT

FULLSCREEN: OFF

Dans le nord-est de la Forêt-Noire, l'ennemi a poursuivi ses efforts pour forcer le passage des montagnes et ouvrir les routes vers le sud. De combats se sont déroulés des deux côtés de Lichenzell et au nord de Freudenstein. L'ennemi a perdu de nombreux hommes.

Les troupes françaises attaquant vers le sud-est et le sud au delà de Laufenburg, dans la plaine du Rhin, ont été immobilisées au sud-est de la ville et à l'ouest de Lahr.

faiblement leurs bases de départ, grâce à la ténacité de nos grenadiers et à leurs contre-attaques continuelles.

Dans le nord-est de la Forêt-Noire, l'ennemi a poursuivi ses efforts pour forcer le passage des montagnes et ouvrir les routes vers le sud. De combats se sont déroulés des deux côtés de Lichenzell et au nord de Freudenstein. L'ennemi a perdu de nombreux hommes.

Les troupes françaises attaquant vers le sud-est et le sud au delà de Laufenburg, dans la plaine du Rhin, ont été immobilisées au sud-est de la ville et à l'ouest de Lahr.

Similar Articles

A L'OUEST D'AIX-LA-CHAPELLE Les Alliés s'emparent des défenses extérieures...
Journal de Genève THURSDAY, SEPTEMBER 14, 1944 ~ p.10

A L'OUEST D'AIX-LA-CHAPELLE Les Alliés s'emparent des défenses extérieures de la ligne Siegfried o. G. Eisenhower, 14. — D'Exchange : Des rapports non article | short text p.10

LOCATIONS Malmédy, Luxembourg, Thionville, France, Belfort, Vesoul, Mosselle, AS Nancy, Metz, Belgium, Meuse

PEOPLE Les Alliés, John Eisenhower

TOPICS IN COMMON

FR front · armée · guerre · ennemi · nord

Les Russes sont prêts à franchir le Dniepr
L'Express FRIDAY, SEPTEMBER 24, 1943 ~ p.1

Les Russes sont prêts à franchir le Dniepr L'armée rouge développe sa grande offensive en Ukraine Les canons soviétiques tirent déjà sur la rive occidentale | short text p.1

LOCATIONS Ukraine, Poltava, Smolensk, Moscow, Rostov, Novomoskovsk, Russia, Russia, Gomel, Berlin

TOPICS IN COMMON

FR front · armée · guerre · ennemi · nord

Les troupes britanniques et canadiennes ont déclenché hier une grande [...]
L'Express TUESDAY, JULY 31, 1944 ~ p.1

Les troupes britanniques et canadiennes ont déclenché hier une grande attaque au sud-ouest de la ville de Caen De violents combats font rage au sud de l'article | short text p.1

LOCATIONS Caen, Hubert Montgomery, Berlin

PEOPLE William Lindsay Osteen, Sr., United Kingdom, Les Canadiens, Les Anglais

TOPICS IN COMMON

FR front · armée · guerre · ennemi · nord

Ähnliche Dokumente mit “Arnhem” und Topic Krieg

Named Entity Recognition (NER) and Linking (NEL)



recognition: detecting boundaries

classification: assignment of a semantic category

disambiguation/linking: to a unique reference

Quelle: Maud Ehrmann

impresso benutzt momentan ein regelbasiertes System für NER und das AIDA-System für NEL.

Fortschritt: Wissenschaftlicher NLP-Wettbewerb (Shared Task) zu NER und NEL (CLEF HIPE 2020) mit impresso Daten.

