semantha®

the semantic platform

Die Künstliche Intelligenz,
die nicht nur liest
sondern auch **versteht**.

#BeADocumentHero

29. Mai 2020

hatched by
THINGS
THINKING

Finde gesuchte Inhalte unabhängig der Formulierung.

Die Künstliche Intelligenz semantha® versteht natürliche Sprache unabhängig von der Wortwahl. Sie ist in der Lage, große Mengen Text zu lesen und findet gesuchte Inhalte - ganz gleich wie diese formuliert wurden.

Das Auto war sehr schnell.

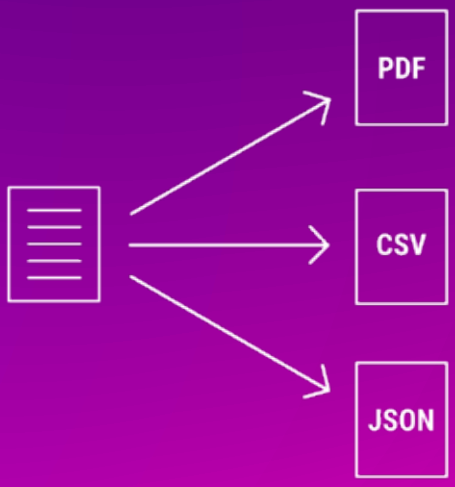
Das Fahrzeug war zügig unterwegs.

1000 Dokumente? Kein Problem!

semantha® kann gesuchte Inhalte in großen Textmengen in wenigen Augenblicken finden und verschiedene Dokumente direkt miteinander vergleichen.

Flexibler Output oder Transfer.

semantha® ist eine flexible Plattform und kann die Ergebnisse in verschiedenen Formaten exportieren (z. B. xlsx, docx, xml etc.) und die gefundenen Inhalte direkt an eine andere Software übergeben. Hierfür stehen verschiedene Benutzeroberflächen und eine standardisierte REST-Schnittstelle zur Verfügung.



Das FORBES Magazin beschreibt das KI-Startup thingsTHINKING aus Karlsruhe mit den Worten Hirn mit **künstlicher Intelligenz (KI).** Nach über 14 Jahren Forschung im Bereich der Sprachverarbeitung und KI gründeten wir 2017 die thingsTHINKING GmbH. Wir haben zahlreiche Preise gewonnen und bedienen Kunden in den Bereichen Automobilbau, Chemie, Pharma, Recht, dem Bauwesen, Wirtschaftsprüfung und Versicherungen. Unsere Plattform **semantha®** ist in der Lage, Dokumente auf Bedeutungsebene zu vergleichen und bleibt nicht bei einem Wortvergleich stehen.

Marc fährt mit dem Zug zur Arbeit.

Um in das Büro zu kommen fährt Marc mit der Eisenbahn.

Mit einer herkömmlichen Suche würde man in diesem Beispiel keine Übereinstimmung finden. semantha® sagt jedoch, dass diese beiden Sätze eine sehr ähnliche Bedeutung haben. Ein Mensch würde zu einem sehr ähnlichen Ergebnis kommen.

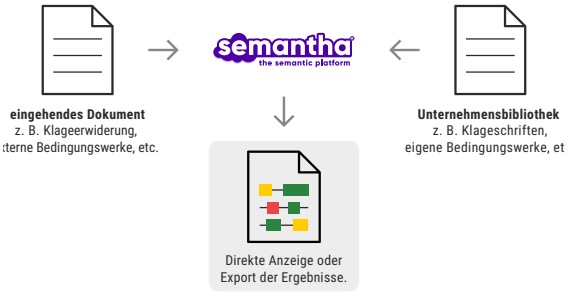
Eine einzigartige Kombination verschiedener KI-Elemente im Kern von semantha® verleiht ihr das Sprachverständnis, das semantische Analysen auf unstrukturierten Daten (Text) erlaubt. Texte kann sie somit unabhängig von ihrer Formulierung lesen, verstehen, vergleichen und die Bedeutung speichern. Dabei bestimmt der Kunde selbst, worauf semantha® bei der Analyse achtet. Dokumentengetriebene Prozesse können dadurch automatisiert und/oder beschleunigt werden. Der Mehrwert ist dort am größten, wo Menschen viel Text lesen müssten um Aufgaben zu erfüllen oder Unternehmen gezielt Informationen manuell aus unstrukturierten Daten ziehen müssen.

