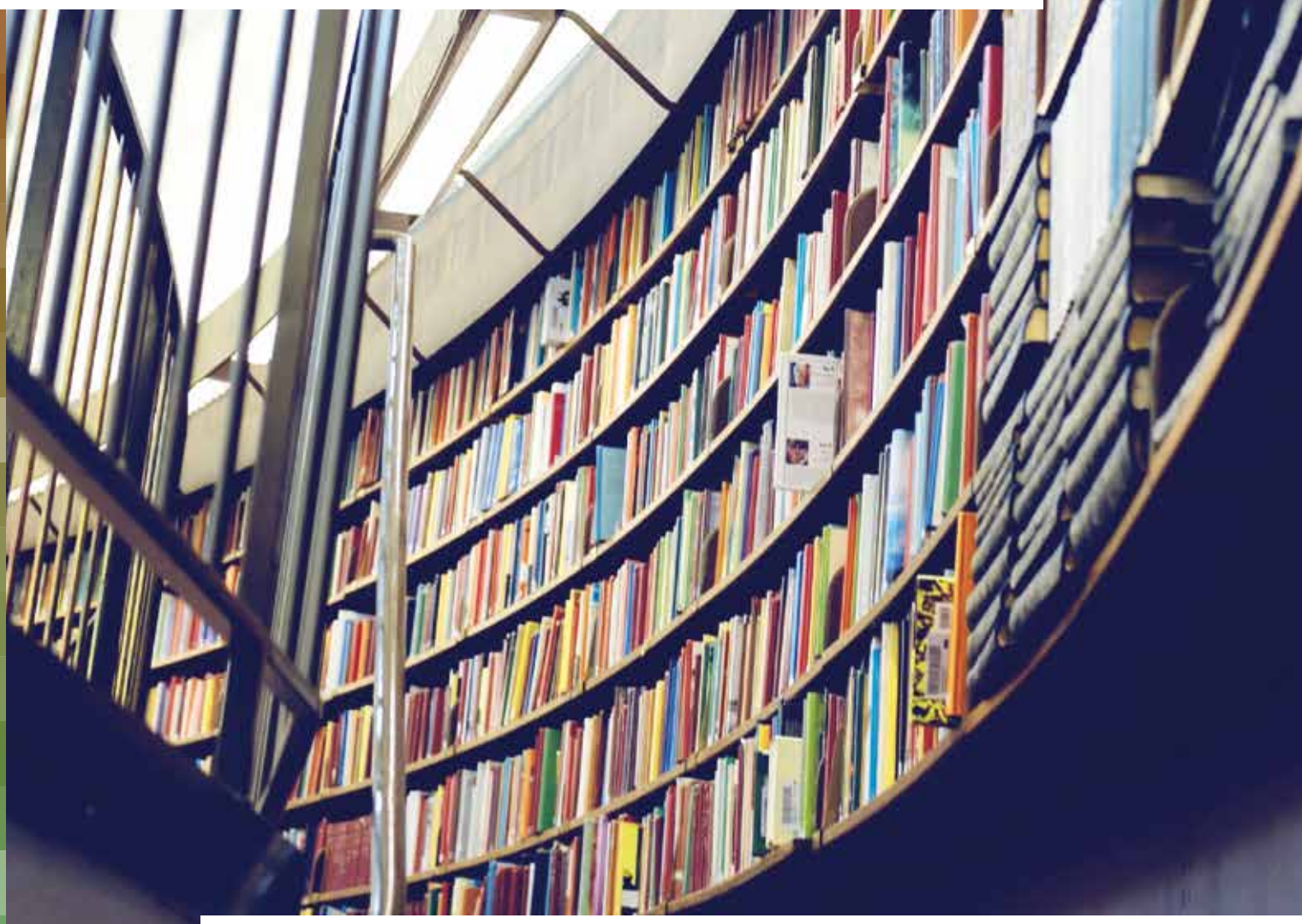
Datenblatt

Linguistik Plugin

Linguistik Plugin für Elasticsearch und OpenSearch

Semantisch-Linguistische Analyse für bessere Suchergebnisse

Mit seiner Linguistik-Erweiterung ermöglicht IntraFind den Nutzern von Elasticsearch und OpenSearch vollständigere und relevantere Suchergebnisse. Dabei ist das Plugin einfach zu integrieren und zu konfigurieren – auch bei Content in mehreren Sprachen. Bei der Elasticsearch-basierten Suchlösung iFinder ist das Linguistik Plugin fester Bestandteil.



