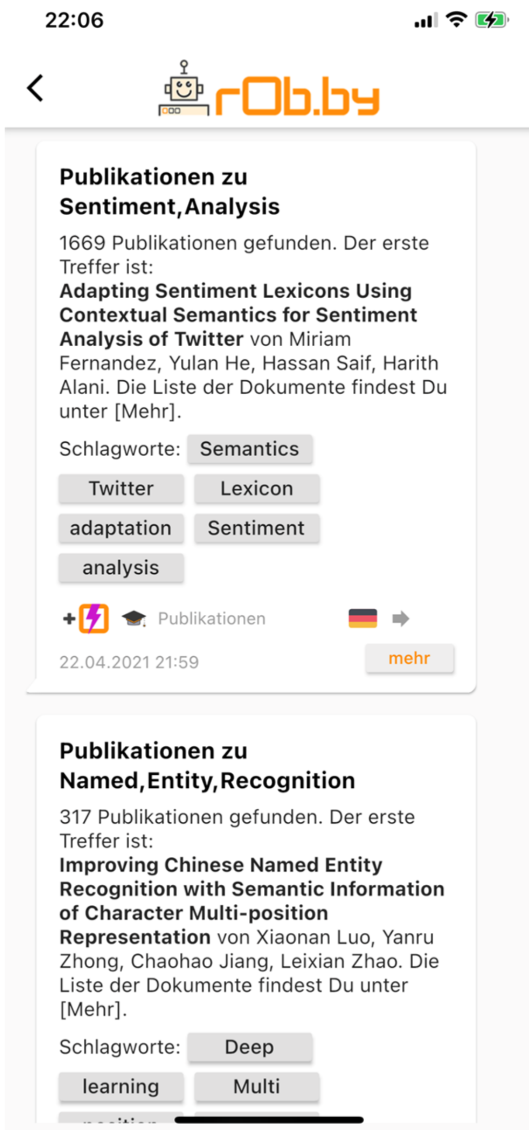
# Finding without Searching: Using Rule-based AI for being automatically updated with relevant Scientific Publications




- Dipl.-Inform. Hermann Bense
  hb@bense.com


  ontology4.us
  rob.by/en/Search/Pubs Examples/
  rob.by/en/Search/Pubs/most_recent/
  ontology4.us/library/Papers/Bens2021b_Pubs/index.html
  schematik.de
  predicator.name
  schreib-maschine.info


  [bense.com] Verlagsgesellschaft für Digitales Publizieren GmbH
  Schwarze-Brüder-Straße 1
  44137 Dortmund

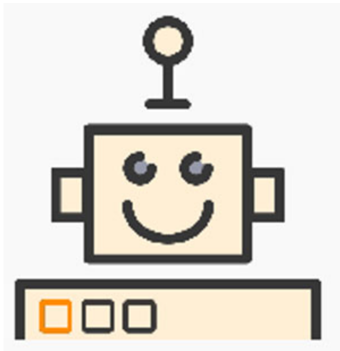



# Finding without Searching:

## Motivation



- Scientists spend a substantial amount of time on
  - Recherching publications
  - Writing papers
  - Creating lists of references
- Publish or perish
  - How can one be sure to know about the most recent publications in the specific research domain?
  - How can one be sure to be the first with a publication
- Solution
  - Apply machine intelligence (AI) for the automatic creation of publication corpora
  - Use smart ChatBot rObby for the automated notification on relevant publications



# Finding without Searching:

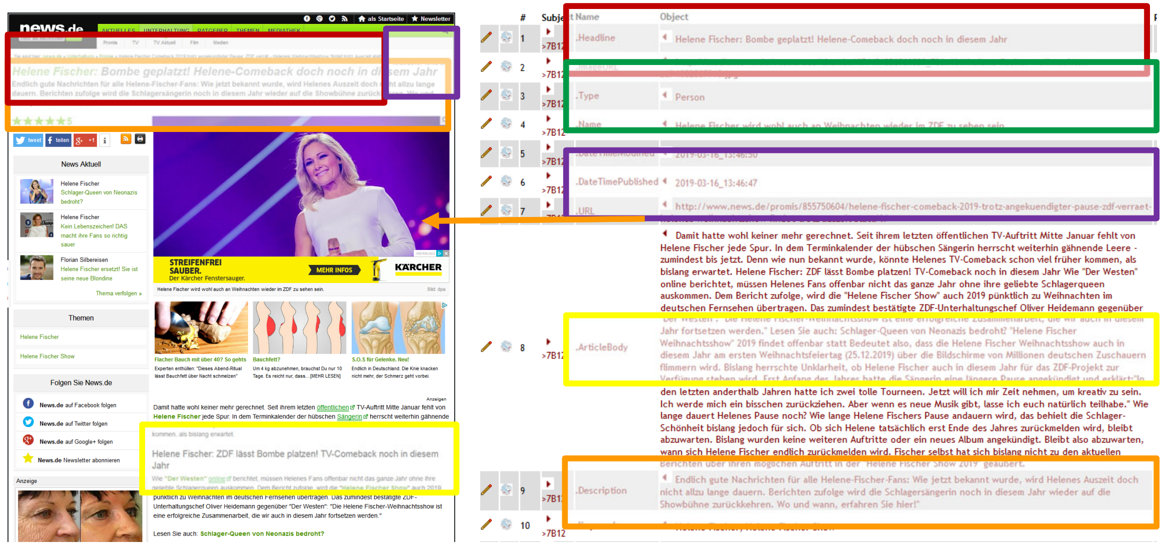

## Agenda

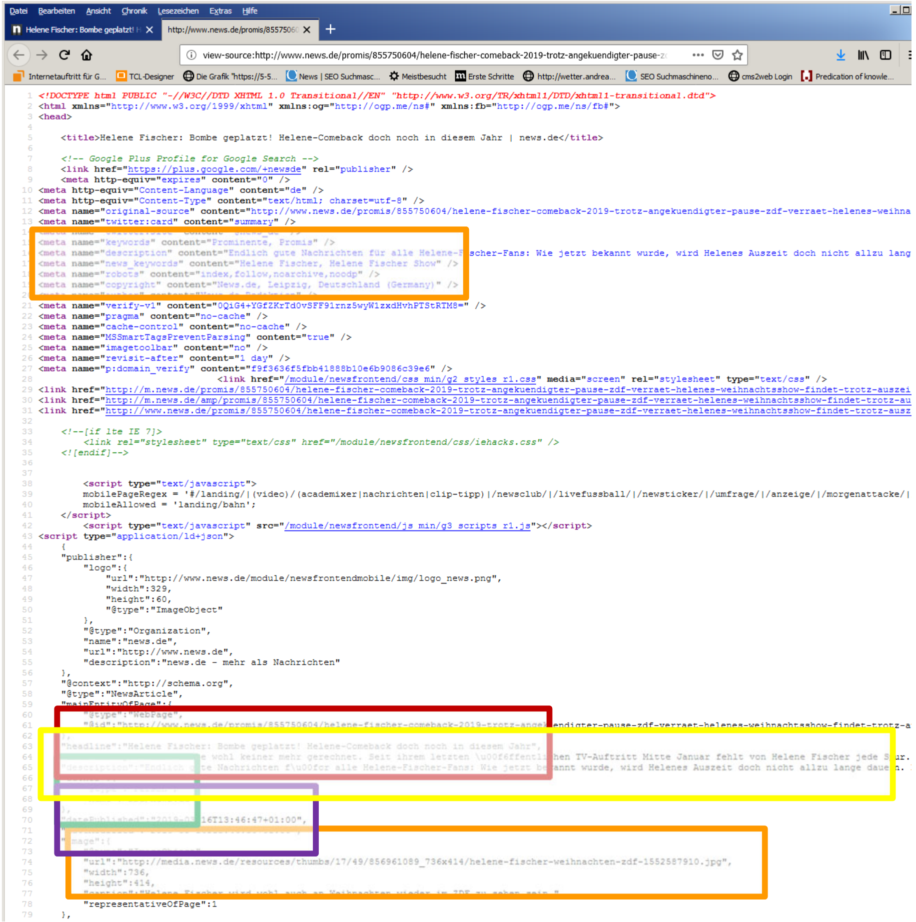
- Corpora
  - Data Model / Import Interface
  - Document Structure
  - Crawling and Indexing
- Searching and Alerting
  - Browsing
- rOb.by-App
  - Rule-based Notification on Publications
- Relevance & Performance Issue
  - Recall & Precision
  - Caching
- Comparison to other Search Engines
  - TIB, Google Scholar, Google
- Summary
  - Benefits

# Publication Imports

## Data Model / Import Interface







- Keywords stored for Authors, Titles, Dates
- Using mySQL-base Triple-store O4Store
- Plus Key-Value-Store (KVS)

