Natural Language Processing in der Praxis: Kategorisierung & Thesaurus-basiertes Indexieren

Stefan Geißler, März 2021

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Einführung

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- NLP (Natural language processing = Natürliche Sprachverarbeitung) ist ein wichtiges Teilgebiet der Informationsverarbeitung
- Viele wichtige Anwendungsfelder für Unternehmen und Organisationen
- Diese Vortragsreihe stellt Teilgebiete der NLP und typische Anwendungen vor

Vorstellung: Stefan Geißler

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- Studium Linguistik / Informatik
- Sieben Jahre Wissenschaftliches Zentrum *IBM* Heidelberg bis 2000
- 2000-2018 Co-Founder von *TEMIS,* frz-dt. Softwareunternehmen
- Verkauf von *TEMIS*, Co-Founder von *Kairntech* (www.kairntech.com) frz-dt. Softwareunernehmen
- DGI-Mitglied (Und bei der Gelegenheit: Danke an die DGI für die reibungslose Zusammenarbeit!)

Fahrplan

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- 10. Feb, 9:00-12:30: NLP Einführung, Anwendungen, Szenarien
- 24.Feb, 16:00-17:30: NLP@Kairntech: Maschinelles Lernen für Endanwender, Teil I: Entitäten-Extraktion.
- 10. März, 16:00-17:30: NLP@Kairntech: Nachbereitung/Fragen zu Entitäten-Extraktion
- 24.März, 16:00-17:30: NLP@Kairntech: Weitere Themen: Dokumenten-Kategorisierung, Annotierung mit Thesauri, …

Kategorisierung

Situation

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

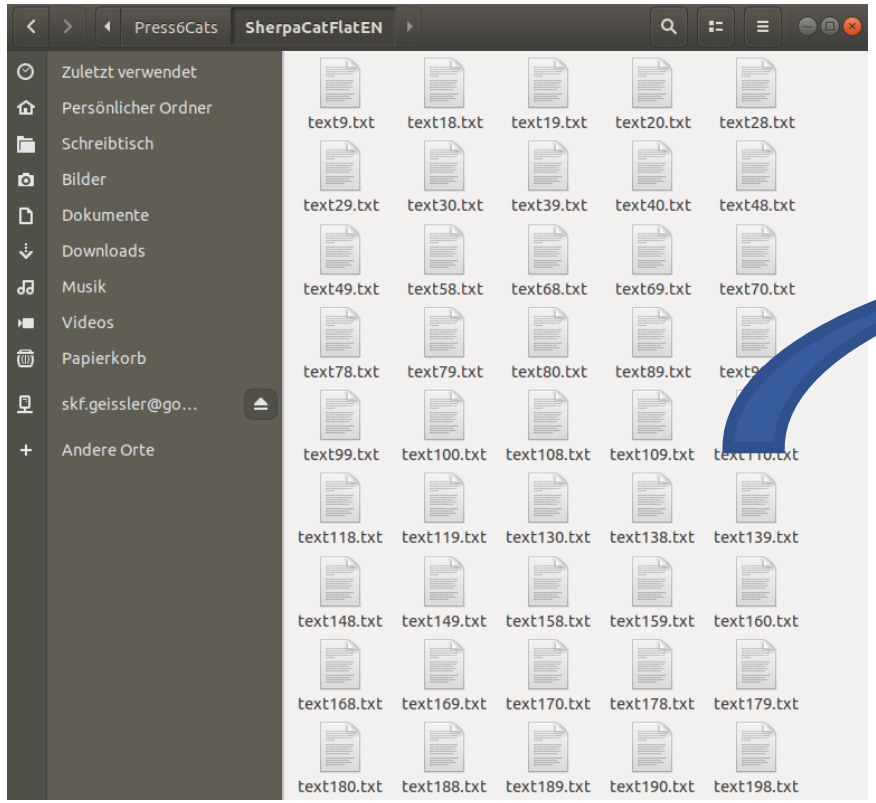
- **Ich muss eine große Anzahl von Dokumenten in entsprechende Ordner oder Kategorien sortieren**
- **Ich habe keinen Trainingskorpus**
- **Ich will nicht programmieren (aehm: Ich weiß auch nicht, wie das geht)**
- **Dennoch möchte ich von diesen neuen cleveren machine learning / AI / Deep Learning Technologien profitieren, von denen alle so begeistert sind**

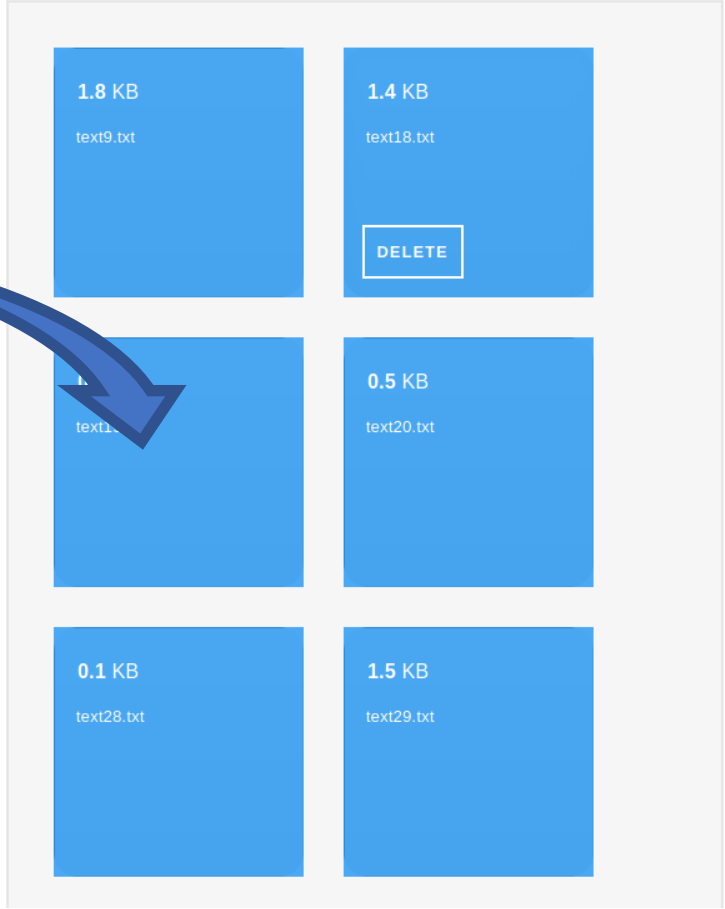
Lösung: Kairntech Sherpa!