INTRAFIND

Linguistik Plugin

INTRAFIND

Ihre Vorteile:

- + Bessere Suchergebnisse: höherer Recall & höhere Precision
- + Nichts übersehen: Vollständige Trefferliste durch Lemmatisierung und Kompositazerlegung auf höchstem Niveau
- + Lexikalisch-Morphologische Normalisierung: sehr hoher Grad an Lexikonabdeckung durch Verwendung von prozeduralen Lexika
- + Unterstützung für über 30 Sprachen
- + Einfach zu installieren, zu konfigurieren und zu benutzen: dank integrierter Spracherkennung keine komplexen Mappings / Settings erforderlich, auch nicht in mehrsprachigen Umgebungen
- + Automatische Sprachsegmentierung gemischtsprachiger Inhalte
- + Optionale Anbindung von Thesauri, auch für die crosslinguale Suche
- + Normalisierung von Zahlenangaben (z.B. Datumsangaben, Preise) und wissenschaftlich-technischen Einheiten (ca. 50 verschiedene Einheiten, z.B. Länge, Fläche, Volumen, Temperatur)
- + Optionale Erkennung von Entitäten wie Personen, Organisationen und Orten
- + Mächtige Query-Operatoren zur genauen Formulierung der Suchanfragen
  - Verschiedene Suchmodi – von exakter Suche bis zur Suche nach phonetisch ähnlichen Wörtern / Namen
  - NEAR-Operatoren, z.B. für die „Suche im Satz“ oder innerhalb einer definierbaren Wortdistanz
  - Operatoren für numerische Suche, Einheitensuche und Suche nach Entitäten

Beste Treffer – so hilfreich ist das Plugin

Das IntraFind Linguistik Plugin wird für die Suche nach „Autor“ auch Treffer mit dem Wort „Kinderbuchautor“ oder „Autorenteam“ liefern, nicht jedoch mit „Autoradio“ oder „Autorad“. Eine Suche nach „Buch“ wird auch Treffer für „Kinderbücher“ liefern, eine Suche nach „Bundesforschungsministerium“ Treffer für „Ministerium des Bundes für Forschung und Technologie“ und zwar allein auf Basis der morphologischen Normalisierung und Kompositazerlegung.

Für den Nutzer wird die Suche komfortabler, da er sich über Wortvarianten und Schreibweisen keine Gedanken mehr machen muss. Die Verwendung von Experten-Queries mit z.B. Wildcards wie *ministerium zur Suche nach allen Ministerien gehört der Vergangenheit an.

Neben der hochwertigen morphologischen Normalisierung gewährleistet die zusätzliche Zerlegung komplexer Token, dass die Produktsuche z.B. nach „iphone12“ und „iphone 12“ gleiche Trefferergebnisse liefert.

Das Plugin stellt verschiedene Suchmodi (z.B. neben der Suche im morphologisch normalisierten Index eine exakte Suche) und zusätzliche mächtige NEAR-Operatoren (z.B. Suche innerhalb eines Satzes) zur Verfügung. Als Bestandteil dieser NEAR-Queries sind beliebig komplexe Queries erlaubt, z.B. auch Wildcard Queries. Hierfür kommt eine erweiterte Query-Syntax zum Einsatz. Es wird selbstverständlich die komplette aus Elasticsearch gewohnte Query-Syntax unterstützt, auch einfache Queries wie die Match Query.