# Crawling of Scientific Publications

- **Sources**
  - Springer Professional
  - TIB
  - PubMed
- **Size of Corpus**
  - Number of Documents: ~ 2.9 Mio.
  - Number of Triples: ~232 Mio.
  - Number of KVS Tripels: ~ 90 Mio.
- **Crawling speed and volume**
  - Daily Rate: approx. 2.500 to 3.400 new publications
- **Structure**
  - DOI, URL
  - Authors, Title, Summary
  - Disciplines
  - Keywords
  - References

List (0 - 40 / 1 found /231.417.305 total number of entries) Importer | OWL/raw-data-Importer

Search SELECT * FROM relation WHERE (`Relation_Subject` = '>PUB_9K1NIK') ORDER BY `Id` DESC

| # | Subject | Name | Object | Ontologies |
|---|---|---|---|---|
| 1 | >PUB_9K1NIK | <>PUB_prev | >PUB_2VM9I4 | SPRPRF |
| 2 | >PUB_9K1NIK | .Description | Our objective in this work is long range understanding of the narrative structure of movies. Instead of considering the entire movie, we propose to learn from the 'key scenes' of the movie, providing a condensed look at the full storyline. To this end, we make the following three contributions: (i) We create the Condensed Movies Dataset (CMD) consisting of the key scenes from over 3 K movies: each key scene is accompanied by a high level semantic description of the scene, character face-tracks, and metadata about the movie. The dataset is scalable, obtained automatically from YouTube, and is freely available for anybody to download and use. It is also an order of magnitude larger than existing movie datasets in the number of movies; (ii) We provide a deep network baseline for text-to-video retrieval on our dataset, combining character, speech and visual cues into a single video embedding; and finally (iii) We demonstrate how the addition of context from other video clips improves retrieval performance. | SPRPRF |
| 3 | >PUB_9K1NIK | <>isi | ^Publication | SPRPRF |
| 4 | >PUB_9K1NIK | <>Subdiscipline | >DSP_Artificial_Intelligence | SPRPRF |
| 5 | >PUB_9K1NIK | .DOC | {BODY{{TITLE{Condensed Movies: Story Based Retrieval with Contextual Embeddings}TITLE} {DOI{978-3-030-69541-5_28}DOI} {PUBLISHEDDATE{2021-02-26}PUBLISHEDDATE} {AUTHOR{Max Bain{AUTHOR}Arsha Nagrani{AUTHOR}Andrew Brown{AUTHOR}Andrew Zisserman}AUTHOR} {KEYWORD{}KEYWORD} {DESCRIPTION{Our objective in this work is long range understanding of the narrative structure of movies. Instead of considering the entire movie, we propose to learn from the 'key scenes' of the movie, providing a condensed look at the full storyline. To this end, we make the following three contributions: (i) We create the Condensed Movies Dataset (CMD) consisting of the key scenes from over 3 K movies: each key scene is accompanied by a high level semantic description of the scene, character face-tracks, and metadata about the movie. The dataset is scalable, obtained automatically from YouTube, and is freely available for anybody to download and use. It is also an order of magnitude larger than existing movie datasets in the number of movies; (ii) We provide a deep network baseline for text-to-video retrieval on our dataset, combining character, speech and visual cues into a single video embedding; and finally (iii) We demonstrate how the addition of context from other video clips improves retrieval performance.}DESCRIPTION} }BODY} | SPRPRF |
| 6 | >PUB_9K1NIK | <>is_Publication_of | >ATHR_TTEXXR | SPRPRF |
| 7 | >PUB_9K1NIK | <>Discipline | >DSP_Computer_Science | SPRPRF |
| 8 | >PUB_9K1NIK | .KVS | F | SPRPRF |
| 9 | >PUB_9K1NIK | <>is_Publication_of | >COM_NFDQ5H | SPRPRF |
| 10 | >PUB_9K1NIK | .CoverImage | https://media.springernature.com/w306/springer-static/cover/book/978-3-030-69541-5.jpg | SPRPRF |
| 11 | >PUB_9K1NIK | .DatePublished | 2021-02-26 | SPRPRF |
| 12 | >PUB_9K1NIK | <>is_Publication_of | >ATHR_UIZ0U2 | SPRPRF |
| 13 | >PUB_9K1NIK | .URL | https://link.springer.com/chapter/10.1007/978-3-030-69541-5_28 | SPRPRF |
| 14 | >PUB_9K1NIK | .DOI | 978-3-030-69541-5_28 | SPRPRF |
| 15 | >PUB_9K1NIK | .DateTimeImported | 2021-04-21_12:57:49.354 | SPRPRF |
| 16 | >PUB_9K1NIK | .Title | Condensed Movies: Story Based Retrieval with Contextual Embeddings | SPRPRF |
| 17 | >PUB_9K1NIK | <>is_Publication_of | >ATHR_KOY3ZP | SPRPRF |
| 18 | >PUB_9K1NIK | <>is_Publication_of | >ATHR_JDSXEX | SPRPRF |
| 19 | >PUB_9K1NIK | .References | 56 | SPRPRF |
| 20 | >PUB_9K1NIK | <>Reference | >REF_LOCLG3 | SPRPRF |
| 21 | >PUB_9K1NIK | <>Reference | >REF_SFZO1Z | SPRPRF |

- Springer Professional: https://www.springerprofessional.de/
- TIB Leibniz-Informationszentrum Technik und Naturwissenschaften: https://www.tib.eu/de/
- NIH National Library of Medicine (PubMed): https://pubmed.ncbi.nlm.nih.gov/

# Indexing of Publications

- **Multiple Language Support**
  - **Translations of Titles from any Language to English and German**
- **Lemmatization**
  - **German and English titles are lexically analysed by Stanza and TreeTagger**
- **Key Value Store (KVS)**
  - **All Author names, keywords and lemmas and the publishing date are stored in the KVS**
  - **Each KVS entry has a count for the number of documents referenced by the entry**

| | |
|---|---|
| >PUB_V4T350 NULL<br>978-3-319-62971-1_15<br>**2017-11-03**<br>isTrans<br>KVS_Pubs_exists<br>EN_Title_exists<br>DE_Title_exists<br>TimeOut \| new<br>start:<br>\| lngEN:EN lngDE:EN lngFR:<br>P7: KB:CRAWLER\|KVS:F<br>PrevId:NULL\|NextId:NULL<br>CD:2021-04-14 | 0:Title Culture of Engagement: Preparing Civic-Minded Public Service Professionals of the Future<br>Description: The emerging arena of community-based research (CBR) and service-learning incorporates a collaborative approach of civic engagement for students to analyze and develop solutions to complex problems and bring about social change and strengthen communities. Focusing specifically on the student populat<br>**10** KWinserted\|Keywords:**Kapucu, Community, based, Research,, CBR, facultyFaculty, Service, learning, Course, Public, Administration, Programs**<br>ANinserted\|AuthNames:**Naim Kapucu, Fatih Demiroz, Brittany Haupt, Mirtha Bailey**<br>**3** TKWinserted\|TitleWordsKVS:**Culture, of, Engagement, Preparing, Civic, Minded, Public, Service, Professionals, of, the, Future**<br>**2** SDinserted\| Subdiscipline:**Higher,Education**<br>**6** TW_DE_inserted\|LemmasDEulO:kultur, engagement, vorbereitung, zukünftig, fachkraft, dienst, bürgersinn<br>**1** TW_EN_inserted\|LemmasENulO:culture, engagement, preparing, civic, minded, service, professional, future<br>URL: Detail:>PUB_V4T350 \| NULL<br>CorAuthor:<br>deepl Results:<br>TREN:EN\|Culture of Engagement: Preparing Civic-Minded Public Service Professionals of the Future<br>TRDE:EN\|Kultur des Engagements: Vorbereitung der zukünftigen Fachkräfte des öffentlichen Dienstes mit Bürgersinn<br>DateTimeImported: |

```
PHP-Time:43.65404009819 PHP-Total:44.574103116989
Query-Time: 2.5348002910614 Queries:68 Total:99 Time/Query:0.025604043344055
```