Step 1: Import

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

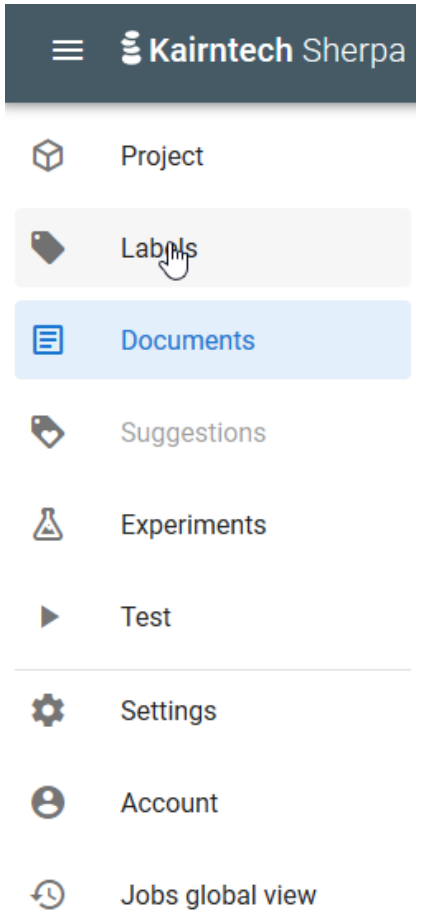
Import documents

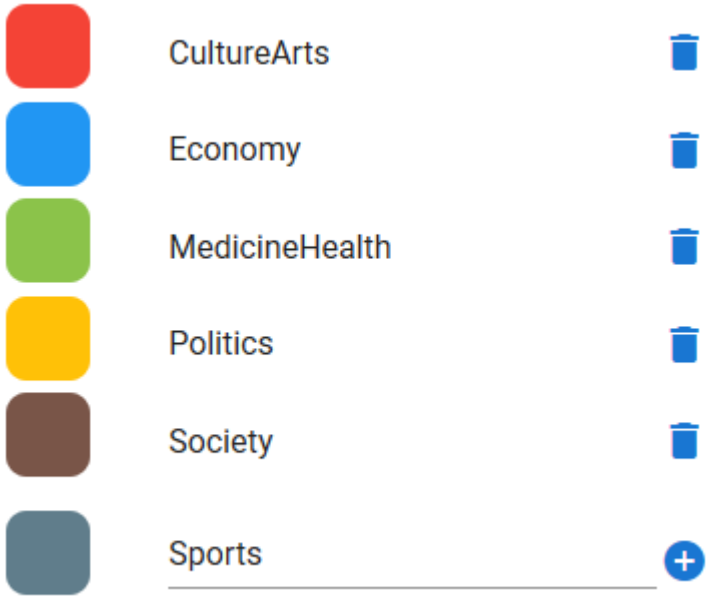




Step 2: Definieren Sie Ihre Kategorien

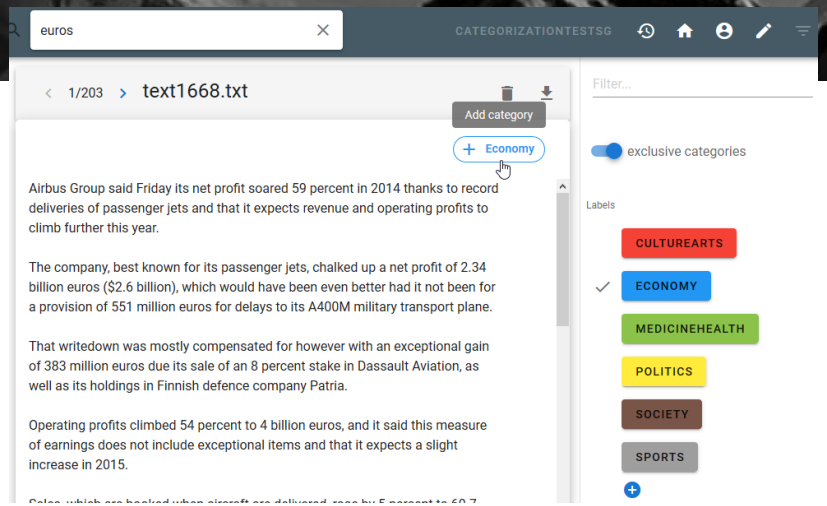
KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

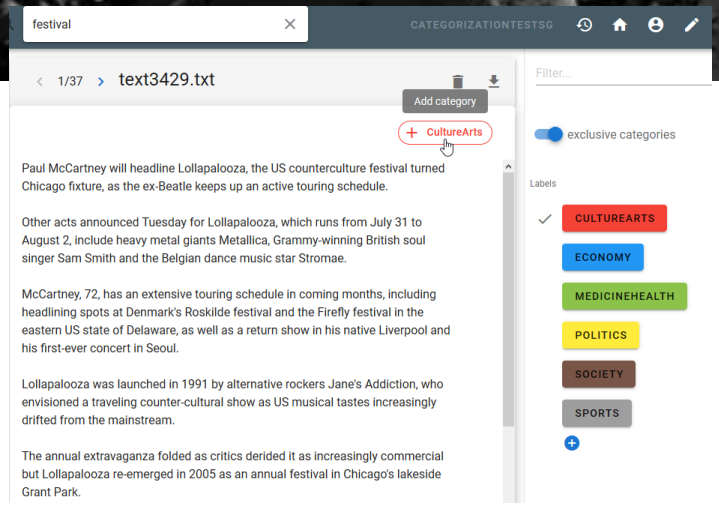


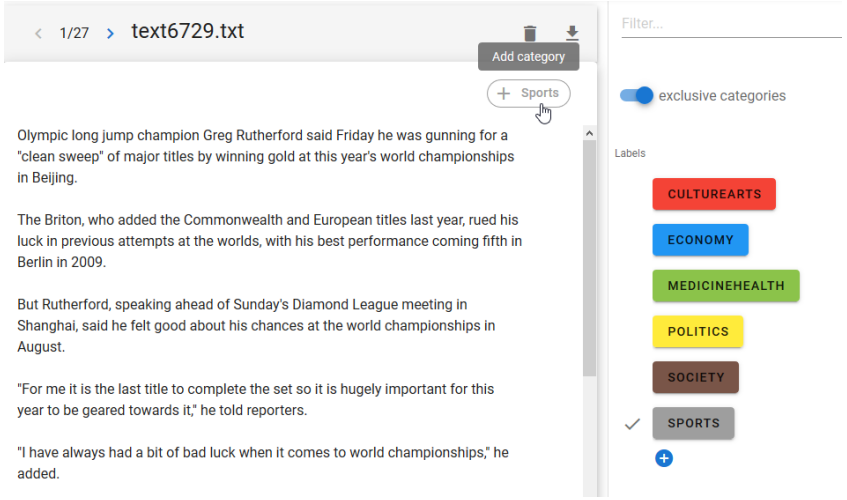


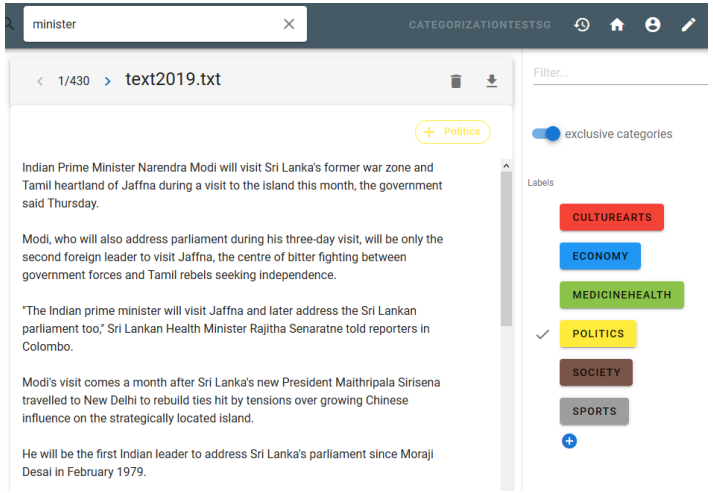
Step 3: Zuweisung von Dokumenten zu Kategorien

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT









Hier ist mein Domain-Know-how gefragt.

Annahme:
Viele (nicht alle) Dokumente, die von "Meisterschaften" sprechen, sind über *Sport*

"Minister" über *Politik*.

"Festival" *CultureArts*

"Euros" vielleicht *Wirtschaft* suggerieren

Ich gehe durch die Suchergebnisse für jeden dieser und sammele schnell 2-3 Dutzend Beispieldokumente für jedes.

Step 4: System beginnt zu lernen …

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

1/1992 text6160.txt

+ Society

Stade Francais president Thomas Savare on Thursday dismissed speculation that France lock Pascal Pape was set to move to Top 14 rivals Lyon next season.

"I am not aware (of a move)," Savare told AFP amid reports the 34-year-old international, who has 59 caps for France, was ready to leave.

"For me he is under contract with Stade Francais until June 2016. And we'll talk about his extension very quickly. Of course we want to keep him."

Lyon-born Pape had been linked with a move back to his home region despite their near certain relegation from the top flight next season.

Meanwhile the Stade Francais boss confirmed the arrival after the World Cup of Australia and Queensland Reds scrum-half Will Genia, 27, adding that they wanted to recruit some further cover for during the tournament.

"It's (scrum-half) a position which we wanted to strengthen," said Savare. "We had this opportunity and we seized it.