Unsere Zielgruppe sind derzeit Unternehmen, die über dokumentengetriebene Prozesse verfügen und diese in Bezug auf Qualität, Wiederholbarkeit, Beschleunigung und Automatisierung verbessern wollen.
Wir bieten eine SaaS-Lösung (Software as a Service) an und beginnen mit einer zweimonatigen Testphase, in der wir in enger Abstimmung mit dem Kunden die Anwendungsfälle ausgiebig testen. Anschließend folgt eine jährliche Lizenz und der Kunde kann die Plattform individuell in die eigene IT-Struktur mittels einer Schnittstelle (REST API) integrieren.

Einsatzmöglichkeiten:

Ⓐ Definiere Inhalte, die wichtig sind und prüfe, ob ähnliche Formulierungen vorhanden sind.

Ⓑ Vergleiche mehrere Dokumente miteinander und finde sofort die Übereinstimmungen.



Um den vollen Funktionsumfang von semantha® kennenzulernen und deine zeitintensiven Prozesse zu beschleunigen, schicke uns einfach eine Mail an **info@thingsTHINKING.net**

Ausführliche Informationen

semantha® verarbeitet unstrukturierte Dokumente in drei aufeinanderfolgenden Schritten:

- Schritt 1 ermittelt semantische Ähnlichkeiten zwischen Textpassagen.
- Schritt 2 extrahiert aus Dokumenten Datenpunkte und strukturiert sie dadurch.
- Schritt 3 verwendet die Datenpunkte, um logische Schlussfolgerungen zu ziehen.

Dafür kommen verschiedene KI-Elemente in Kombination zum Einsatz. Wir sind davon überzeugt, dass sich viele Anwendungsfälle nicht mit nur einem KI-Element lösen lassen, sondern gerade die Kombination verschiedener KI-Zutaten zum Erfolg führt.

Alternative Softwarelösungen ohne KI können abweichende Formulierungen kaum oder nur schwer miteinander vergleichen und sind oft nicht in der Lage, Textdokumenten unabhängig von konkreten Formulierungen miteinander zu vergleichen.

Im Folgenden erläutern wir, welchen KI-Elementen des Periodensystems die Verarbeitungsmechanismen zuzuordnen sind.

Auf der Eingabeseite muss semantha® zunächst ein Dokument einlesen und in eine interne Datenstruktur überführen. Liegen die Daten semistrukturiert vor, z.B. in einem Word-Dokument, greift semantha® auf die vorhandene Dokumentstruktur zurück, um bspw. Kapitelabschnitte anhand von Überschriften einzuteilen. Liegen die Eingabedaten in weniger strukturierter Form, z.B. in einem PDF-Dokument vor, greift semantha® nicht nur auf den textuellen Inhalt zu, sondern auch auf visuelle Eigenschaften **(Element Ir - Image Recognition)**. So müssen beispielsweise Tabellen erkannt und gesondert ausgewertet werden oder Texte in Abbildungen sowie in Kopf- und Fußzeilen übersprungen (also nicht dem Fließtext zugeordnet) werden; im selben Zug werden optische Eigenschaften des Textes genutzt, um Rückschlüsse auf die erweiterte Dokumentstruktur zu ziehen (bspw. Erkennen von Überschriften und Marginalien). Eine derartige Aufbereitung ist notwendig, da der reale Einsatz nicht auf “technisch schöne” Dateiformate eingeschränkt werden kann. Strukturierte Dateiformate, wie bspw. XML-Dialekte, können ebenso als Datenformat verwendet werden.