| | |
|---|---|
| >PUB_2VM9I4 TIB<br>NULL<br>**2018-01-01**<br>isTrans<br>KVS_Pubs_exists<br>EN_Title_exists<br>DE_Title_exists<br>TimeOut \| new<br>start:<br>Language:de \| lngEN:DE lngDE:DE lngFR:<br>P7: KB:CRAWLER\|KVS:F<br>PrevId:>PUB_DFLDDR\|NextId:NULL<br>CD:2021-04-14 | 0:Title „Vergessen, verdrängt, verschwunden“: aufgegebene Kulturen, Beziehungen und Orientierungen in der Balkanromania<br>Description: NULL<br>**5** KWinserted\|Keywords:**Literatur, Kulturwandel, sprachliche, Minderheit, Balkanromanisch**<br>ANinserted\|AuthNames:<br>**3** TKWinserted\|TitleWordsKVS:**Vergessen, verdrängt, verschwunden, aufgegebene, Kulturen, Beziehungen, und, Orientierungen, in, der, Balkanromania**<br>**2** SDinserted\| Subdiscipline:**Historische,Linguistik**<br>**5** TW_DE_inserted\|LemmasDEulO:verdrängen, aufgegeben, kultur, beziehung, orientierung, balkanromania<br>**3** TW_EN_inserted\|LemmasENulO:cultures, relations, orientations, balkanromania<br>URL: Detail:>PUB_2VM9I4 \| https://www.tib.eu/de/suchen/id/TIBKAT:1009949446/Vergessen-verdr%C3%A4ngt-verschwunden-aufgegebene-Kulturen?cHash=84e1612bc960b1afb8325713067c170d<br>CorAuthor:<br>deepl Results:<br>TREN:DE\|"Forgotten, repressed, disappeared": abandoned cultures, relations and orientations in Balkanromania.<br>TRDE:DE\|„Vergessen, verdrängt, verschwunden“: aufgegebene Kulturen, Beziehungen und Orientierungen in der Balkanromania<br>DateTimeImported: |

```
PHP-Time:26.297097921371 PHP-Total:26.39576292038
Query-Time: 0.7346830368042 Queries:70 Total:101 Time/Query:0.0072740894733089
```

- *Stanza (StanfordNLP): https://stanfordnlp.github.io/stanza/pos.html*
- TreeTagger: *https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger/*

# Keyword Frequencies

- The following list represents the relation between the order of magnitude of keywords and the number of documents to which they are associated. The selective in per Mille designates the portion of documents in relation to the corpus of 2.9 Mio. documents
  - ca. 57.000 Keywords with >= 50 findings
  - ca. 34.900 Keywords with >= 100 findings, selectivity: 0,03 %o

- High frequency keywords (HFK)
  - ca. 5400 Keywords with >= 1000 findings, selectivity: 0,3 %o
  - ca. 920  Keywords with >= 5000 findings, selectivity: 1,5 %o

- Very high frequency keywords (VHFK)
  - ca. 340 Keywords with >= 10.000 findings, selectivity: 3 %o
  - ca. 100 Keywords with >= 20.000 findings, selectivity: 6 %o
  - ca. 45 Keywords with >= 30.000 findings, selectivity: 9 %o
  - ca. 25 Keywords with >= 40.000 findings, selectivity: 14 %o, documents: 25 * 40k = 1 Mio.
  - ca. 20 Keywords with >= 50.000 findings, selectivity: 17 %o, documents: 20 * 50k = 1 Mio.
  - ca. 10 Keywords with >= 70.000 findings, selectivity: 25 %o, documents: 10 * 70k = 700k

- With the 10 + 20 + 25 = 55 top keywords 2.7 Mio documents are indexed, which is almost the complete corps
  - The list of keywords frequencies shows that the top keyword *Engineering* is assigned to ca. 130 k documents.
  - It is followed by the keywords ***Systems***, ***Intelligence***, ***Analysis***, ***Management***, ***based***, ***System***, ***Theory***, ***computational*** **and** ***Information*** each of them indexing more than 60k documents.
  - Taking *Systems* and *System* together would even account for 175k documents.
  - The 10 top keywords select ca. 700k documents, the following 11 to 20 about 1 Mio. and again the following 21 to 30 also about 1 Mio.
  - Keywords indexing more than 10k documents are regarded as very high-frequency keywords (VHFK). About 340 keywords fulfill this criteria.
  - The keywords indexing less than 100 documents are coined very low-frequency keywords (VLFK) those with less than 1k documents low-frequency keywords (LFK). About 57k keywords index between 50 to 100 documents

# Plurals, Homonyms, Author Names



## Recommendations for Search Optimization

- **Plurals:**
  - Often the singular and plural forms of nouns are indexed. To find all appearences in publications in search queries both forms should be used in rules/queries using the pipe symbol **|** for the OR-function e.g. `System|Systems` or `Machine|Machines`

- **Homonyms:**
  - Very high frequency homonyms (VHFKs) like *brand* (Englisch and German noun), *can* (English verb and noun), *jet* (English conjunction and noun), *lead* (English adjective and noun), *not/Not* (English negation and German noun for *need*), *may* (verb and name of month), *second* (numeral and noun) , *set* (verb and noun), *song/Song* (noun and named entity), *state/s* (noun for *status* and noun for *country*) and *use* (verb and noun) require special treatments.
  - In best case the meaning can be derived from the context where the words are in. But currently is it not simply possible to make this distinction for the entries in the KVS.
  - A similar problem shows up for named entities. Examples: *Schade* (last name of author and german adjective for *pity*) and *Siegel* (last name of author and german noun for *seal/signet*).

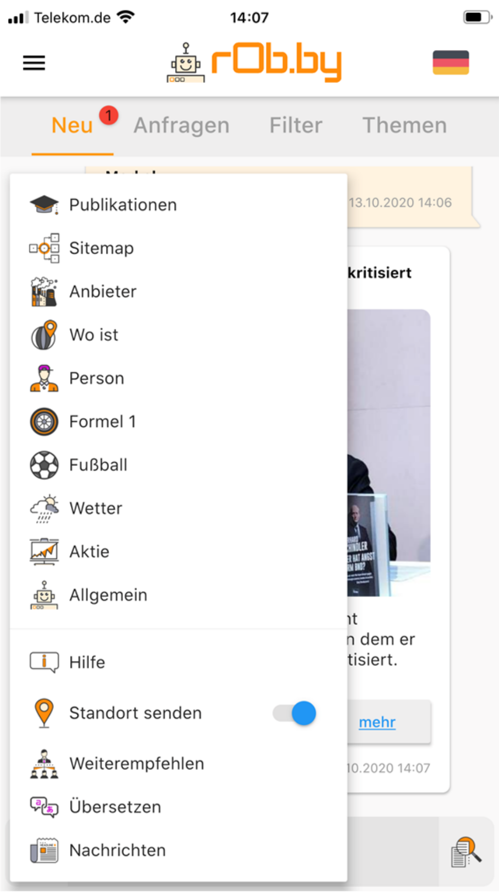
- **Author Names:**
  - Normally author names are very selective especially in combinations.
  - But a lot of Author names belong to the set of very high frequency keywords (VHFK) like *John, Paul* and *Smith*, Asian author names like *Cheng, Gao, Guo, Han, Huang, Jiang, Kim, Lee, Lin, Liu, Lung, Yang, Zhao, Zhang*, *Zheng, Zhou* and *Zhu* and Indian authors names like *Kumar* and *Singh* belong to the VHFKs. Also often cited authors like *Moore* and *Markov* fall into this category.

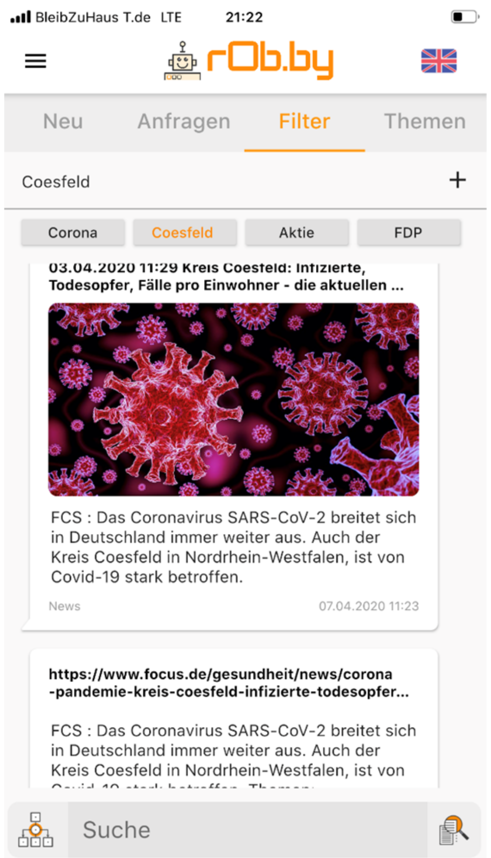
# rOb.by App for iOS and Android

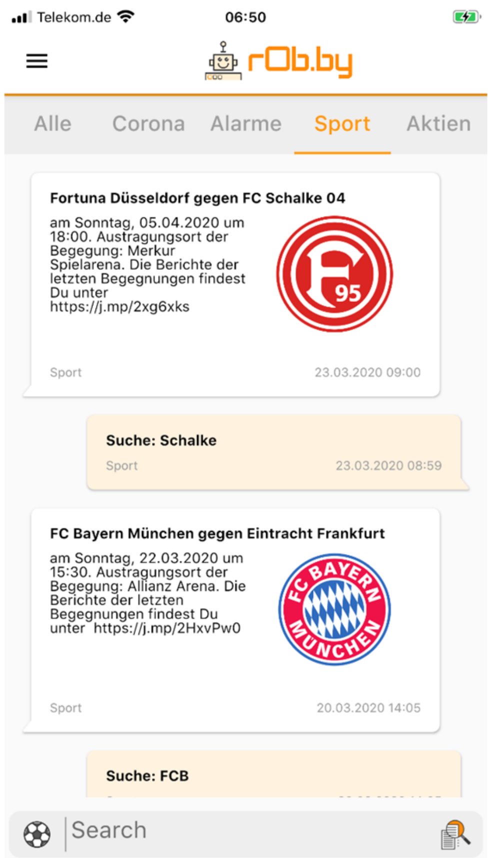
[ontology4.us]