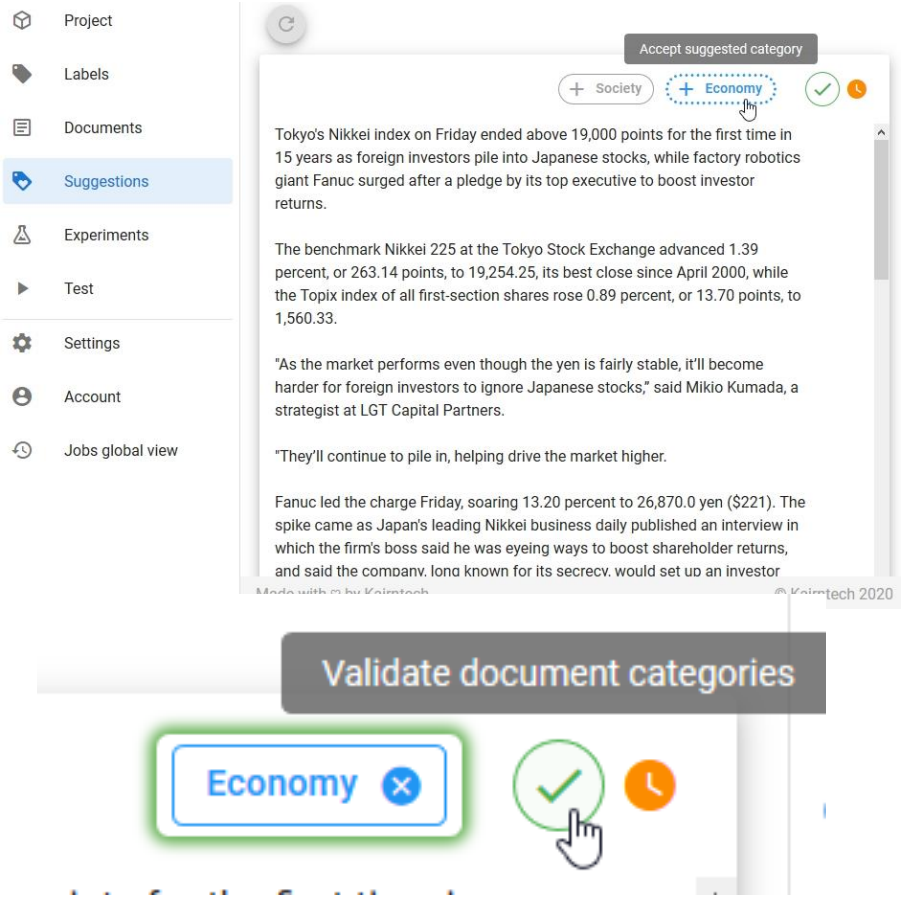
Made with ♡ by Kairntech © Kairntech 2020

Nach einigen Minuten der Zuweisung von Dokumenten zu Kategorien beginnt das System zu lernen.
... und generiert automatisch neue Vorschläge.

New suggestions available in the **suggestions panel**

Step 4: automatisch generierte Vorschläge

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT



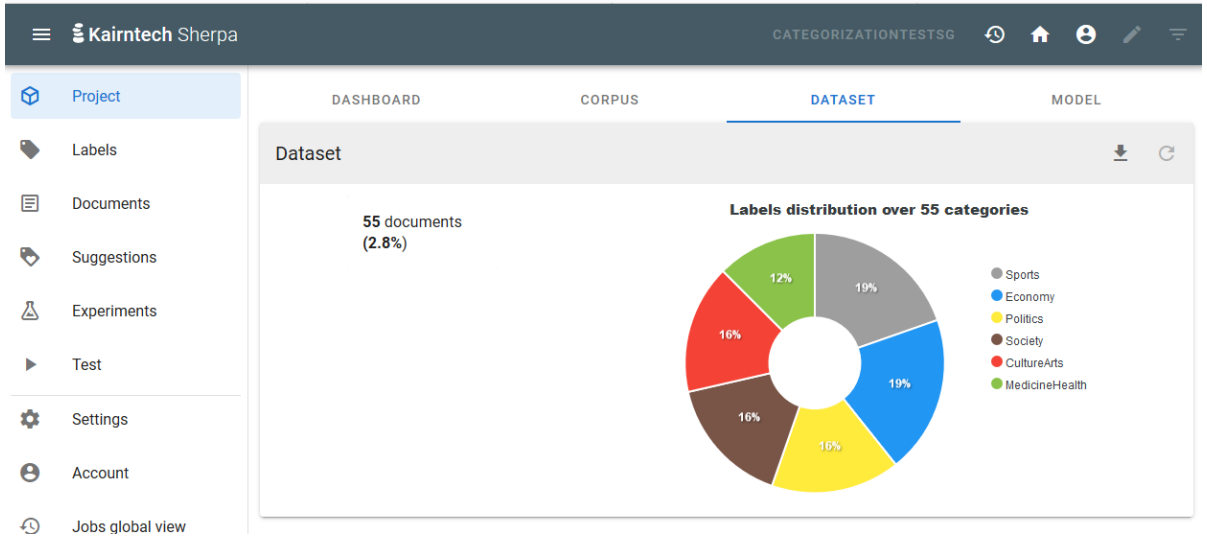
System präsentiert Dokumente, die es "denkt" sind über eine bestimmte Kategorie.

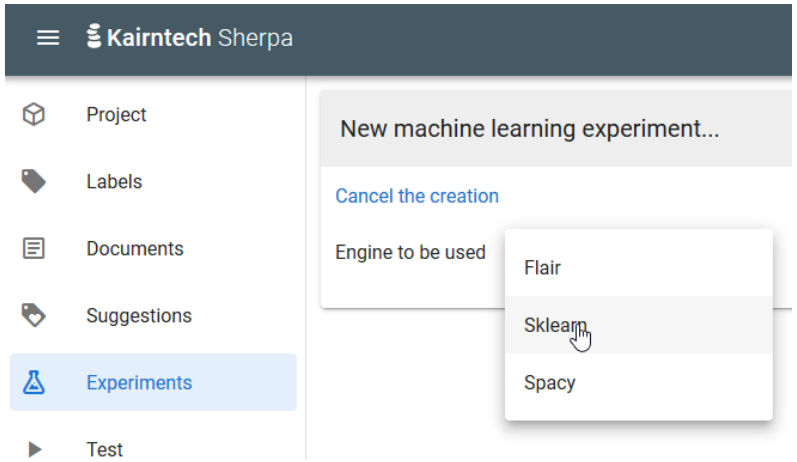
Nutzer/in akzeptieren oder weisen Vorschlag zurück und gehen zum nächsten Vorschlag auf.

Das ist schneller, als selbst Dokumente und ihre Kategorien auswählen zu müssen.

Step 5: Zwischenstand

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT





Wie viele Beispiele habe ich erstellt?

Ok, Zeit für eine erste Bewertung!