Ein nicht zu unterschätzender Vorteil des Linguistik Plugins ist die einfache Konfiguration gerade im multilingualen Umfeld. Der typische Aufwand, für jede Sprache ein eigenes Feld und einen eigenen Analyzer im Schema konfigurieren zu müssen, zusammen mit einer Steuerung der Indexierung mittels einer Spracherkennungskomponente, entfällt komplett. Dies liegt daran, dass die IntraFind Analyzer bereits eine Spracherkennung enthalten und sogar bei gemischtsprachigen Inhalten die in unterschiedlichen Sprachen abgefassten Teile erkannt und automatisch korrekt behandelt werden.

Premium-Linguistik für perfekte Suchergebnisse

Lemmatisierung, Wortstammnormalisierung und Kompositazerlegung

Im Linguistik Plugin wurden verschiedene Verfahren zur linguistischen Texterschließung kombiniert, um flektierte Wörter auf ihre Grundformen zu normalisieren (**Lemmatisierung + Wortstammnormalisierung**) und um zusammengesetzte Begriffe in ihre Grundbestandteile zu zerlegen (**Kompositazerlegung**). Letzteres ist besonders wichtig für die deutsche Sprache. Zusätzlich werden auch einfache Wortkategorien (Substantiv, Verb, Adjektiv, Funktional) geliefert.

Beispiele:
Lemmatisierung:
Bücher → Buch

Wortstammnormalisierung:
Arbeitslosigkeit → Arbeitslose, arbeitslos

Kompositazerlegung:
Bundesumweltminister → Bund, Umwelt, Minister

Die linguistischen Verfahren basieren auf umfassenden Lexikonbeständen und zeichnen sich zusätzlich durch eine stark prozedurale Orientierung in der Lexikon-Analyse aus. Linguistische Ergebnisse insbesondere für die Wortzerlegung werden also hierbei, im Gegensatz zu vergleichbaren Verfahren, stärker über Prozeduren „berechnet" und nicht im Lexikon „nachgeschlagen". Die umfangreichen Lexika sind hochqualitativ und effizient aufbereitet, so dass die Analyse schnell und speichersparend vollzogen wird. Da sich Sprachen dynamisch entwickeln und ein großes kreatives Potential zur Schöpfung neuer Wörter bzw. Wortkombinationen aufweisen, wäre es ineffizient und unpraktisch, sich nur auf starre Lexika zu verlassen.

Vor diesem Hintergrund arbeitet das Linguistik Plugin auf der Grundlage von Basislexika (Vollformen-Grundformen-Wortstamm-Mapping) mit den morphologischen Elementarbausteinen einer Sprache. Diese Bausteine können, ebenso wie die kombinatorischen Regeln für deren Analyse, mit einem hohen Vollständigkeitsgrad ermittelt und angewandt werden. Auf diese Art und Weise werden neue Wortkombinationen, vor allem Komposita, erfolgreich analysiert. Das Plugin bietet ein umfassendes Paket von Lexika und Prozeduren zur Analyse von neuen Wortschöpfungen. Außerdem werden die Lexika durch Neologismen regelmäßig erweitert.

Die Suche basierend auf unseren hochwertigen Lexika für Lemmatisierung und Wortstammnormalisierung und unserer Kompositazerlegung ist wesentlich präziser als die üblicherweise in Suchmaschinen eingesetzte Suche. Diese basiert auf einfachen algorithmischen Verfahren zur Wortstammbildung, die in vielen Fällen zu Übergeneralisierung neigen und Wörter mit völlig unterschiedlicher Bedeutung auf den gleichen künstlichen Stamm zurückführen (Beispiele: Messer → Mess, Messe → Mess).

Die Lemmatisierung bewirkt, dass eine Suche nach „Buch" auch Treffer für „Bücher" liefert, denn Worte werden auf ihre lexikalische Grundform normalisiert. Die darauf aufsetzende Wortstammnormalisierung sorgt zusätzlich dafür, dass für die Suche nach „Arbeitslosigkeit" auch der „Arbeitslose" gefunden wird und mit Hilfe der Kompositazerlegung wird auch der „Langzeitarbeitslose" gefunden.

Die Kompositazerlegung bewirkt nicht nur, dass eine Suche nach „Minister" auch Treffer für „Bundesumweltminister" liefern wird. Enthält die Suchanfrage ein zusammengesetztes Wort wie „Bundesumweltminister", wird automatisch eine Near-Query mit den Bestandteilen erzeugt, die auch einen Treffer für „Bundesminister für Umwelt" liefert.

