Michael Göpfert

# BEYOND CRAWLING: CRAWL-VISUALISIERUNG UND URL INSPECTION API

## (TEIL 3)

In den vergangenen beiden Ausgaben wurde viel auf die Möglichkeiten der Datenverarbeitung mit KNIME eingegangen. In Teil 3 der Serie Beyond Crawling soll es darum gehen, Daten zu visualisieren, Zusammenhänge zu erkennen und Indexierungsprobleme mit Googles neuer URL Inspection API zu entdecken.

Am 31. Januar 2022 hat Google die Search-Console-URL Inspection API freigeschaltet. Diese Schnittstelle ermöglicht es Webseitenbetreibern, automatisiert Informationen aus dem URL-Prüfungs-Tool der Search Console (GSC) abzufragen. So lassen sich wichtige Informationen wie der Indexierungsstatus einer URL für bis zu 2.000 URLs pro Tag (Limitierung durch Google) abfragen.

Viele SEO-Tool-Anbieter haben schnell reagiert und Möglichkeiten, die URL Inspection API abzufragen, in ihre Software eingebaut. So auch der beliebte SEO-Crawler Screaming Frog. Hier kann über den Menüpunkt „Configuration“ › „API Access“ › „Google Search Console“ auf die GSC API zugegriffen werden. Programmierkenntnisse sind nicht nötig, es reicht aus, seinen Google-Account, in dem die Search Console Properties liegen, mit Screaming Frog zu verbinden. Anschließend kann über den Reiter „URL Inspection“ auf die URL Inspection API zugegriffen werden (Abb. 1).

So lassen sich nun die Crawldaten von Screaming Frog mit Performance-Daten wie Klicks und dem Indexierungsstatus anreichern. Diese Daten wiederum können in einem Datenanalyse-Tool wie KNIME weiterverarbeitet und visualisiert werden.

### Crawl mit Screaming Frog

Um möglichst viele Erkenntnisse zu gewinnen, sollte zunächst ein vollständiger Crawl der Website durchgeführt werden. Vorab sollte Screaming Frog wie oben beschrieben der Zugriff auf die Search Console gewährt werden, um den Crawl mit Trafficdaten und Informationen aus der URL Inspection API anzureichern.

Nach Abschluss des Crawls werden die Daten im Reiter „Internal“ aus Screaming Frog als CSV-Datei exportiert.

Zunächst muss ein neuer Workflow in KNIME erstellt werden, alternativ kann der fertige Workflow auch unter *https://kni.me/w/IxEQiCtnArGB61wG* heruntergeladen und in KNIME geöffnet werden.