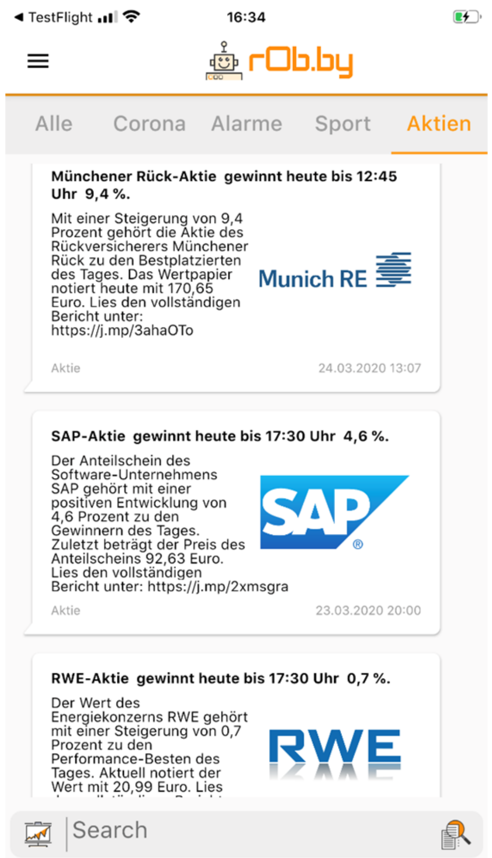
## Examples for User Queries and rule-based Notifications

- **rOb.by Functions:**
  - News, Publication, Event, Weather, Stock Search & Notifications
  - Multi language translations and chat for > 25 languages supported by deepl.com



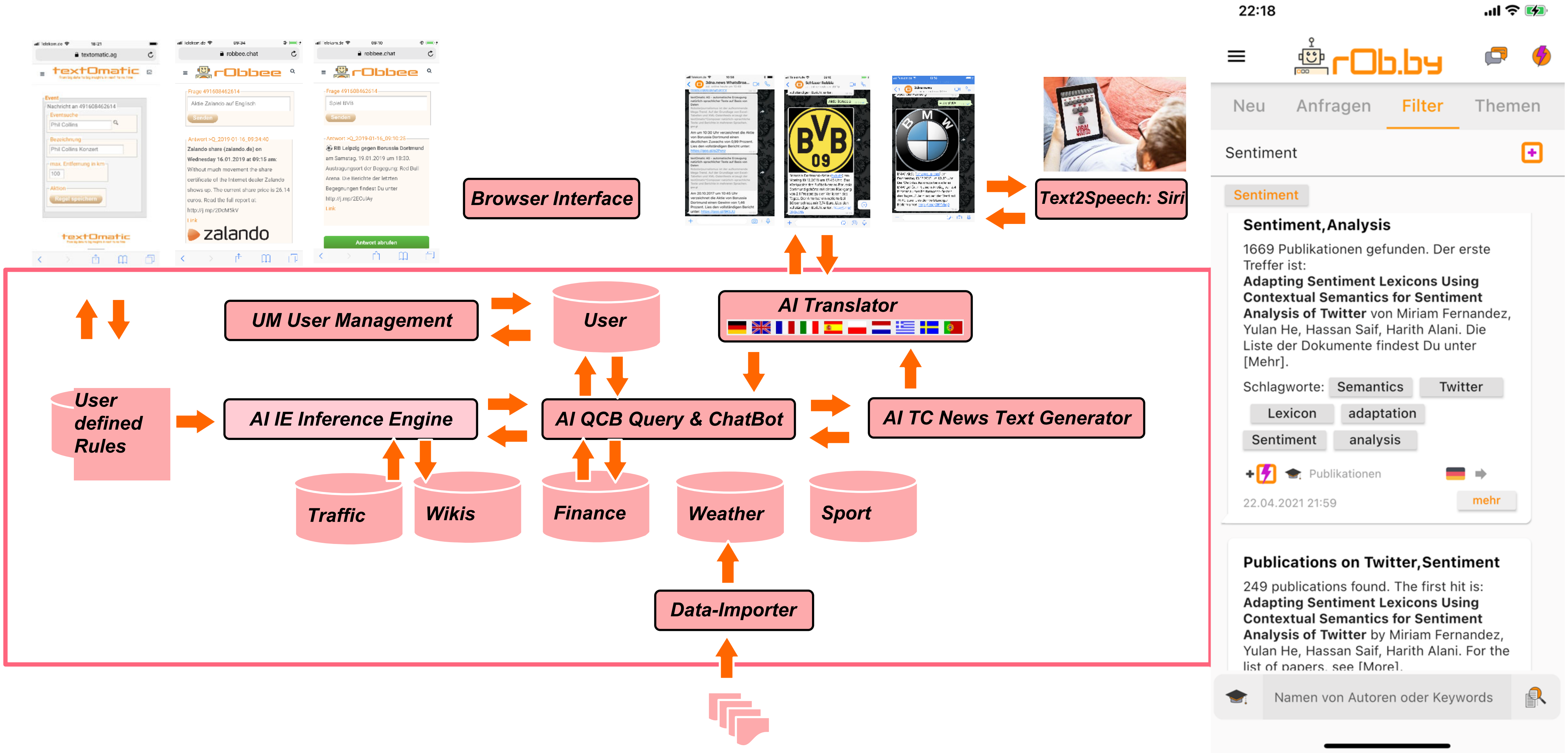




- **URL:** https://rob.by/en/App
- **DeepL:** https://en.wikipedia.org/wiki/DeepL_Translator