Durch die Wortstammnormalisierung würde hier dann auch das „Bundesministerium für Umwelt" gefunden. Die verschiedenen Normalformen sind automatisch so gewichtet, dass exakte Treffer oder Treffer aufgrund der Lemmatisierung stärker gewichtet werden als Treffer, die auf Grund der Wortstammnormalisierung oder Kompositazerlegung entstehen.

Verwendung von Thesauri

In zahlreichen Suchprojekten wird die Suche durch die Verwendung von Thesauri qualitativ deutlich aufgewertet. Viele dieser so erreichbaren Verbesserungen liefert das Linguistik Plugin durch die Verwendung von Lemmatisierung und Kompositazerlegung bereits automatisch. Durch die Anbindung von Thesauri kann das Linguistik Plugin darüber hinaus auch Suchanfragen durch Synonyme und Abkürzungen (wie z.B. Auto,

Kfz, Kraftfahrzeug), Hyponyme (Unterbegriffe) oder Hyperonyme (Oberbegriffe) erweitern.

Das Plugin erlaubt die Verwendung beliebiger Thesauri und unterstützt übliche Thesaurus-Formate, wie z.B. SKOS. Durch die Verwendung von Lemmatisierung und Kompositazerlegung auch für den Lookup im Thesaurus müssen keine Flexionsformen im Thesaurus gepflegt werden, was die Pflege kundenspezifischer Thesauri und die Verwendung bereits existierender Thesauri deutlich vereinfacht. Mehrwortbegriffe werden vollständig unterstützt, auch für den Lookup. Wir bieten optional außerdem allgemeinsprachliche Thesauri für Deutsch, Englisch, Französisch und einen crosslingualen Thesaurus Deutsch-Englisch an.

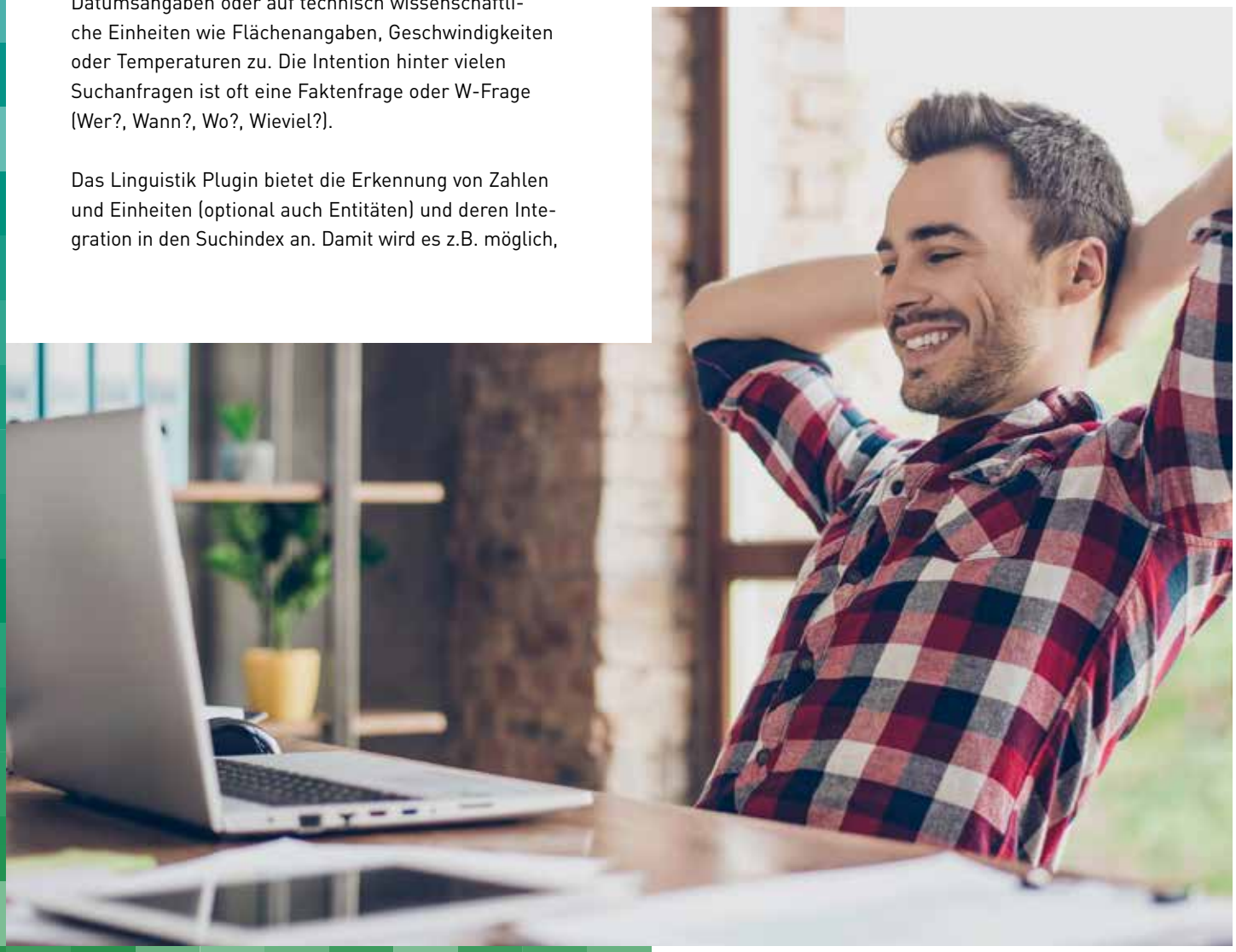
Suche nach Entitäten, Numerische Suche und Einheitensuche