Step 6: Evaluierung

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

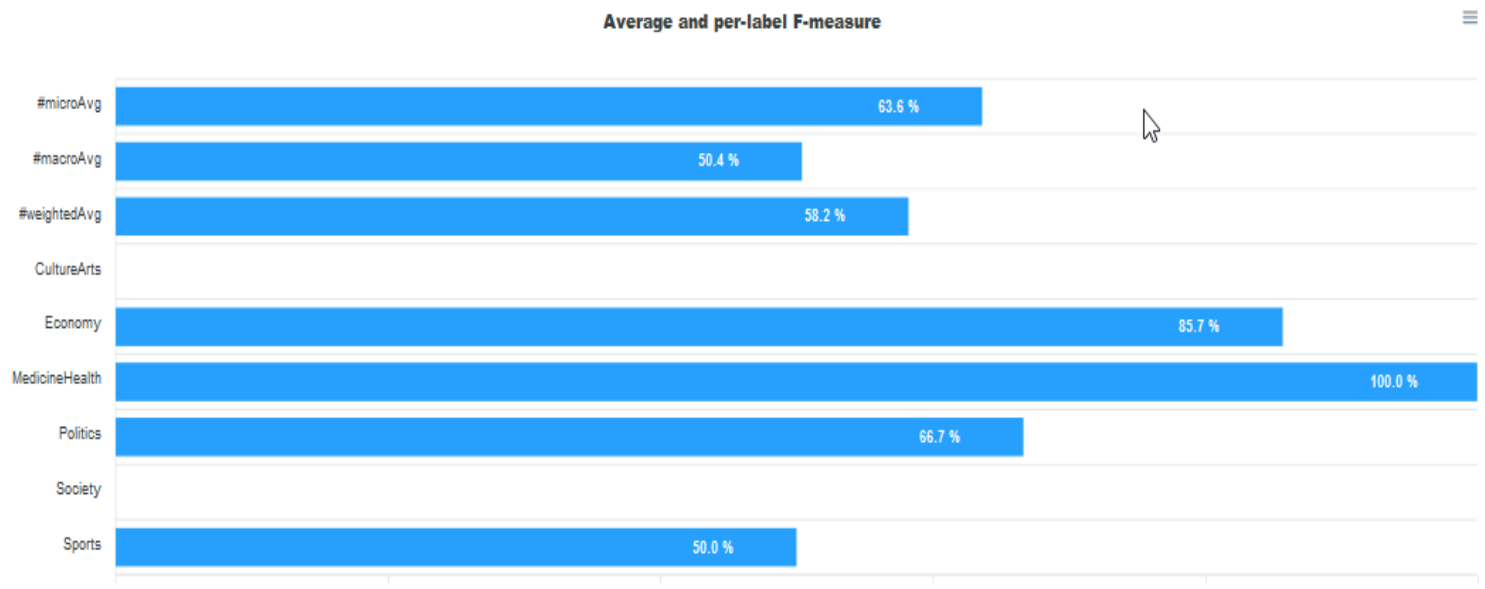
Machine learning experiments

Create new experiment...

Favorite	Name	Engine	Options	F-measure	Up-to-date	Last run	Actions	Test	Quality
☆	sklearn	Sklearn	mongo_fromcol: documents	63.6 %	✓	August 10, 2020 2:48 PM			☑

Quality report

F-measure



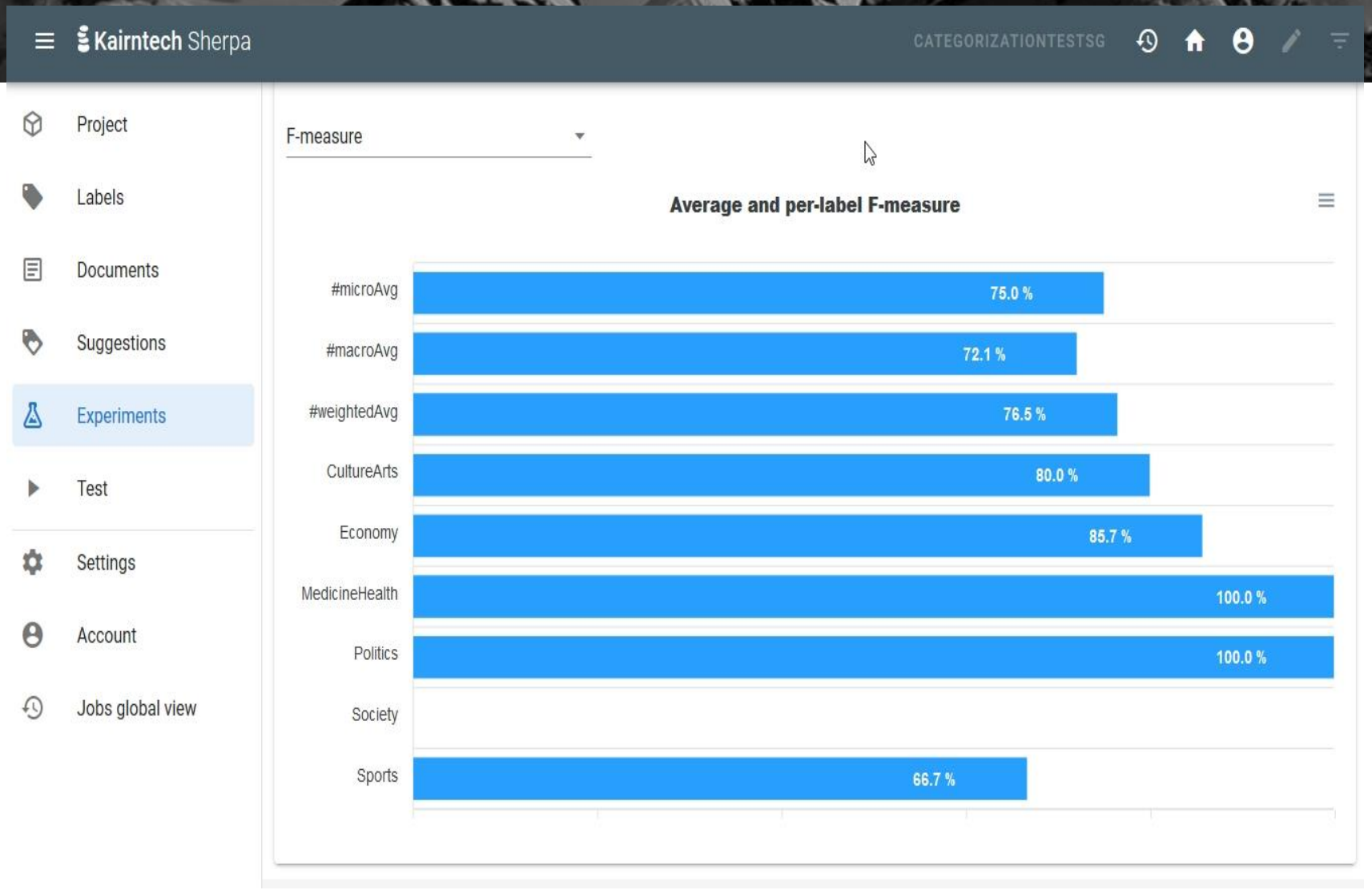
Ok, einige Kategorien sind bereits ganz ok.

"CultureArts" und "Society" scheinen noch einige Arbeit zu erfordern.

Gehen wir zurück, um einige weitere in diesen Kategorien zuzuweisen ...