Die nächste Stufe der Verarbeitung ist die Ermittlung semantischer Ähnlichkeiten **(Element Lu - Language Understanding).** Hierbei werden Textpassagen miteinander abgeglichen und auf überlappenden Inhalt geprüft. So können gleiche oder nahe beieinanderliegende Textpassagen identifiziert werden, um ein Dokument bspw. auf zuvor definierte Hotspots zu überprüfen. Ebenso ist es möglich, zwei Dokumente (oder zwei Versionen eines Dokuments) direkt miteinander zu vergleichen.

In dieser Stufe sind zwei Eigenschaften von semantha® besonders hervorzuheben: Erstens verfügt semantha® über ein vorgegebenes Sprachverständnis. Durch dieses können Aufgaben aus dem Stand heraus bearbeitet

werden, ohne dass zuvor ein anwendungsfall- oder kundenspezifisches Training nötig wäre. Das Sprachverständnis ist jedoch von der Sprache selbst abhängig, weswegen wir je nach Sprache verschiedene Sprachmodelle verwenden. Das Sprachmodell ist jedoch nicht statisch, sondern kann mit verschiedenen Methoden auf den Anwendungsfall angepasst werden. Im einfachsten Fall sind dies konfigurative Änderungen, je nach Fachsprache/Jargon kann jedoch auch ein gesamtes Sprachmodell basierend auf Kundendokumenten "maßgeschneidert" werden. Die zweite Eigenschaft der Textanalysen von semantha® ist die unabhängigkeit vom Wortlaut: das Sprachmodell ermöglicht es semantha®, Texte auf Bedeutungsebene zu erfassen wodurch sie in der Lage ist, Dokumente auf Bedeutungsebene zu analysieren. Der tatsächliche Wortlaut spielt somit eine untergeordnete Rolle was bspw. die Definition der oben genannten Hotspots stark vereinfacht. **Alternative Softwarelösungen ohne KI können abweichende Formulierungen kaum oder nur schwer miteinander vergleichen und sind oft nicht in der Lage, Textdokumenten unabhängig von konkreten Formulierungen miteinander zu vergleichen.**

Ist es für einen Anwendungsfall ausreichend, diese zu identifizieren, können die Fundstellen inkl. der Referenzen im Dokument vermerkt und an den Anwender zurückgemeldet werden. Die Datenbasis, aufgrund derer die Treffer ermittelt werden, können vom Anwender beeinflusst werden. semantha® bietet zudem die Möglichkeit, Feedback zu einzelnen Ergebnissen (bspw. ein Nutzerhinweis "Diese Textpassage passt besser / am besten.") zu geben **(Element Lt - Knowledge Refinement).** Die oben geschilderte Anpassbarkeit des Sprachmodells ist eine weitere Eigenschaft von semantha®, die das Element Lt umsetzt.

Ein einfacher Anwendungsfall ist das Prüfen von Verträgen auf No-Go-Klauseln, wobei semantha® nicht nur die Hotspots identifiziert (die entsprechende Textpassage kann für den Anwender rot hervorgehoben werden), sondern auch spezifische Aufhebungsklauseln erkennen kann (die Passage wäre dann "nur" gelb mit einem Verweis auf die zugehörige Öffnungsklausel). Ein anderer Anwendungsfall ist ein 1:n-Dokumentvergleich, bei dem jeweils ein Dokumentpaar miteinander verglichen wird. So ist es mithilfe von semantha® bspw. einfach möglich, die eigenen Vertragsbedingungen mit denen der Wettbewerber zu vergleichen. semantha® ermittelt dann die Fundstellen in den Fremddokumenten, die inhaltlich überlappende Regelungen enthalten, wie das eigene Dokument. Der 1:n-Vergleich lässt sich zudem auf einen n:m-Vergleich verallgemeinern, wodurch es dem Anwender möglich wird, die hierbei implizit erstellten semantischen Kategorien **(Element Lc - Category Learning)** zu erkunden und für zukünftige Analysen zu persistieren.