Namen, z.B. von Personen, Organisationen und Orten (Adressen), spielen eine besondere Rolle in der Suche. Dies trifft auch auf Zahlenangaben wie Preise oder Datumsangaben oder auf technisch wissenschaftliche Einheiten wie Flächenangaben, Geschwindigkeiten oder Temperaturen zu. Die Intention hinter vielen Suchanfragen ist oft eine Faktenfrage oder W-Frage (Wer?, Wann?, Wo?, Wieviel?).

Das Linguistik Plugin bietet die Erkennung von Zahlen und Einheiten (optional auch Entitäten) und deren Integration in den Suchindex an. Damit wird es z.B. möglich, nach einer beliebigen Person in der Nähe (kleiner Wortab-stand oder gleicher Satz) von „gründen“ und „Ärzte ohne Grenzen“ zu suchen. Wird für diese Suchanfrage ein Dokument gefunden, so enthält das Highlighting Snippet mit hoher Wahrscheinlichkeit die Namen der Gründer des Vereins “Ärzte ohne Grenzen“.

Diese Anreicherung des Index ermöglicht z.B. auch die gezielte Suche nach einer Person mit Namen „Schwarz“ unter Ausschluss von Treffern für die Farbe „Schwarz“. Die Suche nach einem beliebigen Geldbetrag in einem Text in der Nähe von „Miete“ liefert die Antwort auf die Miethöhe eines Objekts.

Die numerische Suche kann nicht nur die Antwort auf Fragen nach „Wieviel?“ liefern. Die Suche nach den Wörtern „Siedepunkt“ und einer „Temperatur von 90 – 110 °C“ im gleichen Satz liefert dank der Einheitennormalisierung auch Treffer für den Satz „Der Siedepunkt von Wasser liegt bei 212 °F“. Die Suche nach „Display 5 Zoll“ wird auch Treffer für die Sätze „Sein neues Smartphone hat eine Displaydiagonale von 4.8 Zoll“ oder „Das Display hat eine Diagonale von 12 cm“ liefern.