Step 7: … nach weiteren Dokumenten

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

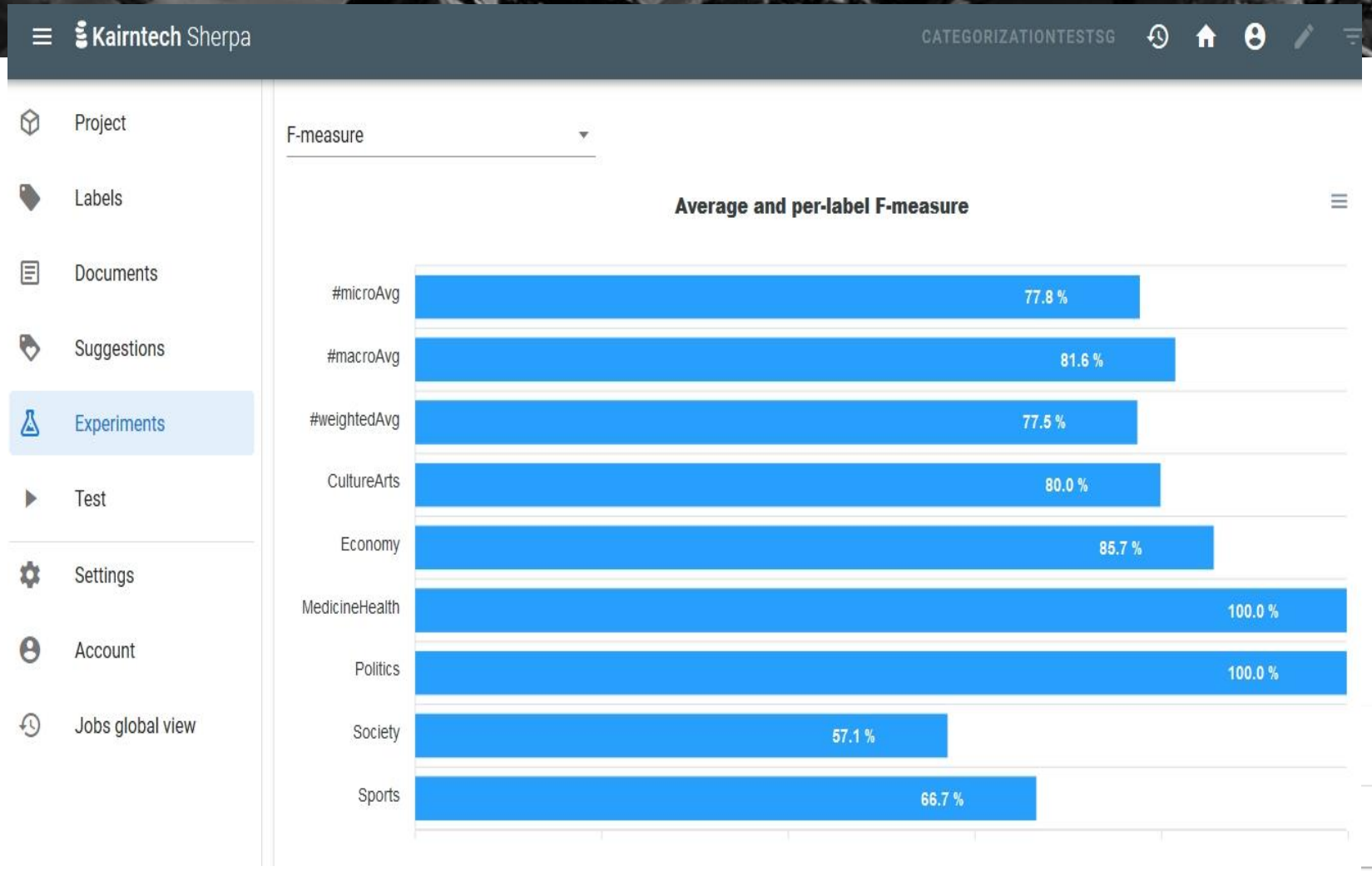


Ah, es wird.

Aber „Society" scheint noch etwas Aufmerksamkeit zu brauchen ...

Step 7: … schließlich

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT



Na also.

Bis hierher haben wir ca. 30-45 min verbracht.

Wir haben bei Null angefangen mit nicht kategorisierten Dokumenten, und nur unser Domain-Wissen verwendet

Nun bei einer Qualität von 78 %.

F-measure	Up-to-date	Last run
77.8 %	✓	August 10, 2020 4:00 PM

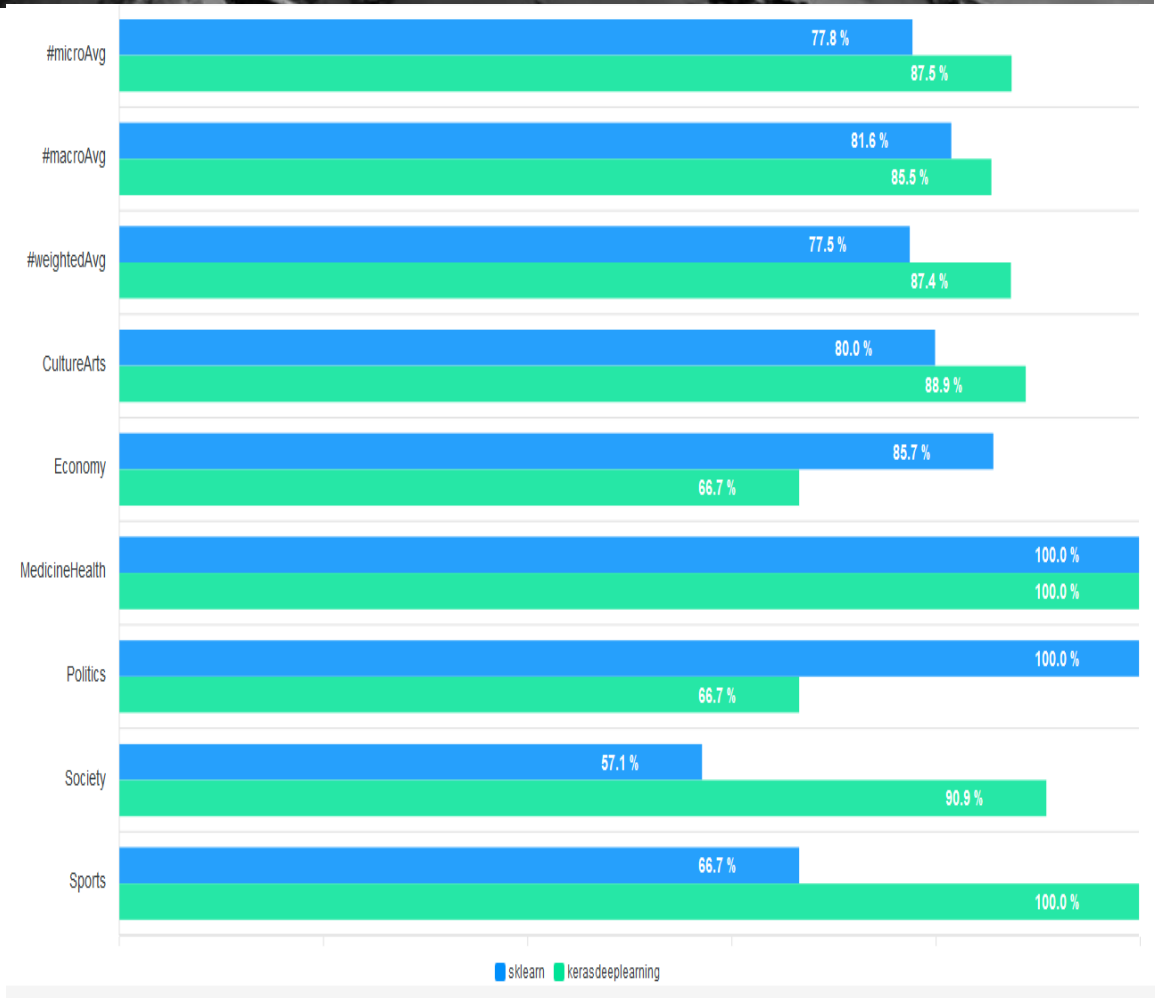
Step 8: Tuning

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- **Bisher haben wir "nur" die Standardkategorisierung verwendet (schnell, anwendbar auf die meisten Szenarien, aber nicht den Top-Ansatz, den Sherpa zu bieten hat)**
-
- **Wenden wir jetzt einen Deep Learning-basierten Ansatz an – alles verfügbar über die Web-GUI**
-

kerasdeeplearning	Sklearn	train_on: train, test_on: test, mongo_fromcol: documents, classifier: KerasMLPClassifier, featurizer: tfidf	87.5 %

- **Der Deep Learing-basierte Ansatz steigert die Qualität von standardmäßigen 78% bis zu 87% auf wenig manuell erstellten Trainingskorpus.**
-
- **Experimentier-/Trainingszeit bisher: weniger als 1 Stunde.**
-



Zeit für Praxis …

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Thesaurus-basiertes Indexieren

Einführung

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- Das Finden und Extrahieren von Konzepten und benannten Entitäten ist ein wichtiger Bestandteil vieler Dokumentanalyseprozesse
- Heute gibt es umfangreiches öffentliche Resourcen, die für diesen Zweck verwendet werden können
- Erforderlich: Umwandlung dieser Ressourcen in qualitativ hochwertige, umfangreiche Dienstleistungen
- Wir beschreiben Herausforderungen und Lösungen

Anreicherung von Dokumenten mit Weltwissen

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Anforderungen:

- Breite Abdeckung: Biowissenschaften, IT, Handel & Unternehmen, Recht, Allgemeines Interesse, ...
- Mehrsprachig
- Aktuell, Up-to-date

Entitätsextraktion ist mehr als nur das Abgleichen einer Zeichenfolge

- Scoring: Zuweisen von Konfidenzwert
- Typisierung: Unterscheiden Sie Orte, Personen, Substanzen, Körperteile, Krankheiten, ...
- Disambiguieren: Bestimmng der geeigneten Bedeutung unter mehreren Optionen
- Verknüpfen: Verbinden von Entitäten mit Hintergrundinformationen
- Hoher Durchsatz

Our choice: Wikidata

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- Community-Project : wikidata.org
- Entwickelt als Grundlage für andere Projekte, darunter Wikipedia
- 90Mio-Konzepte (und rasch wachsend)
- Wikidata ist eine Obermenge von vielen domänenspezifischen Vokabeln:
- MeSH, Geonames, DrugBank, ...
- ... und enthält Bezeichner/Links zu diesen
- CC-Lizenz
- Warum Wikidata? (Es gibt andere große öffentliche Vokabulare: DBPedia, Freebase, Yago, OpenCyc)