DER AUTOR

**Michael Göpfert** arbeitet gerne mit Rohdaten, um diese in maßgeschneiderten Analysen für seine Kunden aufzubereiten.

Foto: cybernaut / gettyimages.de

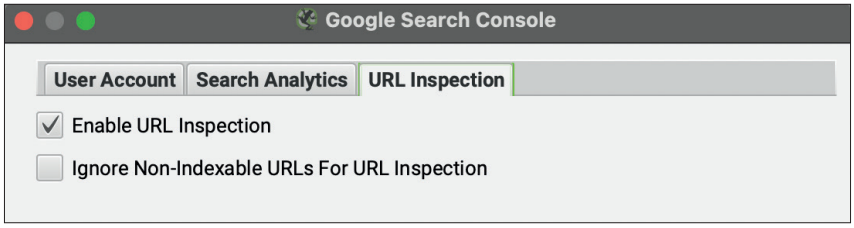



Abb. 1: Zugriff auf die URL Inspection API mit Screaming Frog

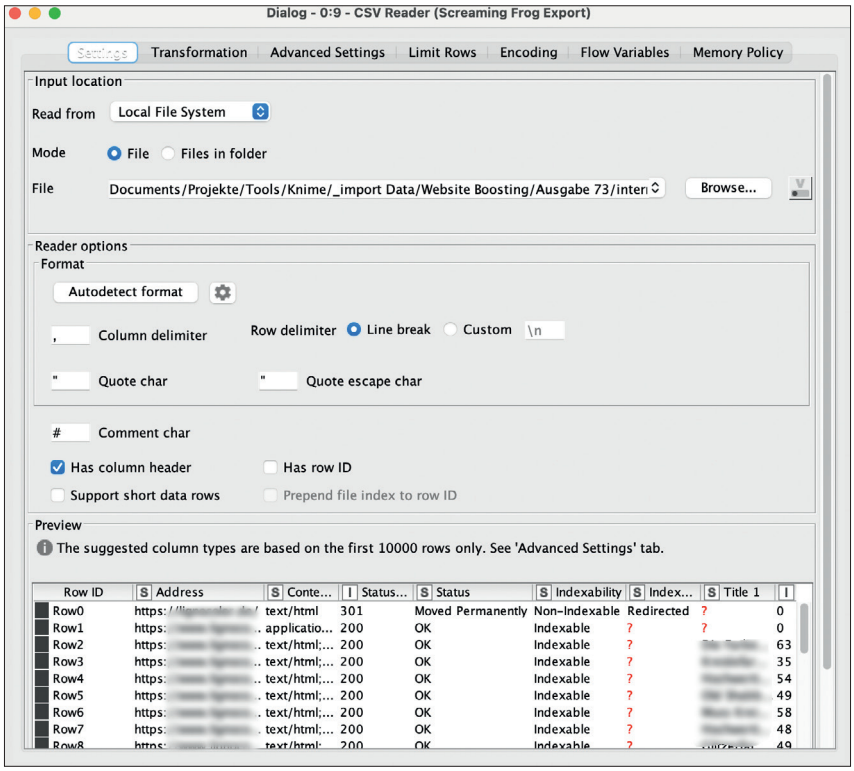



Abb. 2: Zugriff auf die URL Inspection API mit Screaming Frog

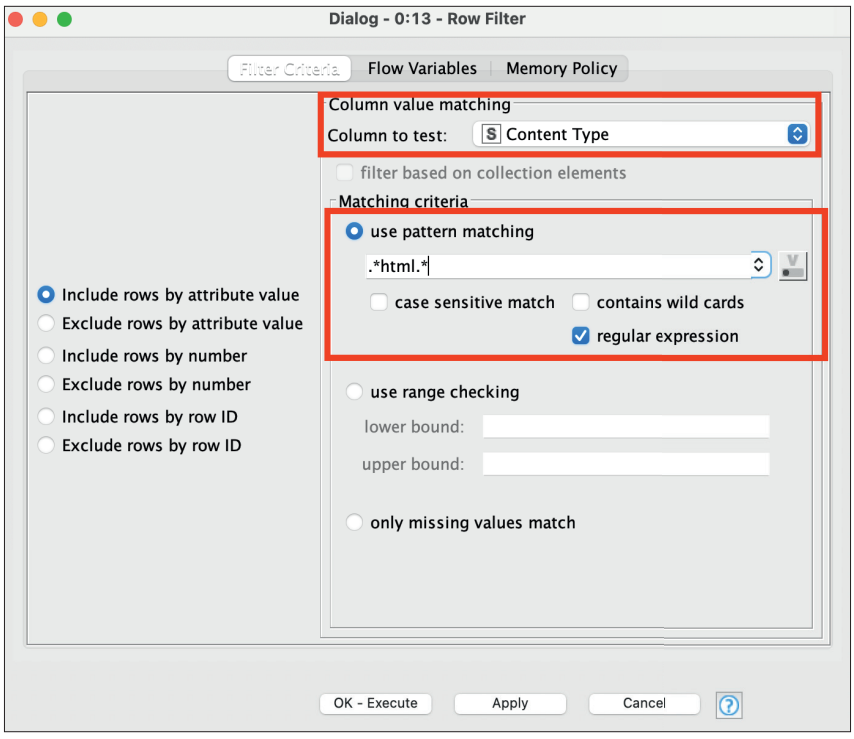
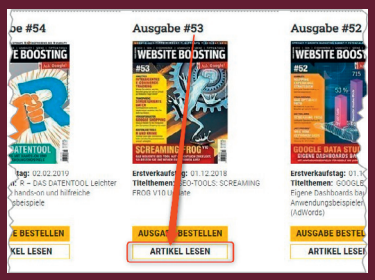



Abb. 3: Auf HTML-Dokumente filtern

**TIPP**

Wie Sie an das kostenlose Tool KNIME kommen und wie es prinzipiell funktioniert, finden Sie in der Ausgabe 53 oder online frei als HTML oder PDF unter *http://einfach.st/knime53*.