NER-Performance in HIPE 2020 auf historischen Daten

(a) Literal	French						German					
	Strict			Fuzzy			Strict			Fuzzy		
	P	R	F	P	R	F	P	R	F	P	R	F
CISTERIA	-	-	-	-	-	-	.745	.578	.651	.880	.683	.769
EHRMAMA	.793	.764	.778	<u>.893</u>	.861	.877	.697	.659	.678	.814	.765	.789
ERTIM	.435	.248	.316	.604	.344	.439	-	-	-	-	-	-
INRIA	.605	.675	.638	.755	.842	.796	-	-	-	-	-	-
IRISA	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-
L3I	.831	.849	.840	.912	.931	.921	.790	.805	.797	<u>.870</u>	.886	.878
LIMSI	<u>.799</u>	<u>.829</u>	<u>.814</u>	.887	<u>.909</u>	<u>.898</u>	-	-	-	-	-	-
NLP-UQAM	.705	.634	.668	.828	.744	.784	-	-	-	-	-	-
SBB	.530	.477	.502	.765	.689	.725	.499	.484	.491	.730	.708	.719
SINNER	.788	.802	.795	.886	.902	.894	.658	.658	.658	.775	<u>.819</u>	<u>.796</u>
UPB	.693	.686	.689	.825	.817	.821	.677	.575	.621	.788	.740	.763
UVA-ILPS	.656	.719	.686	.794	.869	.830	.499	.556	.526	.689	.768	.726
WEBIS	.731	.228	.347	.876	.273	.416	.695	.337	.454	.833	.405	.545
Baseline	.693	.606	.646	.825	.721	.769	.643	.378	.476	.790	.464	.585

<https://impresso.github.io/CLEF-HIPE-2020/>

Neueste BERT-basierte “Labor”-Ansätze sind erstaunlich gut mit alten und OCR-fehlerbehafteten Texten: Bestes System für Französisch findet 94.9% der NEs (**Recall**); 83.1% der gefundenen NEs sind korrekt (**Precision**).

Named Entities in der impresso Artikelsicht

Luxemburger Wort — Saturday, July 27, 1946

14 p.8

article DE | p.8

FACSIMILE

TRANSCRIPT

FULLSCREEN: OFF

Der „Verräter von Arnheim“ gestorben

Der „Verräter von Arnheim“ gestorben London, 27. Juli. (Reuter.) Der Korrespondent des „Daily Express“ in Amsterdam berichtet seiner Zeitung, daß Chris

LOCATIONS: London, Germany

PEOPLE: Marika Röck, Kurt Singer

TOPICS

DE film - woche - regie - kino - john

DE london - armee - angriff - luftwaffe -

general

DE regierung - berlin - pol - partei -

wien

DE regierung - afp - sowjetunion -

reuter - moskau

DE gericht - gefängnis - polizei -

angeklagt - fall

DE polizei - brand - feuer - nacht -

mann

DE arzt - mittel - krankheit - gesundheit

- wirkung

DE familie - beerdigung - trauer - herr -

hinschied

DE juli - juni - august - mai - samstag

ADD TO COLLECTION ... Arnheim 2944 K

Der „Verräter von Arnheim“ gestorben

London, 27. Juli. (Reuter.) Der Korrespondent des „Daily Express“ in Amsterdam berichtet seiner Zeitung, daß Christian Lindemann, genannt „King-Kong“, der geschickteste Spion, der den Deutschen den Plan über Landung alliierter Luftlandetruppen bei Arnheim

Der „Verräter von Arnheim“ gestorben

© London, 27. Juli. (Reuter.) Der Korrespondent des „Daily Express“ in Amsterdam berichtet seiner Zeitung, daß Christian Lindemann, genannt „King-Kong“, der geschickteste Spion, der den

1 TEXT REUSE

PASSAGE AVAILABLE

Distribution von Linked Entities über Korpus

Luxemburger Wort — Saturday, July 27, 1946 14 p.8

article DE | p.8 FACSIMILE TRANSCRIPT FULLSCREEN: OFF

gestorben

Der „Verräter von A
seiner Zeitung, daß

LOCATIONS London 📍
PEOPLE Marika Röck 👤
TOPICS
DE film · woche · regie ·
DE london · armee · ang
general 📄
DE regierung · berlin · p
wien 📍