WIKIDATA

Wikidata

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- Wikidata ist 'FAIR'
  - FAIR data: 'Findable, Accessible, Interoperable, and Reusable'
  - Cf. Waagmester et al. 2019: 'Wikidata as a FAIR knowledge graph for the life sciences', doi: https://doi.org/10.1101/799684

- Wikidata schneidet gut ab im Vergleich einer Reihe von >30-Kriterien mit anderen öffentlichen Wissensquellen:
- Cf. Färber et al. 2017: „Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO ", doi: https://doi.org/10.3233/SW-170275

- Open source software package (Einsatzbereite Entity Recognition software) liegt vor (https://github.com/kermitt2/entity-fishing), übrigens implementiert vom Kairntech Chief ML Expert Patrice Lopez

Wikidata …

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Wikidata verfügbar

- Alleine der englische Teil: 13,5 GB komprimiert
- Aktualisierte Dumps werden regelmäßig (oft sogar wöchentlich) veröffentlicht.
- >90Mio Entitäten und >750 mio Verlinkungen
- Standardmäßig berücksichtigt Kairntech EN, DE, FR, SP, IT: 37 Mio. Entitäten
- Umwandlung der Dumps in den Inhalt unseres Dienstes bei Kairntech: 2 Tage Kompilierung
- Wikidata-Entitäten können mit Hunderten von spezifischen Eigenschaften mit Anmerkungen angezeigt werden. Z.B. Eigenschaft „P351" verbindet ein Gen mit seinem Identifikator in Entrez Gene

Wikidata: Life Science Beispiel

kairntech
From Documents to Business Impact

- Example taken from https://www.biorxiv.org/content/10.1101/799684v1.full.pdf

  "consider a pulmonologist interested in **identifying candidate chemical compounds** for testing in disease models. […] She may start by identifying genes with a genetic **association to any respiratory disease**, with a particular interest in **genes that encode membrane-bound proteins**. She may then look for chemical **compounds that either directly inhibit** those proteins, or finding none, compounds that **inhibit another protein in the same pathway**. Finally she may specifically filter for **proteins containing a serine-threonine kinase domain**"

```
SELECT DISTINCT ?compound ?compoundLabel
where
{
  # gene has genetic association with a respiratory disease
  ?gene wdt:P31    wd:Q7187 .
  ?gene wdt:P2293  ?diseaseGA.
  ?diseaseGA wdt:P279*  wd:Q3286546 .

  # gene product is localized to the membrane
  ?gene     wdt:P688          ?protein .
  ?protein  wdt:P681          ?cc .
  ?cc       wdt:P279*|wdt:P361*  wd:Q14349455 .

  # gene is involved in a pathway with another gene ("gene2")
  ?pathway  wdt:P31   wd:Q4915012 ;
        wdt:P527  ?gene ;
        wdt:P527  ?gene2 .
  ?gene2    wdt:P31   wd:Q7187 .

  # gene2 product has a Ser/Thrprotein kinase domain AND known enzyme inhibitor
  ?gene2     wdt:P688  ?protein2 .
  ?protein2  wdt:P129  ?compound ;
         wdt:P527  wd:Q24787419 ;
         p:P129    ?s2 .
  ?s2        ps:P129   ?cp2 .
  ?compound  wdt:P31   wd:Q11173 .

  FILTER EXISTS {?s2 pq:P366 wd:Q427492 .}

  SERVICE wikibase:label { bd:serviceParam wikibase:language"en". }
}
```

Wikidata: Beispiel aus den Life Sciences

KAIRNTECH

- Ergebnis auf https://query.wikidata.org/

?compound	?compoundLabel
<http://www.wikidata.org/entity/Q4850048>	balanol@en
<http://www.wikidata.org/entity/Q7376181>	ruboxistaurin@en
<http://www.wikidata.org/entity/Q5381479>	enzastaurin@en
<http://www.wikidata.org/entity/Q426386>	ingenol mebutate@en
<http://www.wikidata.org/entity/Q5957181>	staurosporine@en
<http://www.wikidata.org/entity/Q15269707>	palbociclib@en
<http://www.wikidata.org/entity/Q4063441>	alvocidib@en
<http://www.wikidata.org/entity/Q3494619>	seliciclib@en
<http://www.wikidata.org/entity/Q423111>	vemurafenib@en
<http://www.wikidata.org/entity/Q15708292>	cobimetinib@en
<http://www.wikidata.org/entity/Q4545713>	pyrazolanthrone@en
<http://www.wikidata.org/entity/Q421136>	sorafenib@en
<http://www.wikidata.org/entity/Q15723091>	dinaciclib@en
<http://www.wikidata.org/entity/Q19903515>	binimetinib@en
<http://www.wikidata.org/entity/Q7833138>	trametinib@en
<http://www.wikidata.org/entity/Q3891664>	regorafenib@en
<http://www.wikidata.org/entity/Q6584634>	Y-27632@en
<http://www.wikidata.org/entity/Q407699>	arachidonic acid@en
<http://www.wikidata.org/entity/Q3011604>	dabrafenib@en
<http://www.wikidata.org/entity/Q5007568>	PKR inhibitor@en
<http://www.wikidata.org/entity/Q5437132>	fasudil@en
<http://www.wikidata.org/entity/Q7448840>	selumetinib@en
<http://www.wikidata.org/entity/Q8063129>	ZM-447439@en
<http://www.wikidata.org/entity/Q6605919>	tozasertib@en
<http://www.wikidata.org/entity/Q5745805>	hesperadin@en
<http://www.wikidata.org/entity/Q7394936>	SU6656@en
<http://www.wikidata.org/entity/Q7119049>	PF-4800567@en
<http://www.wikidata.org/entity/Q198728>	ponatinib@en
<http://www.wikidata.org/entity/Q6554172>	linifanib@en
<http://www.wikidata.org/entity/Q894611>	bosutinib@en
<http://www.wikidata.org/entity/Q15633917>	alisertib@en

Wikidata annotation

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Szenario:

- Verwenden des Annotationsdienstes (REST Api) zum Anreichern von Dokumentinhalten
- Information ist getypt, mit Konfidenzwert, desambiguiert and verlinkt

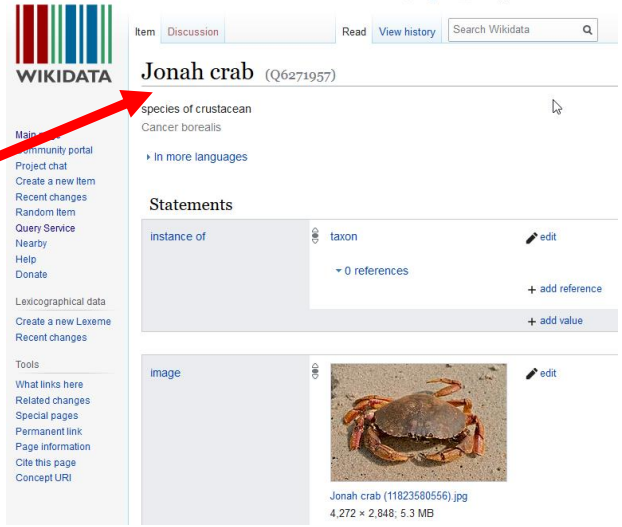
REST API

Annotation Result


```
{
  "start" : 236,
  "end" : 251,
  "labelName" : "misc",
  "score" : 1.0,
  "properties" : {
    "wikidataId" : "Q6271957",
    "preferredTerm" : "Jonah crab",
    "wikipediaExternalRef" : 6098993
  },
  "text" : "Cancer borealis",
  "label" : "Wikidata Concept"
}
```