## Auswertung des Crawls mit KNIME

Für den Import der Daten wird die Node CSV Reader aus dem Node Repository in den Workflow gezogen (Drag-and-drop). Mit einem Doppelklick auf die Node lässt diese sich konfigurieren.

Hier reicht es, den Pfad zum CSV-Export aus Screaming Frog anzugeben. In Bereich „Preview“ lässt sich sehen, wie die Daten importiert werden (Abb. 2). Treten hier Fehler auf, lässt sich das Problem in den meisten Fällen durch Anklicken der Checkbox „Support short data rows“ beheben.

Nach dem Abschließen der Konfiguration kann die Node via Rechtsklick „Execute“ ausgeführt werden. Springt die „Ampel“ der Node auf Grün, wurden die Daten erfolgreich importiert. Anzeigen lassen sich die Daten per Rechtsklick und Auswahl des Menüpunktes „File Table“.

## Daten filtern und fehlende Werte ersetzen

Screaming Frog fragt den Indexierungsstatus bei Google nur für HTML-Dokumente ab, deshalb ist es sinnvoll, den Crawl auf HTML-Dateien zu filtern. Dazu kann die Node „Row Filter“ aus dem Node Repository in den Workflow gezogen werden und mit dem CSV Reader an den schwarzen Dreiecken per Drag-and-drop verbunden werden. Nun können die Daten vom CSV Reader in den Row Filter fließen.

Um auf HTML-Dokumente zu filtern, kann mit einem regulären Ausdruck gearbeitet werden. Zunächst muss jedoch die Spalte, die gefiltert werden soll, definiert werden. Dazu wird bei „Column to test“ die Spalte „Content Type“ ausgewählt. Anschließend wird im Bereich „Matching Criteria“ der reguläre Ausdruck „.*html.*“ eingetragen. Zusätzlich muss die Checkbox „regular expression“ ausgewählt werden (Abb. 3). Nach Ausführen der Node sind nur noch HTML-Dokumente im Crawl enthalten.

Auch sollen in der Auswertung nur Seiten betrachtet werden, die auch indexierbar sind. Deshalb kann mit einer zweiten Row-Filter-Node die Spalte „Indexability“ auf den Wert „Indexable“ gefiltert werden.