Leistungsmerkmale des Linguistik Plugins am Beispiel Deutsch:

- Derzeit über 700.000 Einzelwortlexeme (Simplex-Lexeme) im deutschen Lexikon und über 100 prozedurale Regeln, z.B. für die Behandlung von Straßennamen wie Berliner Straße oder Berlinerstraße
- 50.000 Fachterme und Eigennamen für Deutsch
- Optionale Suche nach Eigennamen (Personen, Organisationen, Orte, Adressen, etc.)
- Jährliche Aktualisierung mit Wortneubildungen (Neologismen)
- Unknown Word Recognition: unbekannte Wörter (meist seltene Eigennamen) werden als solche markiert und einer speziellen „Plural-s"-Normalisierung unterzogen.
- Unterstützt neue und alte deutsche Orthographie
- Numerische Suche und Suche nach Einheiten
- Optional: allgemeinsprachlicher Synonymthesaurus Deutsch
- Optional: Deutsch-Englisch-Lexikon für crosslinguale Suche
- Multilingualität: sehr einfache Konfiguration im multilingualen Umfeld (Content in mehreren Sprachen)
- Mixed Language Documents: werden korrekt prozessiert, Feststellung des Sprachwechsels im Text
- Satzgrenzen-Erkennung für noch bessere Relevanz und Suche innerhalb eines Satzes
- Verschiedene Suchmodi
- Beliebig geschachtelte NEAR-Operatoren (Phrase mit Wildcard Queries: NEAR/1(N* York))

Neben einfachen Volltext-Queries erlauben mächtige Suchoperatoren die Formulierung sehr präziser Suchanfragen:	
MODE/E&MAN	exakte Suche nach der Firma MAN (groß geschrieben) -> keine Treffer für man (klein geschrieben)
MODE/P(Tschaikowsky)	phonetische Suche nach ähnlich klingendem Namen
NEAR/0(func* AND NOT function)	Suche nach Wörtern, die mit func beginnen; ignoriere Treffer für das Wort function
NEAR/0(Schwarz AND ENTIY/PERS)	Suche nach einer Person mit Namen Schwarz
UNIT/(300 tausend km/s; 10.000 km/s)	Suche nach einer Geschwindigkeit in der Nähe der Lichtgeschwindigkeit
NEAR/S(UNIT/*(Hektoliter) AND Bier)	Suche nach einem Satz, der Bier und eine Volumenangabe enthält
NEAR/S(Miete AND UNIT/(1.000 €; 100€))	Suche nach dem Wort Miete und einem Geldbetrag von 1000 € +/- 100 € im selben Satz.
NEAR/20((ENTITY/PERS OR ENTITY/ EMAIL) AND "SAP Account")	Suche nach einem Personennamen oder einer Email-Adresse in der Nähe des Mehrwortbegriffs SAP Account. Das könnte die Frage beantworten, von wem man einen SAP Account bekommen kann.

Verfügbare Sprachen

Deutsch	Griechisch	Türkisch	Japanisch	**Fehlt Ihnen eine Sprache? Sprechen Sie uns an!**
Englisch	Tschechisch	Ukrainisch	Thailändisch	
Französisch	Portugiesisch	Norwegisch	Indonesisch	
Italienisch	Ungarisch	Dänisch	Malaiisch	
Spanisch	Rumänisch	Schwedisch	Hindi	
Niederländisch	Slowakisch	Finnisch	Chinesisch*	
Polnisch	Slowenisch	Russisch		
Kroatisch	Arabisch	Koreanisch		

* (Festland + traditionell / Taiwanesisch inklusive Normalisierung)

Lieferumfang

IntraFind liefert ein komplettes Softwarepaket als **Plugin** für **Elasticsearch und OpenSearch**, bestehend aus dem qualitativ hochwertigen **Linguistic Analyzer** für über 30 Sprachen und einem erweiterten **Query Parser**. Im Linguistic Analyzer ist die Funktionalität zur Erkennung von Zahlenangaben und Einheiten bereits integriert. Optional wird eine Thesaurus-Komponente zur Query-Expansion und eine Tagging-Komponente zur Erkennung von Entitäten (Standard-Entitäten wie Personen, Organisationen, Orte, Adressen, aber auch kundenspezifischen Entitäten) angeboten.

INTRAFIND

IntraFind Software AG
Landsberger Straße 368
80687 München
Deutschland

+49 89 3090446-0
sales@intrafind.com
www.intrafind.com