DB record

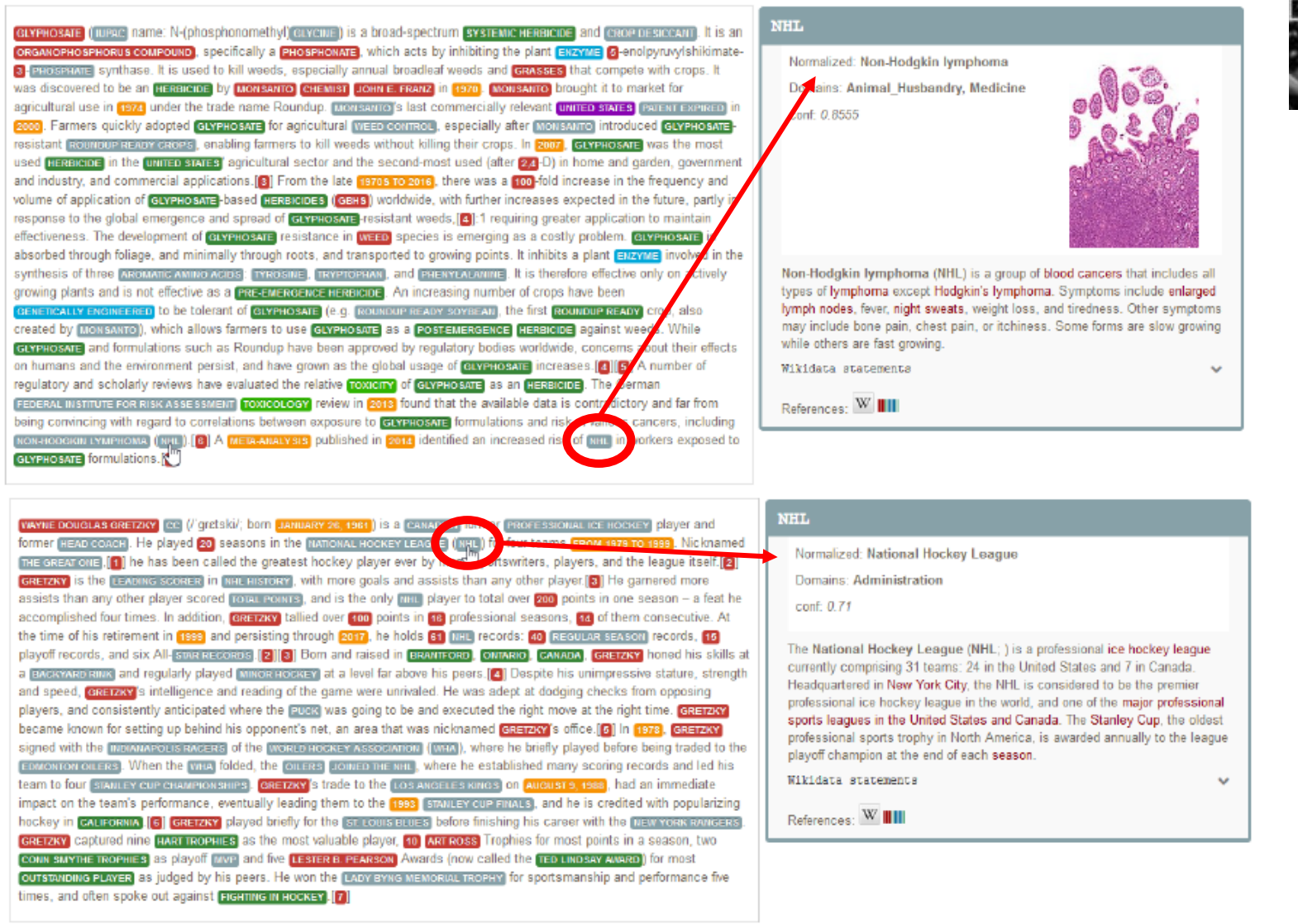
(mit links nach Freebase, NCBI, SealifeBase und vielen anderen)



Wikidata annotation

KAIRNTECH

DOCUMENTS TO BUSINESS IMPACT



Praktisch jeder nicht triviale Thesaurus ist massiv mehrdeutig: Terme können unterschiedliche Bedeutungen haben

Wann wird ausgewählt, welche Bedeutung?

Beispiel "NHL"

Unklarheit je nach Kontext: "NHL": das "Non-Hodgkin-Lymphoma" oder die "National Hockey League"?

Entity Extraktion innerhalb von Kairntech löst diese und viele andere, basierend auf automatisch erlernten Kontextinformationen auf (keine manuelle Wartung erforderlich)

Verlinkung: Das Vorkommen der extrahierten Begriffe wird mit Hintergrundinformationen aus Wikidata oder kundenspezifischem Wissen verknüpft

Weitere Beispiele

KAIRNTECH

FROM DOCUMENTS TO BUSINESS IMPACT

The EUROFIGHTER TYPHOON purchase will feature 55 units slated to replace the TORNADO, plus 38 additional TYPHOONS covered under the QUADRIGA PROJECT meant to phase out older TRANCHE 1 EUROFIGHTERS in the GERMAN AIR FORCE fleet with newer models.The SUPER HORNETS will be used in the nuclear-delivery role currently charged to the TORNADO fleet. The TORNADO is declared to NATO as a DUAL CAPABLE AIRCRAFT (DCA), meaning it may be tasked with the nuclear-delivery role utilizing the B61 nuclear GRAVITY BOMB. The EA-18G GROWLERS, meanwhile, would provide a replacement for the ELECTRONIC WARFARE MISSION CAPABLE TORNADO ECR variants. The GERMAN AIR FORCE operates [illegible] of these platforms. GERMANY wants to begin a PHASED WITHDRAWAL of its TORNADO fleet in 2025, around which time the newer fighters would ideally begin to arrive. In order to ensure a capability gap does not emerge, the TORNADO fleet would continue to serve with the GERMAN AIR FORCE until 2030, by which time the new fighters are to become operational.

TORNADO

Normalized: **Panavia Tornado**

Domains: **Aviation, Transport**

conf: *0.4496*

WtTableThe **Panavia Tornado** is a family of twin-engine, variable-sweep wing multirole combat aircraft, jointly developed and manufactured by Italy, the United Kingdom, and West Germany. There are three primary Tornado variants: the Tornado IDS (interdictor/strike) fighter-bomber, the suppression of enemy air defences Tornado ECR (electronic combat/reconnaissance) and the Tornado ADV (air defence variant) interceptor aircraft.

Wikidata statements

References: W

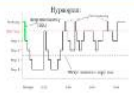
RAPID EYE MOVEMENT SLEEP (REM or RMS) is a unique phase of sleep in MAMMALS and birds, characterized by random/rapid MOVEMENT OF THE EYES, accompanied by LOW MUSCLE TONE throughout the body, and the propensity of the sleeper to dream vividly. The REM PHASE is also known as PARADOXICAL SLEEP (PS) and sometimes desynchronized sleep because of PHYSIOLOGICAL similarities to waking states, including rapid, low-VOLTAGE desynchronized BRAIN WAVES. Electrical and CHEMICAL ACTIVITY regulating this phase seems to originate in the BRAIN STEM and is characterized most notably by an abundance of the NEUROTRANSMITTER ACETYLCHOLINE, combined with a nearly complete absence of MONOAMINE NEUROTRANSMITTERS HISTAMINE, SEROTONIN, and NOREPINEPHRINE.[1]

REM

Normalized: **Rapid eye movement sleep**

Domains: **Medicine**

conf: *0.36*



Rapid eye movement sleep (REM sleep or REMS) is a unique phase of sleep in mammals and birds, distinguishable by random/rapid movement of the eye, accompanied with low muscle tone throughout the body, and the propensity of the sleeper to dream vividly.

Wikidata statements

References: W

Three people are dead and more than 172 injured after a rare TORNADO ripped through CUBA's capital, HAVANA. With WIND SPEEDS of up to 100km/h (62 mph), the first TORNADO to hit the city in decades UPROOTED TREES and CUT POWER in poor areas late on SUNDAY. President MIGUEL DIAZ-CANEL met emergency crews on the streets before dawn on MONDAY, and TWEETED that the damage was severe. Pictures posted on TWITTER showed homes destroyed and trucks overturned.