# NAS (News Alert System)
## Architektur & Technologie

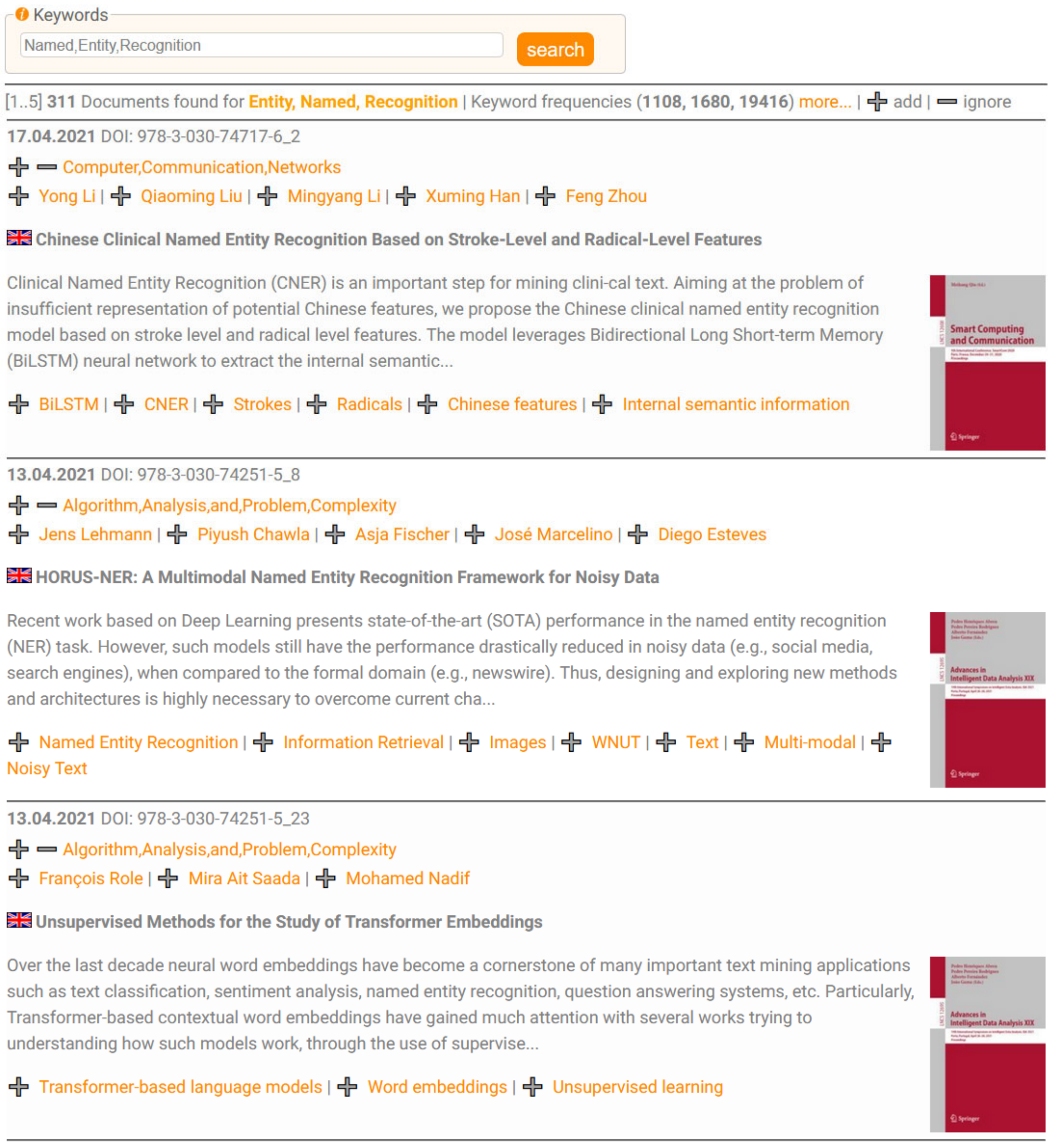





- *Big Data Sources for Events, Sport, Finance, Weather, Traffic , Scientific Publications*

# Search Example
## Named Entity Recognition

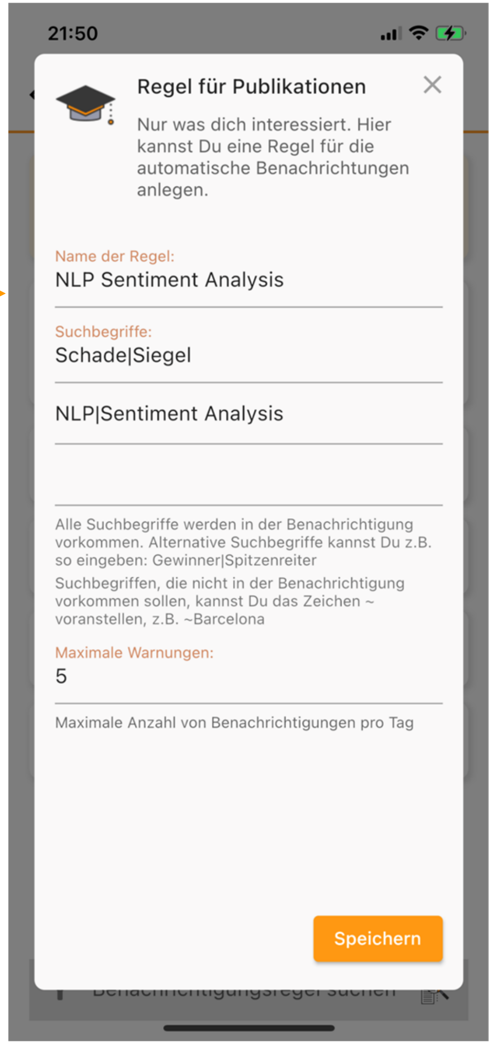
- Keywords Frequencies are displayed
  - Entity: 1108
  - Named: 1680
  - Recognition: 19416
- Search Algorithm
  - Keywords are sorted by descending number of publications
  - The set of all publication Ids is retrieved
  - For more then one keyword, the intersection of the set of publication Ids is computed
- Easy to use User Interface (UI)
  - Click on Keyword(s) to start a new search
  - Click on ＋ to add a keyword to search to make query more selective / precise
- Results (311)
  - are ordered by descending actuality




- URL: https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

# Query Types

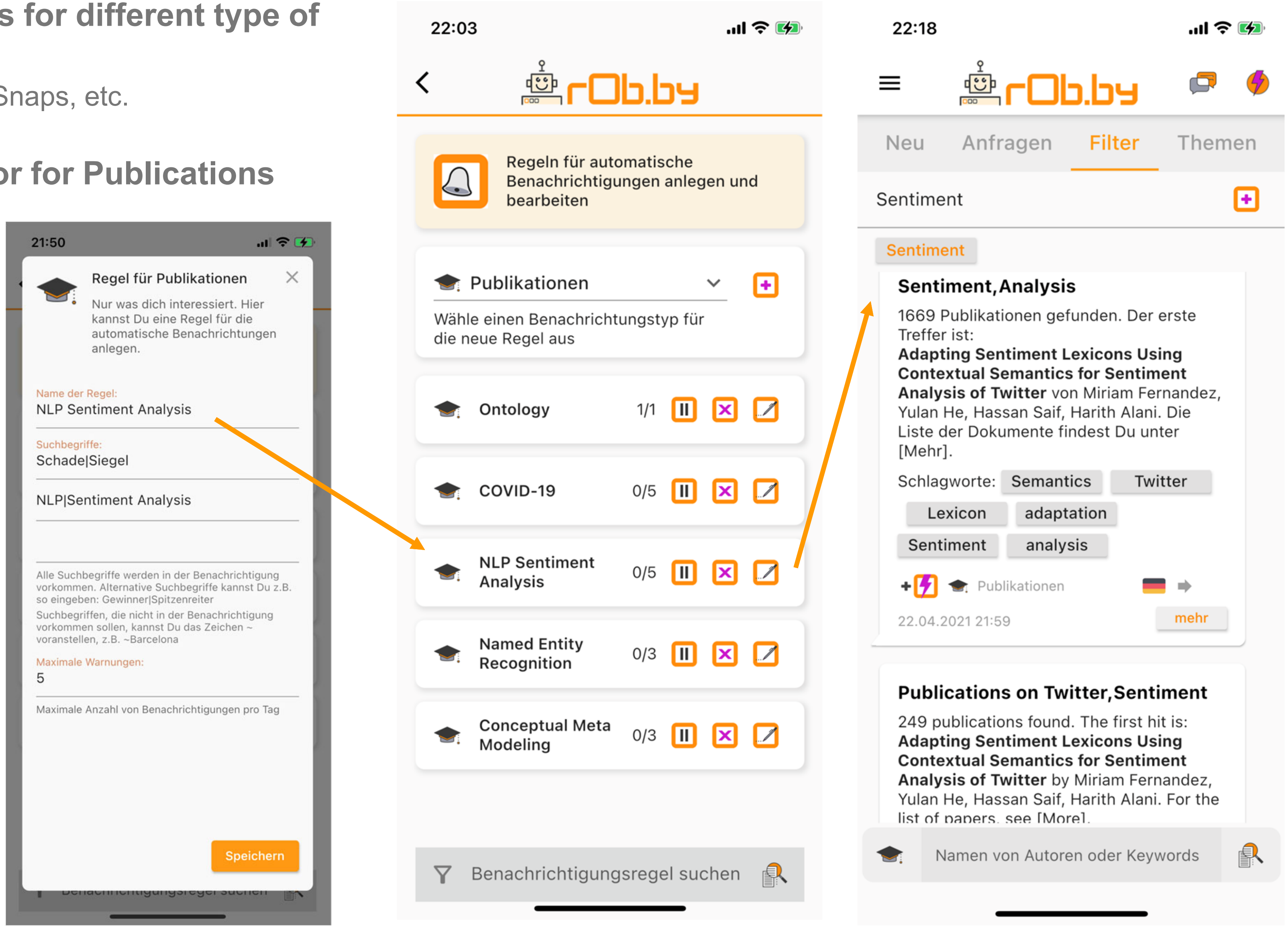
- **Logical AND**
  - Entity,Named,Recognition
  - Bense,Reibold,Hoppe,Humm
- **Logical OR**
  - Learning|Intelligence
  - Schade|Siegel
- **Logical OR and AND**
  - Machine|Deep,Learning
  - Schade|Siegel,NLP|sentiment analysis
- **Logical NOT**
  - ~Blockchain
  - From the result set of a query those Ids a removed where the keyword ~Blockchain is assigned
- **Time**
  - [2020-07-01
  - Retrieves only documents published from 01.07.2020 on




- **Melanie Siegel:** https://rob.by/en/Search/Pubs/Melanie%C2%A0Siegel.html
- **Ulrich Schade:** https://rob.by/en/Search/Pubs/Ulrich%20Schade.html
- **Sentiment Analysis:** https://rob.by/en/Search/Pubs/Sentiment,analysis.html
- **Schade|Siegel,NLP|sentiment analysis** https://rob.by/en/Search/Pubs/Schade|Siegel,NLP|Sentiment%20analysis.html

