Inhalte erschließen mit

LCIndex

Sie wollen wissen, welche Begriffe in einem Dokument die entscheidenden sind und um welche Hauptthemen es geht? Sie wollen eine optimale Verschlagwortung Ihrer Dokumente, um Inhalte später einfach wiederfinden zu können? Sie wollen bei Ihrer Suchfunktion weg von der Volltextsuche und hin zu einer semantischen Suche? Sie wollen in Ihrem Verlag für Sachbücher effizient Indexe erstellen?

LCIndex bietet Ihnen diese Funktionalitäten auf allerhöchstem Niveau. Sie können schnell und preiswert große Dokumentmengen inhaltlich erschließen und so die Basis für eine effektive Suche schaffen.

Automatische Indexierung

LCIndex basiert auf den linguistischen Technologien von **LCCore** und **LCTerm** und kombiniert diese mit statistischen Verfahren. Aufgrund der ermittelten morpho-syntaktischen Informationen über die Wörter Ihrer Texte kann **LCIndex** wichtige von unwichtigen Wörtern unterscheiden und garantieren, dass Textpassagen auch dann gefunden werden, wenn sie nicht nur 1:1-Entsprechungen der Suchbegriffe enthalten, sondern beispielsweise auch flektierte Formen und Schreibvarianten.

Linguistisch charakteristische Einheiten wie Komposita oder Nominalphrasen aus Adjektiv und Substantiv werden zunächst als potenzielle Schlagwörter ermittelt. Da linguistische Faktoren allein nicht ausreichen, um allgemeinsprachliche von fachgebietsspezifischen Wörtern zu unterscheiden, werden die ermittelten potenziellen Schlagwörter anschließend mithilfe ausgereifter statistischer Verfahren gewichtet. Allgemein anerkannt sind die Annahmen, dass Schlagwörter, die in einem Dokument häufig vorkommen, wichtiger sind als weniger häufige Schlagwörter (Termfrequenz) und dass Schlagwörter, die in vielen Dokumenten des Dokumentbestands vorkommen, weniger wichtig sind als die, die nur in wenigen Dokumenten vorkommen (inverse Dokumenthäufigkeit). Ein interessanter Faktor, der zusätzlich in die Gewichtung einfließt, ist die Häufigkeit semantischer Merkmale, die in einem Text vorkommen.

Berücksichtigung Ihres Thesaurus

Bei Durchführung der Indexierung können wir Ihren bereits vorhandenen Thesaurus berücksichtigen. Die Einbindung eines Thesaurus garantiert die Erkennung und ermöglicht eine hohe Gewichtung des definierten Fachwortschatzes. Darüber hinaus kann das im Thesaurus hinterlegte Wissen über Begriffsrelationen wie Ober- und Unterbegriff genutzt werden, um beispielsweise von einem Heckantrieb auf einen Fahrzeugantrieb zu schließen.

Linguistische Operationen und intelligente Matchingstrategien ermöglichen erweiterte Suchräume für Schlagwörter, die auf Basis eines Thesaurus vergeben werden, wie die nebenstehenden Beispiele illustrieren.

Features

- thesaurusbasierte Verschlagwortung
- freie Verschlagwortung
- Statistische Gewichtung
- Dokumentklassifikation
- Wortwolkenberechnung
- Dokumentähnlichkeit

Sprachen

- Deutsch
- Englisch

Gesucht, gefunden

Anlagenbetreiber
→ *Anforderungen, die das Gesetz an den **Anlagebetreiber** stellt ...*

Solarstromanlage
→ *Informationen zur Nachrüstung von **Solarstrom-Anlagen** ...*

Solarkraftwerk
→ *Ausbaustopp für **Solar-** und Wind**kraftwerke** ...*

Netzstabilität
→ *Bisher war die **Stabilität des Netzes** durch…*

Netzbetreiber
→ *Anzahl der Strom**netzbetreiber** in Deutschland …*

Emissionsreduktion
→ ***Emissionsminderung** für prioritäre Stoffe der …*

Wortwolken

Wortwolken sind Sammlungen semantisch ähnlicher Begriffe, die durch gemeinsames Vorkommen in großen Dokumentbeständen ermittelt werden. Ihr Nutzen ist sehr vielfältig. Den Anwender unterstützen Wortwolken bei der so genannten "assoziativen" Suche: Wenn die verwendeten Suchbegriffe nicht zum gewünschten Ergebnis führen, erlaubt die Wortwolke, mit semantisch ähnlichen Begriffen weiterzusuchen und auf diese Weise die Suchanfrage zu präzisieren.

Im Indexierungsprozess können Wortwolken bei der Auflösung von Mehrdeutigkeiten eingesetzt werden: Enthält ein zu indexierender Text beispielsweise ein mehrdeutiges Schlagwort wie *Virus*, kann man auf Basis der Ähnlichkeit des Texts mit den Wortwolken der verschiedenen Bedeutungsinstanzen entscheiden, ob es sich im vorliegenden Text eher um einen Organismus oder um einen Computervirus handelt. Auf ähnliche Weise können Bedeutungen unbekannter Abkürzungen identifiziert werden, deren zugehörige Langform typischerweise in der Wortwolke der Abkürzung besonders prominent enthalten ist (z. B. *WKA* für *Windkraftanlage*).

Klassifikation

Auf Basis seiner identifizierten Schlagwörter können wir einen Text einer oder mehreren Domänen zuordnen. Auf einer generellen Ebene wird z. B. entschieden, ob es sich um einen Text aus der Chemie oder dem Maschinenbau handelt. Feinere Klassen sind z. B. Halbleitertheorie oder Künstliche Intelligenz.

Bei der Ermittlung von Klassifikationssystemen stellen wir verschiedene Methoden zur Verfügung. Wir verwenden aber auch gern ein von Ihnen bereits gepflegtes Klassifikationssystem.

Dokumentähnlichkeit

Mit der Indexierung und der Klassifikation ist der Grundstein gelegt, um ähnliche Dokumente automatisch in Gruppen zusammenzufassen. Die eingesetzten Clusterverfahren gehen von einer einfachen Annahme aus: Je größer die Übereinstimmung der ihnen zugewiesenen Schlagwörter, desto ähnlicher sind sich Dokumente.

Das Wissen über Dokumentähnlichkeit kann unter anderem die Effizienz von Suchverfahren verbessern.

Weitere Informationen erwünscht?

Kontaktieren Sie uns mit Ihrem Anliegen zur inhaltlichen Erschließung Ihrer Texte. Wenn Sie spezielle Anforderungen an die automatische Inhaltserschließung haben, können wir Ihnen sicherlich weiterhelfen.

IAI Linguistic Content AG

Martin-Luther-Str. 14
66111 Saarbrücken
0681 38951-0

www.iailc.de
kontakt@iailc.de