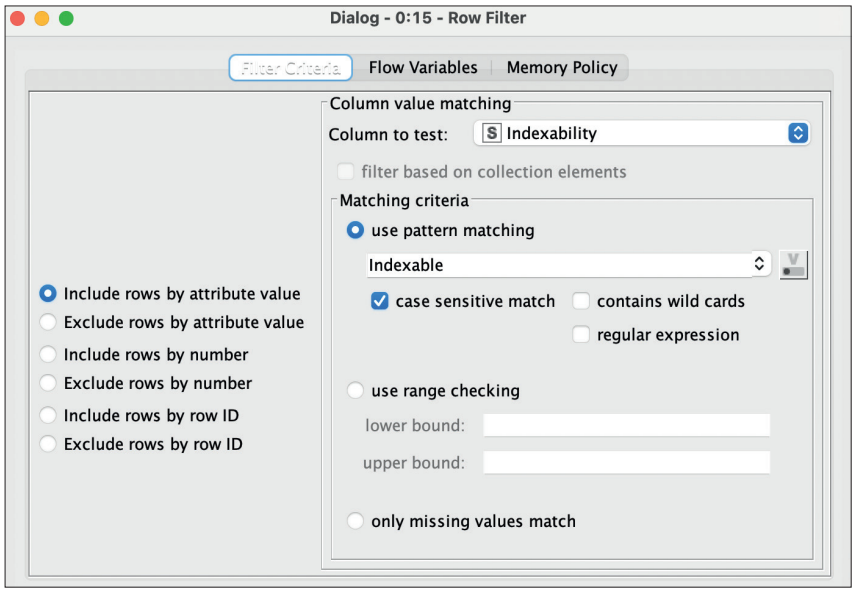



Abb. 4: Auf indexierbare URLs filtern

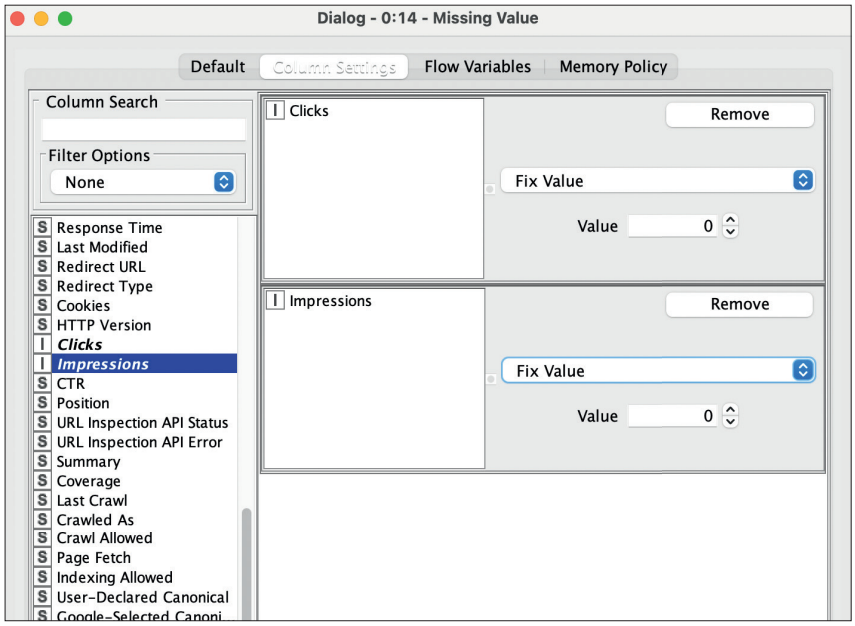



Abb. 5: Fehlende Werte in KNIME ersetzen

## Daten bereinigen: fehlende Werte ersetzen

Bevor die Daten ausgewertet werden können, müssen sie noch etwas bereinigt werden. Das gilt besonders für URLs, welche die Search Console nicht kennt und für die deshalb keine Traffic-Daten zur Verfügung stehen. Sind keine Daten vorhanden, zeigt KNIME ein rotes Fragezeichen in der betroffenen Zelle an (siehe Abb. 2). In den meisten Fällen ist das kein Problem, bei Traffic-Daten empfiehlt es sich jedoch, diese sogenannten „Missing Values“ mit 0 zu ersetzen.

Dazu kann die Node „Missing Values“ aus dem Node Repository in den Workflow gezogen und mit dem Row Filter per Drag-and-drop verbunden werden.

Für die Konfiguration der Node gibt es verschiedene Möglichkeiten. So können fehlende Werte in Spalten, die Text (Datentyp = String), und Spalten, die Ganzzahlen beinhalten (Datentyp = Integer), automatisch mit einem beliebigen Wert ersetzt werden.

Im Reiter „Column Settings“ lassen sich diese Einstellungen auch einzeln pro Spalte vornehmen. Um die Spalte mit den Klicks aus der Search Console anzupassen, können diese per Doppelklick ausgewählt werden und erscheinen anschließend im rechten Feld. Über das Dropdown kann nun für jede einzelne Spalte definiert werden, was mit den fehlenden Werten passieren soll. Hier gibt es zahlreiche Möglichkeiten. Mit der Option „Fix Value“ lassen sich die fehlenden Daten durch eine 0 ersetzen (Abb. 5). Dasselbe kann mit der Spalte Impressions durchgeführt werden.