Im nächsten Schritt können - entweder basierend auf den Fundstellen des vorangegangenen Schritts oder auf dem gesamten Dokument - spezifische Datenpunkte aus einem Dokument extrahiert werden **(Element Te - Text Extraction)**. Hier kommen sowohl klassische Verfahren, als auch Verfahren des Maschinellen Lernens zum Einsatz, um aus den Textbestandteilen typisierte Datenpunkte zu extrahieren. So liefert semantha® bspw. nicht "24.12.2019" als Zeichenkette zurück, sondern "2019-12-24" mit dem Datentyp "Datum", sodass auf den extrahierten Datenwerten gerechnet werden kann - zum Beispiel plus oder minus eine Woche. In diesem Schritt können auch tabellarische Informationen aus dem Dokument extrahiert werden, wodurch eine Nutzung der Daten in einem nachgelagerten Schritt erheblich vereinfacht wird. Immer wenn semantha® einen Datenpunkt aus einem

Dokument extrahiert, liefert sie zudem die Fundstelle (bspw. bei einem PDF-Dokument die Seitennummer sowie X/Y-Koordinaten) mit, sodass ein Anwender in einer grafischen Benutzeroberfläche direkt an die Fundstelle geführt werden kann. Dies wiederum erleichtert es, Anwendungsfälle mit einem human in the loop umzusetzen. Diese Nachvollziehbarkeit - und Belegbarkeit der Ergebnisse - zeigt sich immer wieder als wertvoller Baustein, da in vielen Fällen eine vollständig dunkle Verarbeitung nicht gewünscht und/oder menschliche Kontrolle vorgeschrieben ist (bspw. durch berufsständische Sorgfaltspflichten). Die Ergebnisse dieses Schrittes können zum Einen konsolidiert zurückgegeben werden (um z.B. eine tabellarische Übersicht über viele Dokumente zu erzeugen oder um so gewonnene Datenpunkte als Metadaten in ein Dokumentenmanagementsystem einzuspeisen). Zum Anderen dienen die Ergebnisse als Eingabe für den folgenden Verarbeitungsschritt.

Der letzte Verarbeitungsschritt, die semantische Interpretation, erzeugt aus den gewonnenen Daten neue Informationen wie bspw. Handlungsempfehlungen inkl. einer Begründung **(Element Sy - Synthetic Reasoning)**. Hierbei kommt nicht Maschinelles Lernen zum Einsatz, sondern der Anwendungsfall wird entsprechend modelliert, damit logische Schlussfolgerungen möglich werden. Hierbei greift semantha® auch auf externes Wissen zurück, das nicht explizit im Dokument steht, sondern sich über logische Schlüsse ermitteln und/oder verknüpfen lässt. Beispielsweise kann für einen juristischen Anwendungsfall die Regel hinterlegt werden, dass eine gewisse Klausel nicht unterzeichnet werden darf, wenn die Haftungssumme unbegrenzt oder einen Wert größer als T€ 500 aufweist und der Gerichtsstand außerhalb der Europäischen Union liegt. Offensichtlich einfach ist hierbei die Betragsprüfung, die aufgrund der Datenextraktion in der vorangegangen Stufe durchgeführt wurde. Umfangreicher ist die Herleitung, dass der extrahierte Gerichtsstand “New York” in New Jersey, USA liegt und diese nicht Teil der Europäischen Union ist. Stehen die Informationen für eine semantische Interpretation bereit, liefert semantha® nicht nur das Ergebnis, sondern auch die logische Kette, die zum Ergebnis geführt hat. So wird es einerseits dem Anwender ermöglicht, die Entscheidung der KI nachzuvollziehen, und andererseits können Entscheidungsketten programmatisch beeinflusst werden, im Gegensatz zu einem trainierten System.

Nicht jeder Anwendungsfall benötigt alle Verarbeitungsschritte, weswegen semantha® eine umfangreiche API zur Verfügung stellt, die die gewünschten Auswertungen ermöglicht.