TORNADO

Normalized: **Tornado**

Domains: **Meteorology**

conf: *0.5844*

A **tornado** is a rapidly rotating column of air that is in contact with both the surface of the Earth and a cumulonimbus cloud or, in rare cases, the base of a cumulus cloud. The windstorm is often referred to as a **twister**, **whirlwind** or **cyclone**, although the word cyclone is used in meteorology to name a weather system with a low-pressure area in the center around which, from an observer

LOSING MY RELIGION is a song by the AMERICAN ALTERNATIVE ROCK band REM. The song was released as the first single from the group's 1991 album OUT OF TIME. BUILT ON A MANDOLIN RIFF, LOSING MY RELIGION was an unlikely hit for the group, garnering extensive airplay on radio as well as on MTV and VH1 due to its critically ACCLAIMED MUSIC video. The song became R.E.M.'s highest-CHARTING HIT in the UNITED STATES, reaching No. 4 on the BILLBOARD HOT 100 and expanding the group's popularity beyond its original fanbase. Out of Time garnered R.E.M. seven nominations at the 1992 GRAMMY AWARDS, the most nominations of any artist that year. The band won three awards: one for BEST ALTERNATIVE MUSIC ALBUM and two for LOSING MY RELIGION, Best SHORT FORM MUSIC VIDEO and Best POP PERFORMANCE BY A DUO or Group with Vocal

REM

Normalized: **R.E.M.**

Domains: **Administration, Commerce, Enterprise**

conf: *0.5537*

R.E.M. was an American rock band from Athens, Georgia, formed in 1980 by drummer Bill Berry, guitarist Peter Buck, bassist Mike Mills, and lead vocalist Michael Stipe. Additionally, many liner notes from the band's albums list attorney Bertis Downs and manager Jefferson Holt as non-musical members. One of the first alternative rock bands, R.E.M. was noted for Buck's ringing, arpeggiated guitar style, Stipe's distinctive vocal

Was sagen Kunden?

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Olivier Deguernel (Sealk.co)

> "Our processes require the **enrichment** of document content **with a wide range of named-entities**. We have analysed the available APIs on the market and have decided to integrate the Kairntech Named-Entity Extraction API into out offering. The **clear API, the superior quality and the wealth of information** returned by the API made it a valuable completion of our processes."

SEALK

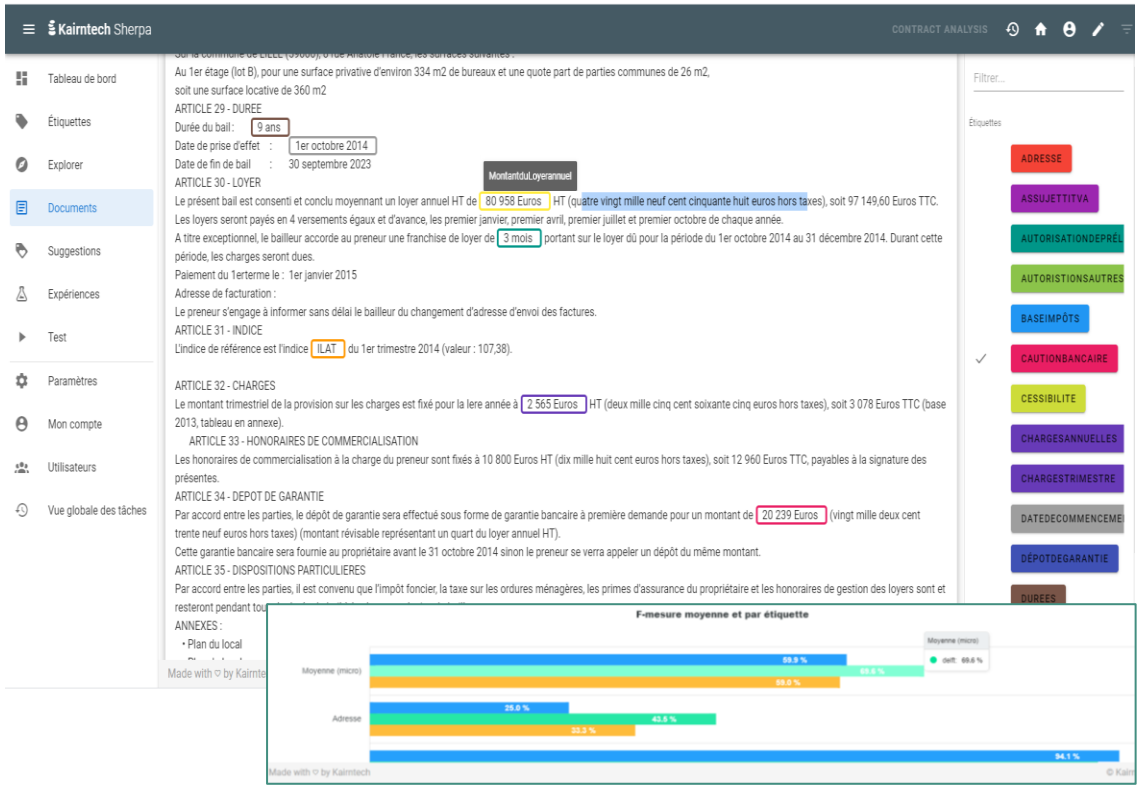
A Recommendation System Dedicated to M&A and Private Equity

Zusammenfassung

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Bei Kairntech verbinden wir in einer intuitiven Umgebung die Möglichkeiten,

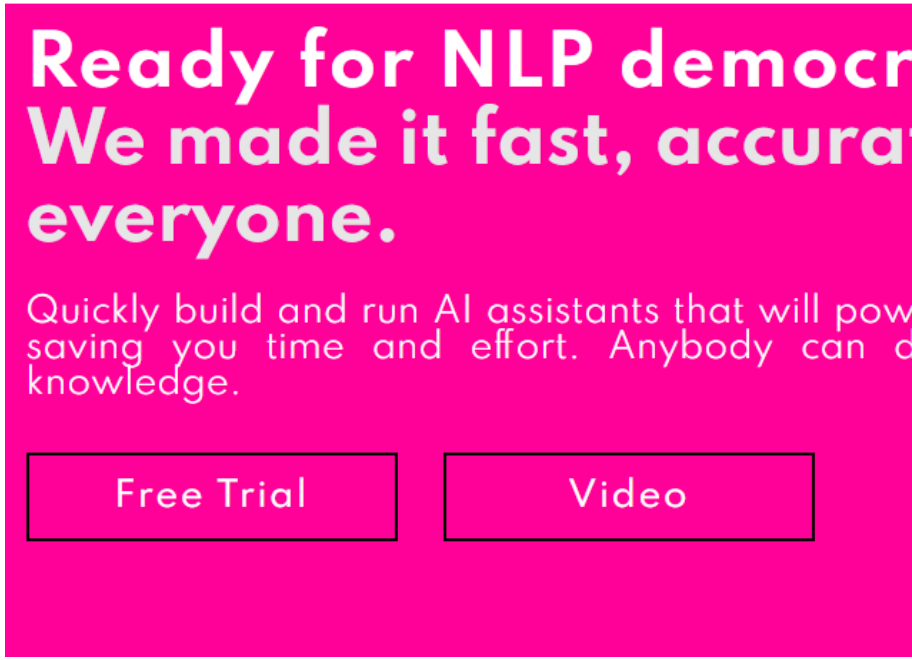
- ❖ Learn powerful models processing your own specific analysis demands
- ❖ With the analysis according to vast existing knowledge on just about any domain



Nächste Schritte

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

- Interesse, mit eigenen Daten an eigenen Fragen zu arbeiten? Fragestellung überlegen und Beispielsdaten vorbereiten
  - Fragestellung: z.B. Extraktion bestimmter Entitäten aus Dokumenten
  - Beispielsdaten: einige Dutzend Beispielsdokumente (bleiben Ihr Eigentum, werden nicht geteilt!)
  - Absprache unter info@kairntech.com (Subject: 'DGI Webinar') für Details zu Login, Formaten etc oder über die Kairntech Website
  - Bei Fragen: stefan.geissler@kairntech.com
- Zugang zu Kairntech-Software (unverbindlich+kostenfrei) bis Ende März 2021.



Q/A

KAIRNTECH
FROM DOCUMENTS TO BUSINESS IMPACT

Fragen?

Kommentare?

Diskussion?

Kritik?

Vorschläge?