## Insights aus Googles neuer API gewinnen

Um erste Insights aus Googles neuer API zu bekommen, lassen sich einzelnen Werten Farben in der Tabelle zuweisen, um später in einer Visualisierung einen schnellen Überblick zu bekommen. Dazu eignet sich die Node „Color Manager“. Auch hier muss zunächst die Spalte, der eine Farbe zugewiesen werden soll, ausgewählt werden. Um mit dem Indexierungsstatus zu arbeiten, bietet sich die Spalte „Summary“ an. Sie gibt in Form von drei Werten Rückschlüsse über den Indexierungsstatus:

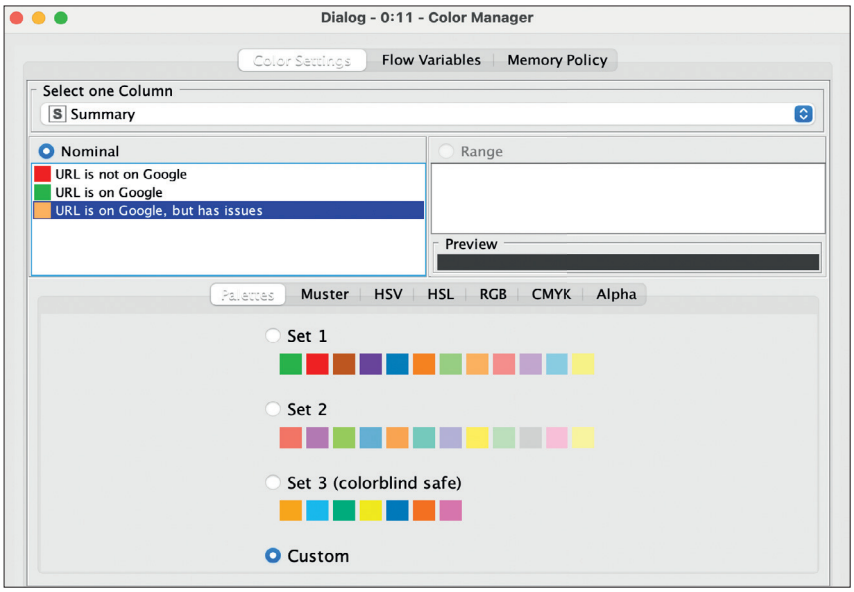



Abb. 6: Farben für bestimmte Werte vergeben

**HINWEIS**

Das Ersetzen fehlender Werte für die CTR und die Position ist etwas komplexer, da hier die Daten von Screaming Frog nicht als numerischer Datentyp, sondern als String exportiert werden. Das kann mit der Node „String to number" behoben werden. Die durchschnittliche Position sollte allerdings nicht mit 0 ersetzt werden, da sonst Seiten ohne Rankings in der Sortierung über den Seiten mit Rankings geführt werden.

- URL is on Google: Es ist alles in Ordnung und die URL ist indexiert.
- URL is on Google, but has issues: Die URL ist indexiert, es gibt aber Probleme. Diese können ein Fehlen der URL in der Sitemap oder Probleme mit der Mobilfreundlichkeit der Seite sein.
- URL is not on Google: Die Seite wurde noch nicht von Google indexiert. Beispielsweise, weil Google die URL als ein Duplikat betrachtet.

Für die Visualisierung mit Farben bietet sich ein Ampelsystem an: rot = URL is not on Google, gelb = URL is on Google, but has issues und grün = URL is on Google (Abb. 6).

Dazu können die Farben im Color Manager durch Auswählen des Werts und der Farbe entsprechend angepasst werden.

## Visualisierungen in KNIME

KNIME bietet zahlreiche Möglichkeiten, Daten zu visualisieren. Die Bordmittel dazu sind allerdings begrenzt. Glücklicherweise gibt es zahlreiche (kostenlose) Erweiterungen, unter anderem Visualisierungen von Plotly. Diese können im Menü „File" › „Install KNIME Extensions" installiert werden. Dazu einfach im Textfeld nach Plotly suchen, auswählen und die Installation starten (Abb. 7).