# Robby-App Rule Editor

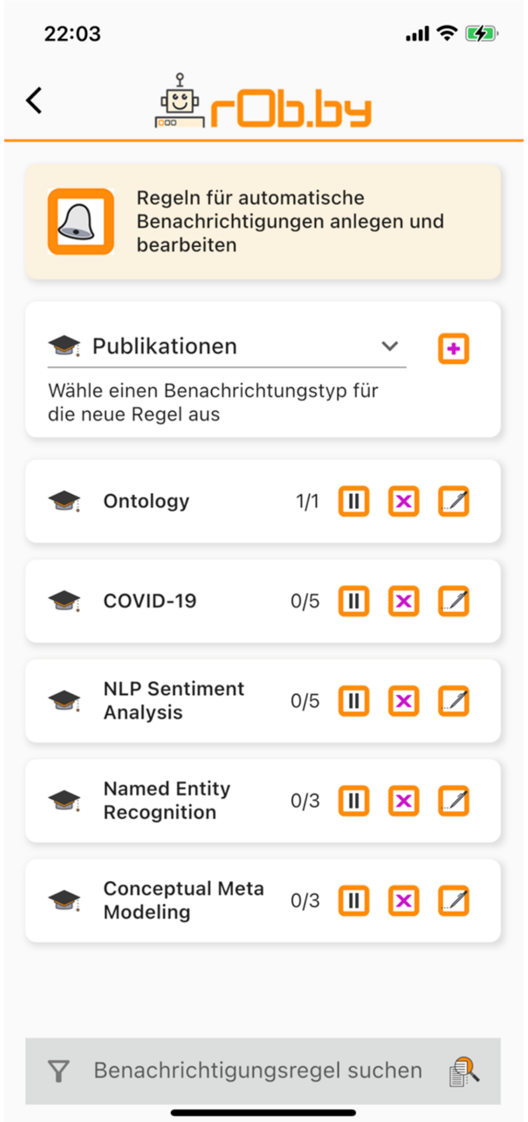
- **User friendly Editors for different type of Notifications:**
  - Weather, Events, Snaps, etc.
- **Example: Rule Editor for Publications**

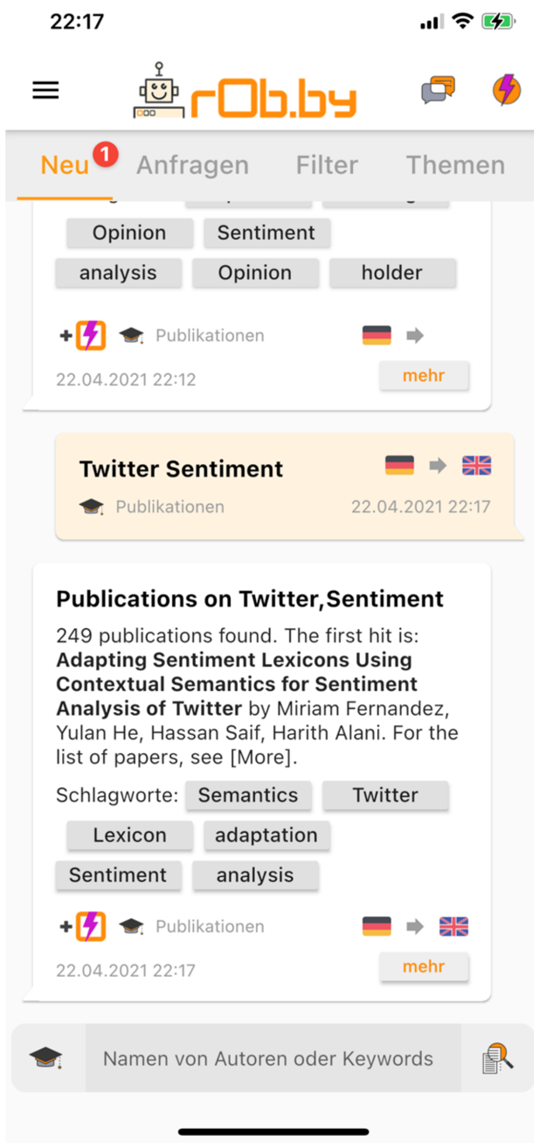



- URL: https://rob.by/en/Search/Pubs/Sentiment,analysis.html

# Processing Robby-App Rule Alerts
## Named Entity Recognition

- **The backend rule processor of the rOb.by App permanently checks all the rules of a user.**
- **If a rule fires, then a notification is generated and send to the user as**
  - Push alert in the rOb.by-App or
  - As E-Mail
- **In the notification the user can click on [more/mehr] to see all results**
- **A single result has an URL to visit the page of the Publisher (TIB, Springer Professional etc.)**