ADD TO COLLECTION

Der „Verräte

London, 2
Korresponden
Amsterdam b
Christian, Ti

CURRENT SELECTION

Marika Röck PERSON



☐ apply current search filters (2 filters)

847 results in total (from 1935 to 2006)

ADD AS SEARCH FILTER EXCLUDE FROM CURRENT SEARCH



Marika Röck (Q48502)
HUMAN Cairo, 1913 - Baden bei Wien,
2004

W Hungarian-born actress, singer, dancer
SOURCE: WIKIDATA W/Q48502

hem" gestor

uter.) Der
Daily Express

Named Entities in der Suche

Autovervollständigung

SEARCH ARTICLES SEARCH IMAGES NGRAMS

Etter

...Etter Search in article contents

...Etter Search in article titles

Etter Search in article contents

G. Etter, notaire

M. Etter

G. Etter

Philipp Etter

Petermann Etterlin

Andrew Von Etter

... Etter More Persons ...

Gelinkte und ungelinkte Namen

Dynamische Suchfilter

Philipp Etter

search for ...

FILTER BY PERSON (2,476 OPTIONS)

results are filtered when:

MENTIONING

☒ Philipp Etter (1,470 results)

☐ Max Petitpierre (80 results)

☐ Enrico Celio (46 results)

☐ Hans Streuli (44 results)

☐ Paul Chaudet (41 results)

☐ Ludwig von Moos (39 results)

☐ Giuseppe Motta (38 results)

☐ Thomas Holenstein (38 results)

☐ Giuseppe Lepori (33 results)

☐ Markus Feldmann (25 results)

MORE... (2,466 MORE OPTIONS)

FILTER BY LOCATION (1,603 OPTIONS)

check one or more locations to filter results

☐ Switzerland (833 results)

☐ Bern (424 results)

☐ Auch (304 results)

☐ Lausanne (281 results)

Zusammenfassung

- ▶ Das Impresso-Interface reichert eine Standard-Such-Applikation mit semantischer Information an.
- ▶ Verschiedene Text-Mining-Methoden unterstützen die Nutzenden beim Erfüllen ihres “Informationsbedürfnisses”.
- ▶ Für jede Art von semantischer Anreicherung (NER, Topics) gibt es spezifische Explorationsmöglichkeiten (heute gar nicht angesprochen von mir).

Vielen Dank für die Aufmerksamkeit!
Fragen, Kommentare, Anregungen willkommen

Danke an alle, die zu <https://www.impresso-project.ch> beigetragen haben.

Literaturangaben

BENNANI-SMIREN, KAMIL, C. MUSAT, A. HOSSMANN, M. BAERISWYL und M. JAGGI (2018).
Simple Unsupervised Keyphrase Extraction using Sentence Embeddings, Proceedings of the 22nd Conference on Computational Natural Language Learning, <http://dx.doi.org/10.18653/v1/k18-1022>.

BLEI, DAVID M. (2012).
Probabilistic topic models, Commun. ACM, 55(4):77–84, <http://doi.acm.org/10.1145/2133806.2133826>.

LUI, MARCO und T. BALDWIN (2012).
Langid.Py: An Off-the-shelf Language Identification Tool, In: *Proceedings of the ACL 2012 System Demonstrations*, ACL '12, S. 25–30, Stroudsburg, PA, USA. Association for Computational Linguistics, <http://dl.acm.org/citation.cfm?id=2390470.2390475>.

MIKOLOV, TOMAS, W.-T. YIH und G. ZWEIG (2013).
Linguistic Regularities in Continuous Space Word Representations, In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, S. 746–751, Atlanta, Georgia. Association for Computational Linguistics, <http://www.aclweb.org/anthology/N13-1090>.

SOCHER, RICHARD und C. D. MANNING (2013).
Deep learning for NLP (without Magic),
<http://nlp.stanford.edu/courses/NAACL2013/NAACL2013-Socher-Manning-DeepLearning.pdf>.