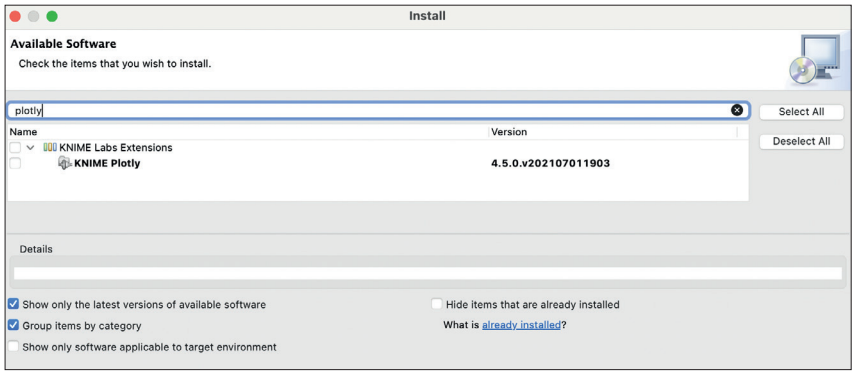



Abb. 7: Plotly installieren

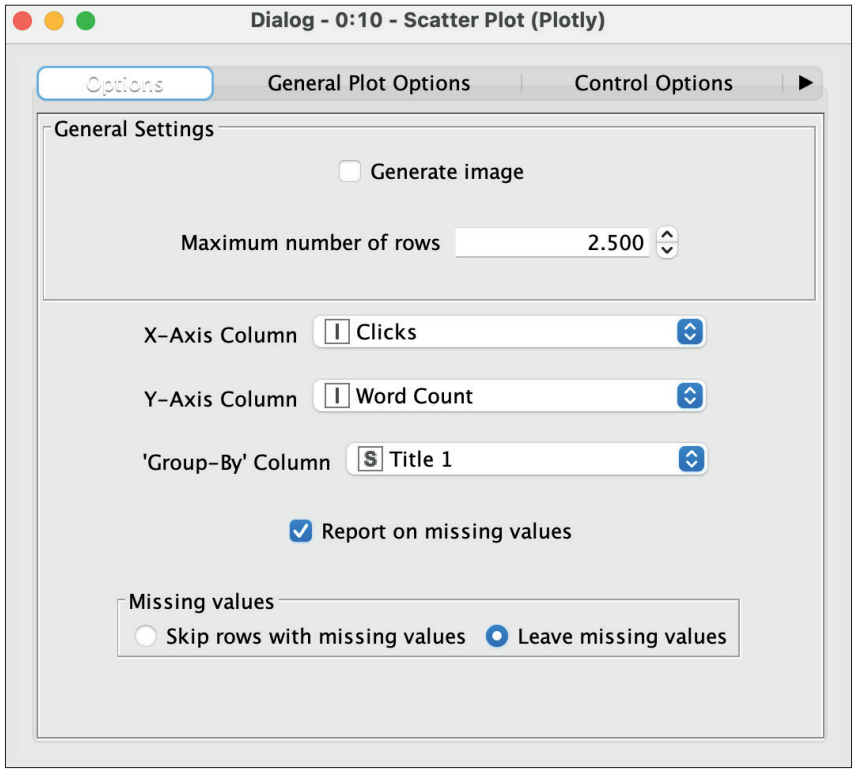



Abb. 8: Einrichten des Scatter Plot

Anschließend kann im Node Repository nach „Plotly“ gesucht werden und es stehen zahlreiche Nodes, mit denen Diagramme erzeugt werden können, zur Verfügung.

Um mögliche Ausreißer oder Zusammenhänge schnell zu erkennen, eignet sich ein Streudiagramm (Scatter Plot). So lässt sich beispielsweise einfach prüfen, ob Inhalte mit viel Text tatsächlich mehr Klicks erzeugen oder nicht. Zunächst muss jedoch die Node „Scatter Plot (Plotly)“ auf die Arbeitsfläche gezogen und mit dem Color Manager verbunden werden. Anschließend können die Einstellungen für das Streudiagramm vorgenommen werden. Je nach Größe des Crawls muss der Wert unter „Maximum number of Rows“ erhöht werden, um alle URLs darzustellen.