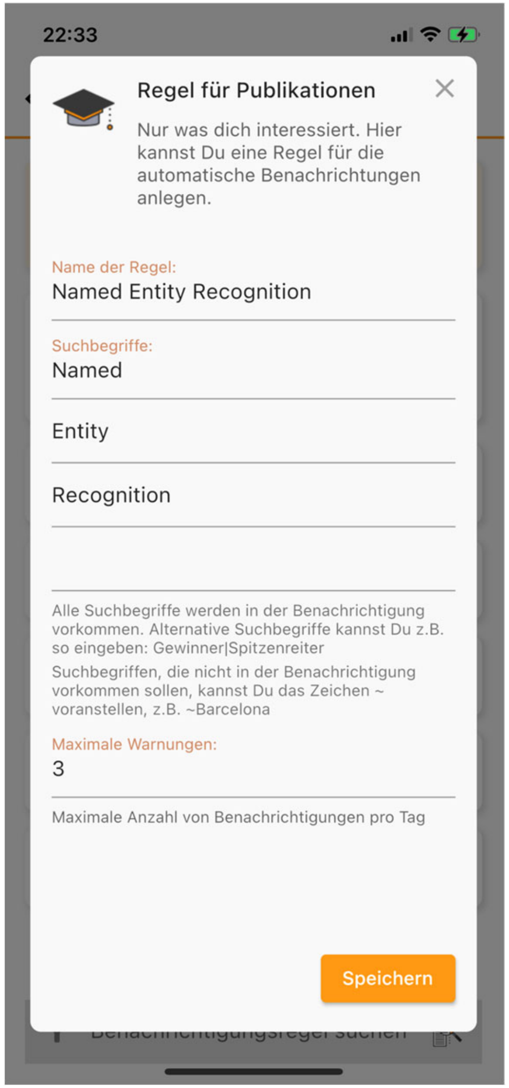



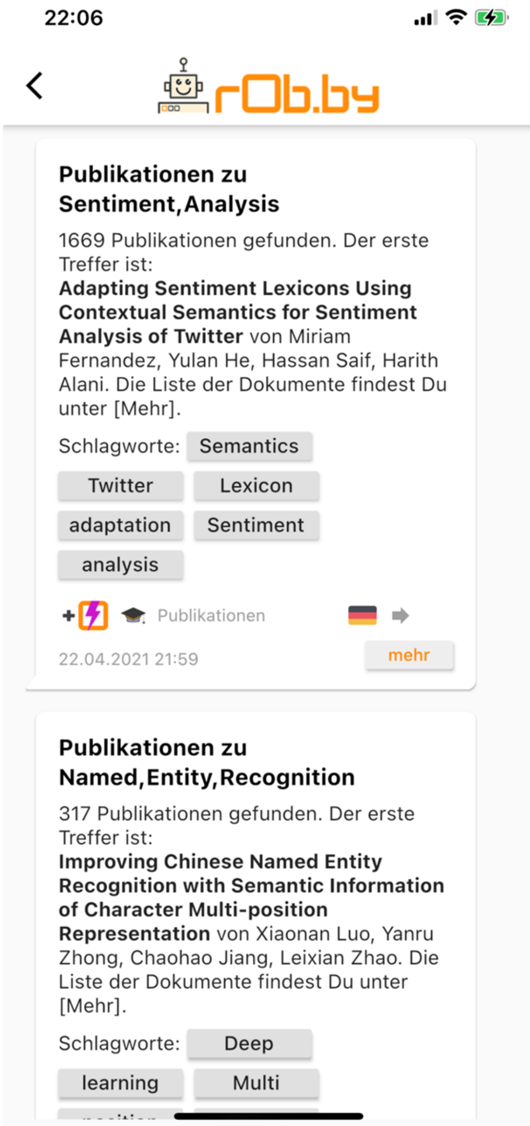

- **URL:** https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

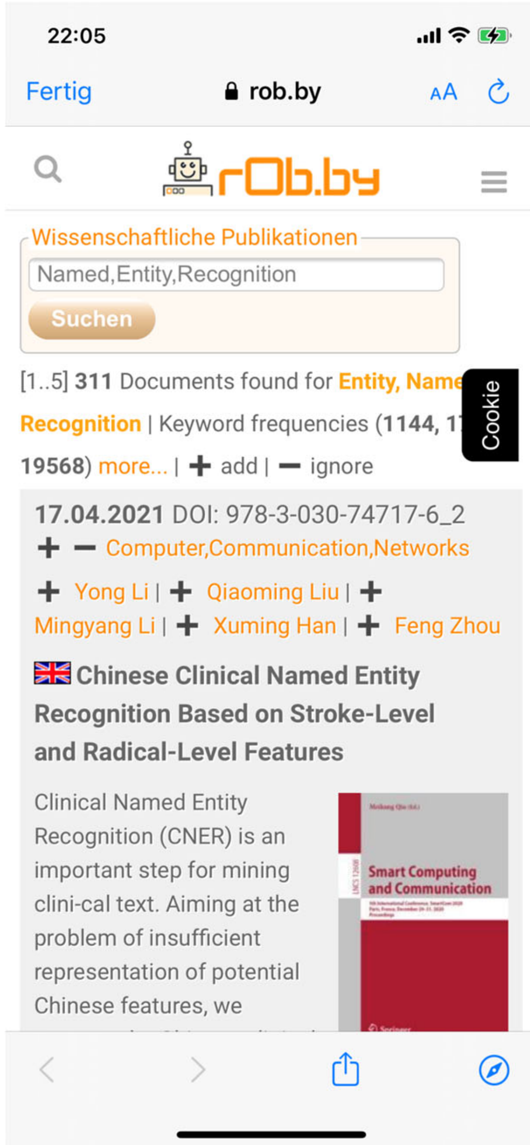
# Robby-App Rule Alerts for Publications

## Named Entity Recognition






- URL: https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

# Search Speed Optimization by Caching of Queries Results

[ontology4.us]

- Average answer time on cached Queries is << 1 Second
- Cache Time is 30 days

###~update_QC QID:> used:1

QC_Id:**>QC_7I45R46M**

cached ✅ | ❌ createQC| ❌ updateQC Count:33

Keywords: Writer,identification

DateTime_created: **2021-04-18_11:58:29**

DateTime_updated:2021-04-19_17:04:02 + CachePeriod: **0000-01-00_00:00:00** = DateTime_4_Update: **2021-05-19_17:04:02**

❌ abgelaufen | 2021-04-23_09:53:56 > 2021-05-19_17:04:02

QC_Id:>QC_7I45R46M | ~create_QC###

- If the user uses keyword combinations, for which the cache time has expired, the cache entry is automatically updated
- This can last a few seconds to 25-30 seconds depending on the number of documents found for each search key

###~update_QC QID:> used:18

QC_Id:**>QC_7I45RVK5**

cached ✅ | ❌ createQC| ✅ updateQC Count:796

Keywords: Cyber,Security

DateTime_created: **2021-04-18_13:58:19**

DateTime_updated:2021-04-19_19:57:20 + CachePeriod: **0000-00-01_00:00:00** = DateTime_4_Update: **2021-04-20_19:57:20**

✅ abgelaufen | 2021-04-23_09:56:44 > 2021-04-20_19:57:20

QC_Id:>QC_7I45RVK5 | ~create_QC###

- **URL:** https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

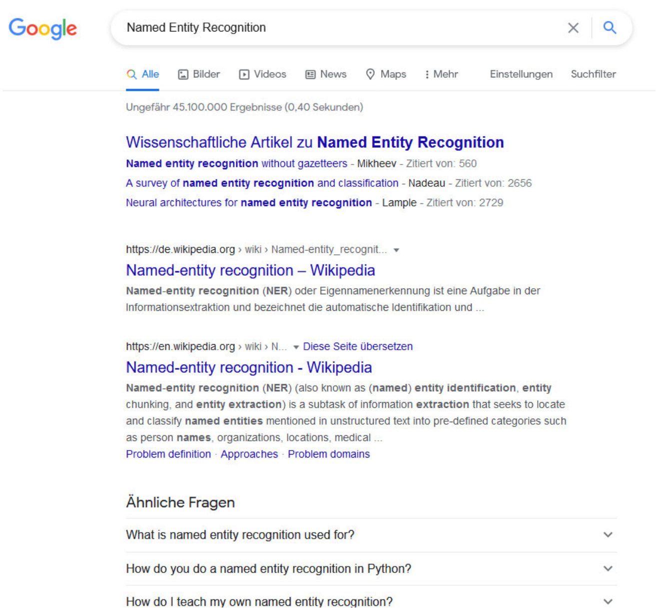
# Comparision with other Search Engines / Portals

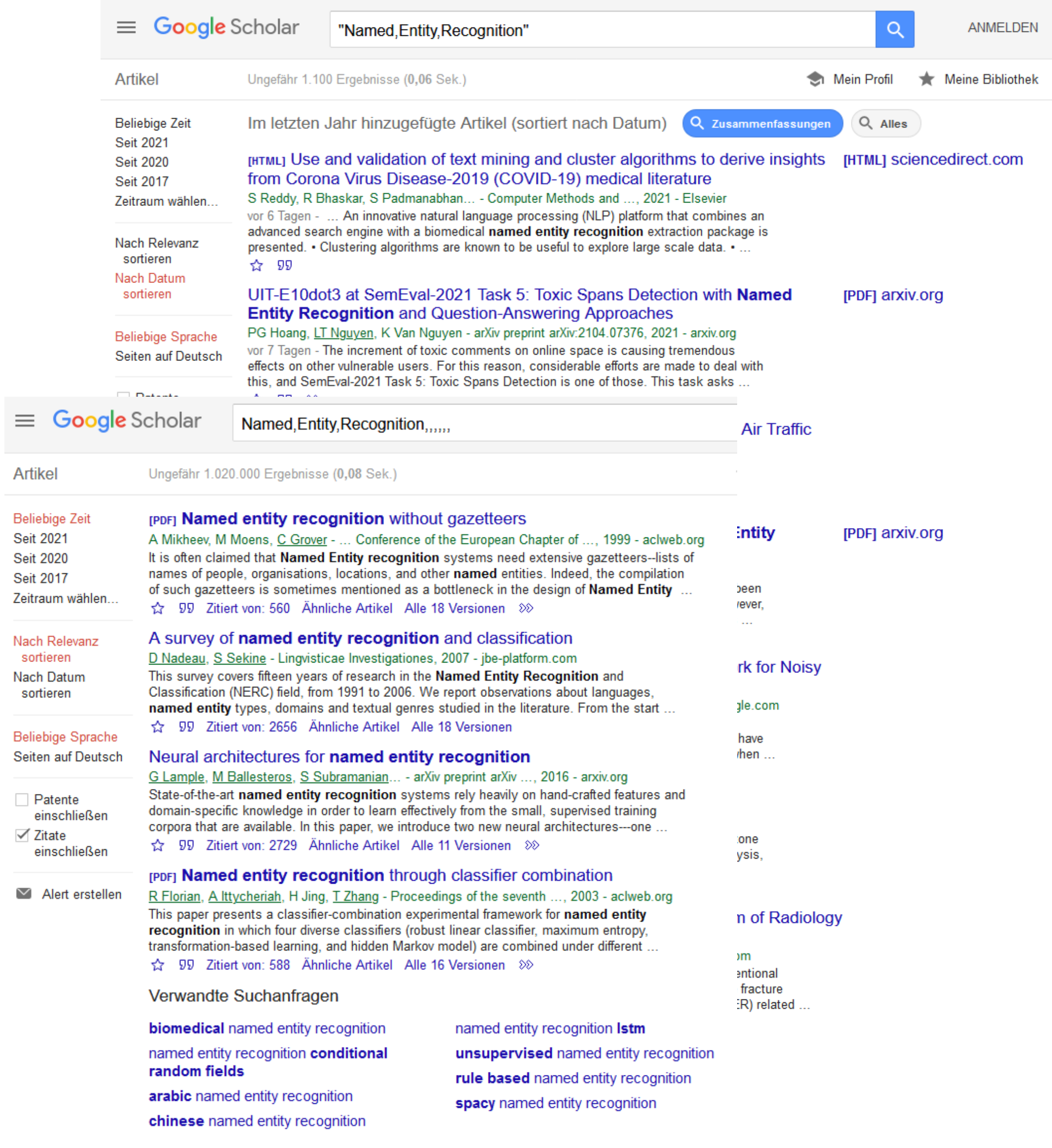


## Google Scholar

- Very fast, very large corpus
- Default Setting for search: order by relevance
- No indication for
  - Publishing Date
  - Keyword Frequencies

