IBM® Watson Content Analytics Crawler für Online Media

Ermöglicht Analyse und Volltextsuche in Online und Social Media Daten

Highlights

Suche und Analyse von:

- Daten aus RSS Feeds
- Social Media Daten (Twitter, BoardReader Forum Search Engine)
- Daten aus Internetsuchen (BING Web Search, Google Site Search)

Erfassen der relevanten Inhalte, nicht relevanter Inhalt kann herausgefiltert werden

Schnelles Crawlen von Änderungen

Reduntante Artikel können erkannt und ignoriert werden

IBM Watson Content Analytics bietet Funktionalitäten für eine skalierbare inhaltsbezogene Textanalyse und stellt vorgefertigte Integrationen zur Indizierung von Datenbanken, Collaborationsanwendungen, Dateisystemen und Webseiten zur Verfügung.

Der *IBM Watson Content Analytics Crawler für Online Media* erweitert die vorhandenen Such- und Analyseszenarien um die Suche in Inhalten, die entweder durch RSS Feeds oder durch Abfragen an externe Datenquellen wie BoardReader, BING oder Twitter bereitgestellt werden.

Dabei kann durch die Auswahl entsprechender Abfragebedingen erreicht werden das nur die relevanten Online und Social Media Einträge in die Analyse eingehen.

Funktionalitäten

Der *IBM Watson Content Analytics Crawler für Online Media* kann verwendet werden um Inhalte zu crawlen die über RSS Feeds zur Verfügung gestellt werden. Dabei werden die Einträge in den Feeds genutzt um den Inhalt des Artikels der Textanalyse von *IBM Watson Content Analytics* hinzuzufügen.

Darüber hinaus kann der *IBM Watson Content Analytics Crawler für Online Media* genutzt werden um Daten von Anbietern wie Twitter oder BoardReader Forum Search Engine zu indizieren. Durch die Auswhal der Abfragebedigungen kann hier sichergestellt werden kann das aus den Milliarden von Social Media Beiträgen nur die relevanten Beiträge indiziert und analysiert werden.

Gleiches gilt für die Suche in Internetseiten. Hierfür nutzt der *IBM Watson Content Analytics Crawler fürOnline Media* als Datenquelle eine Websuche bei BING um nur die Internetseiten zu analysieren die relevanten Inhalt enthalten.

Dabei folget der Crawler generell nicht den im Artikel enthaltenen weiteren Verweisen, dadurch wird sichergestellt das nur relevanter Inhalt dem Volltextindex hinzugefügt wird. Nicht relevante Bestandteile der Internetseiten selbst können zusätzlich über reguläre Ausdrücke oder andere Filteralgorithmen ausgefiltert werden

Technische Informationen

Die zu crawlenden Online und Social Media Quellen werden entweder in einer Konfigurationsdatei oder in einer Datenbank zur Verfügung gestellt, zusammen mit weiteren vorgegebenen Metadaten für die neuen Dokumente, zum Beispiel die Sprache.

Der Crawler bietet verschiedene Optionen um eine hohe Qualität der Dokumente im *IBM Watson Content Analytics* Index zu erreichen:

- Verweise innerhalb der gecrawlten Dokumente werden nicht verfolgt, dadurch wird nur der relevante Inhalt erfasst.
- Die Dokumente werden mit zusätzlichen Metadaten angereichert (aus der Quelle selbst oder feste Metadaten aus der Quellenkonfiguration).
- Durch die Verwendung von regulären Ausdrücken oder Filteralgorithmen kann die Qualität der Dokumente weiter erhöht werden, beispielsweise können Kopf- und Fußzeilen ausgefiltert werden.

Da der Crawler sowohl URLs und Inhalt der Dokumente ähnlich dem Standard Web Crawler speichert können die Benutzer der WCA Such- bzw. Analyseanwendung diese wie gewohnt anzeigen.

Unterstützte Versionen

- *IBM Watson Content Analytics* 3.0 und 3.5
- *IBM Watson Explorer Advanced Edition* 10.0 (Analytical Components).

Unterstützte Formate

- RSS 0.9x, RSS 1.0 / RDF, RSS 2.0, Atom 0.3, Atom 1.0
- BoardReader Forum Search API
- Twitter Search API
- BING Web Search API
- Google Site Search API

Für Abklärung des Supports von abweichenden Versionen kontaktieren sie bitte das Germany Asset Support Center des ECM Software Services Team unter der E-Mail: **gerasc@de.ibm.com**

Service Offering

- Runtime Version je *IBM Watson Content Analytics* System
- Unterstützung bei Installation und Konfiguration

IBM Deutschland GmbH
IBM-Allee 1
71139 Ehningen
ibm.com/de

Die IBM Homepage erreichen Sie unter:
ibm.com