- **Google Search**
  - Limited possibility to only find publications



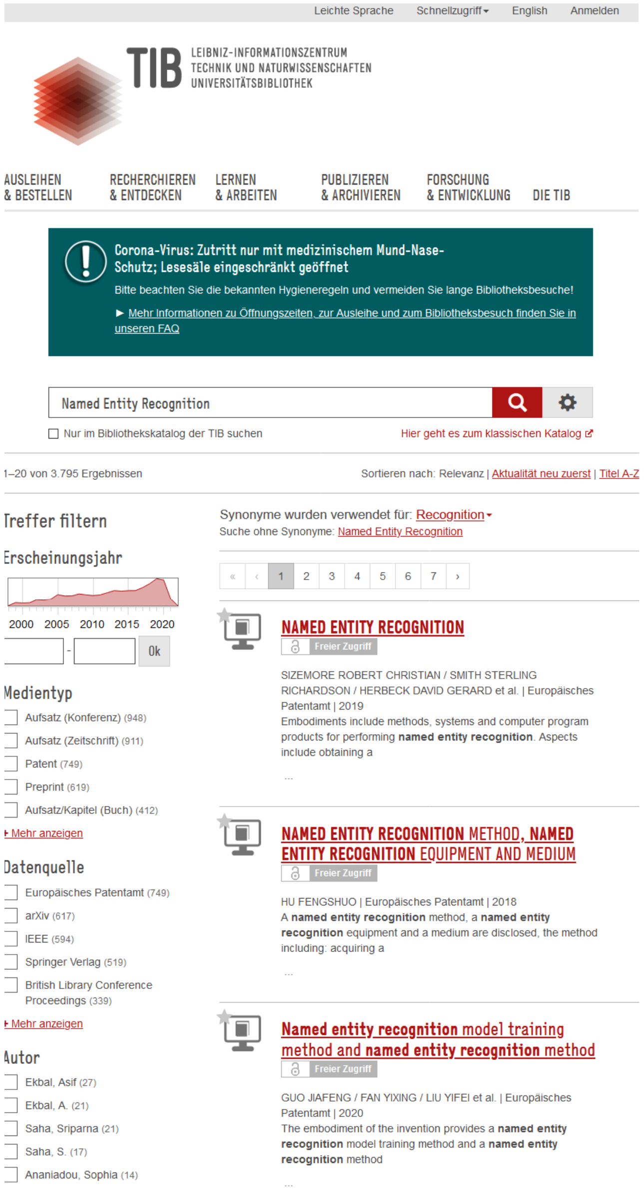

- **URL:** https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

# Comparision with other Search Engines / Portals

[ontology4.us]

## TIB

- Prefers Patents
- Do not have Publication Dates,
  - only Publication Years



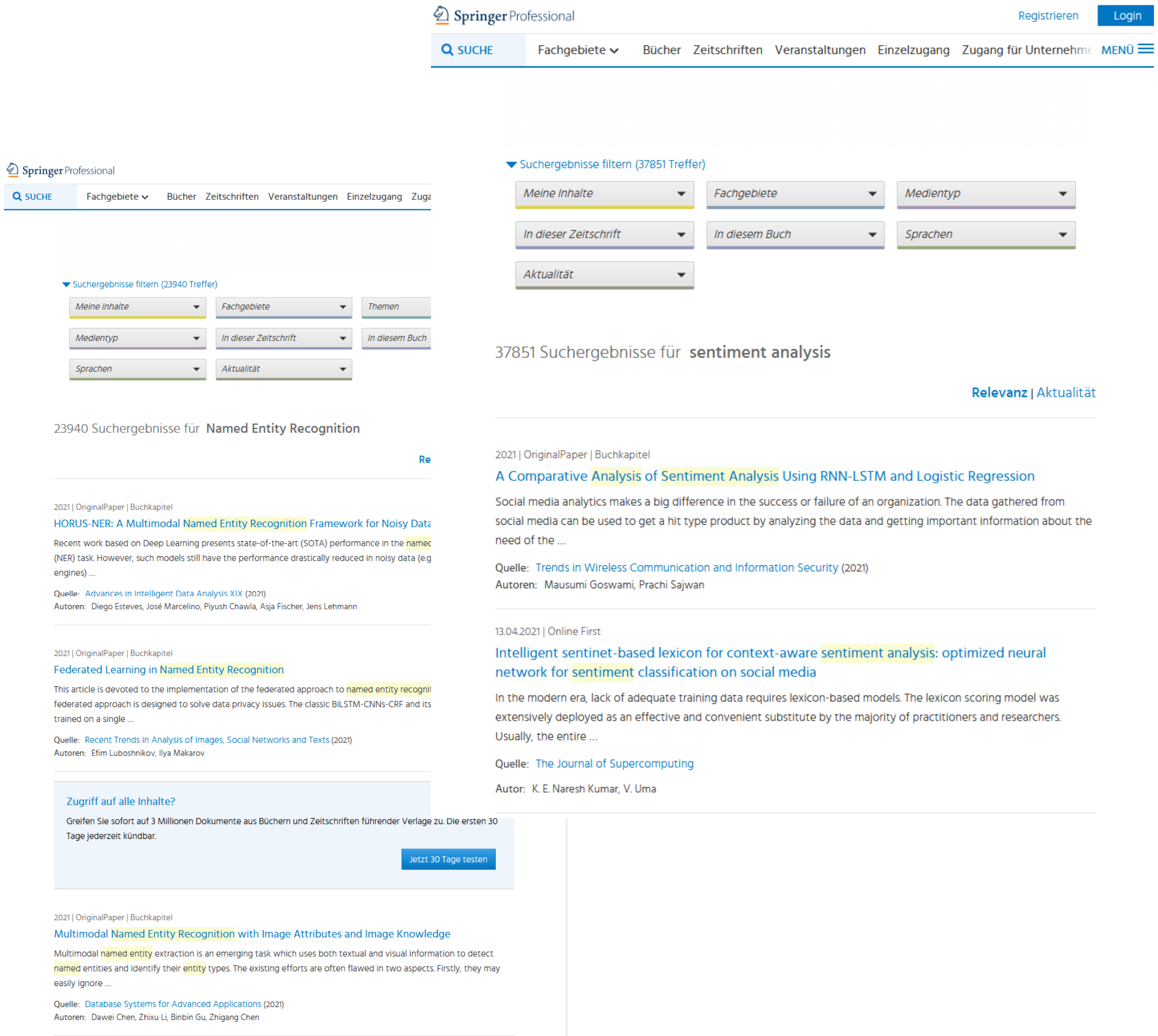

- URL: https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

# Comparision with other Search Engines / Portals



## Springer Professional

- **Very fast, very large corpus of 3 Mio. docs**
- **Default Setting for search:**
  - order by relevance
- **No indication for**
  - Keyword Frequencies
- **Sometimes**
  - Online First displayed




- **URL:** https://rob.by/en/Search/Pubs/Named,Entity,Recognition.html

# Comparision with other Search Engines / Portals



## Summary

- Google (Scholar), TIB, Springer Professional
  - Very fast, very large corpora
  - No indication for
    - Keyword Frequencies
  - No possibility to defined rule based triggers
  - Googe Alerts is Keyword based, but Search Results cannot by restricted by Actuality
  - UIs are non intuitive and require knowledge on advanced query techniques

- rOb.by – Finding without Searching
  - Fast, large corpus
  - Indication for
    - Keyword Frequencies
    - gives Valuable Feedback for Researchers
  - Allows to define rule based triggers
  - UI is intuitive and requires less expertise
  - Higher recall and precision based on lemmatizations and translations

- rOb.by-App with rule-based Alerting
  - With Deepl Multi-language Support
    - For translations
    - Group Chats
    - Query-results
  - iOS and Android platform support

- More Search Examples from Multiple Scientific Domains: https://rob.by/en/Search/Pubs/Examples/index.html

# Thank you …



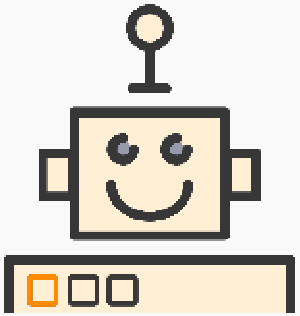
- Download rOb.by-App at
  - www.rob.by/de/App

## Die kostenlose rOb.by-App aus dem Apple Appstore oder von Google Play herunterladen

Lade die **rOb.by**-App herunter und lege noch gleich los — sie ist kostenlos!

Erstelle dein Benutzerkonto und erhalte den Zugang auf allen deinen mobilen Geräten.

Wir freuen uns immer, von dir zu hören! Sende uns bei Feedback, Fragen oder Anregungen an **feedback@rob.by**



Jetzt rOb.by-App downloaden und immer rechtzeitig informiert werden.

Download für Apple iOS | Download für Android