Als x-Achse können nun die Klicks und als y-Achse die Anzahl der Wörter (Spalte „Word Count“) aus dem Crawl gewählt werden. Im Bereich Group-By sollte eine Spalte gewählt werden, die möglichst genau erkennen lässt, um welche Seite es sich handelt (z. B. der Title oder die H1). Leider werden nur die ersten paar Zeichen dieses Werts im Diagramm angezeigt, sodass die Ableitung, um welche Seite es sich handelt, etwas mühselig werden kann. Allerdings wird beim Überfahren eines Punktes mit der Maus die Nummer der Zeile angegeben, in der sich die betreffende URL in der Tabelle befindet. Nach der Konfiguration kann die Node über „Execute and open Views“ ausgeführt werden.

Nun gibt das Streudiagramm einen schnellen Überblick darüber, ob es viele Probleme bei der Indexierung (grüne Punkte vs. gelbe oder rote) gibt und wie sich die Klicks auf die einzelnen URLs verteilen. Je nachdem, wie sich die Punkte auf der y-Achse verteilen, können mögliche (!) Zusammenhänge mit der Textlänge abgeleitet werden (Abb. 10).

Über das Burger-Menü oben rechts im Diagramm können die Spalten, die

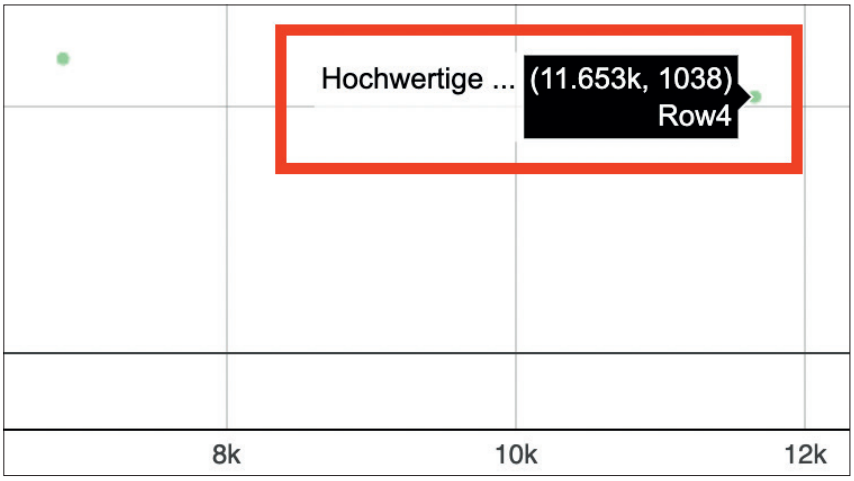



Abb. 9: Bei Mouseover auf einen Datenpunkt wird die Zeile, in der sich dieser in der Tabelle befindet, angezeigt

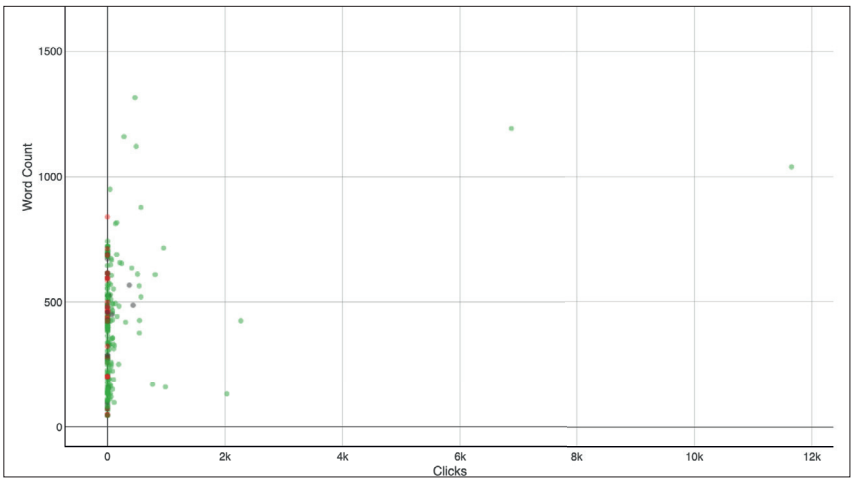



Abb. 10: Anzahl Wörter vs. Klicks im Streudiagramm

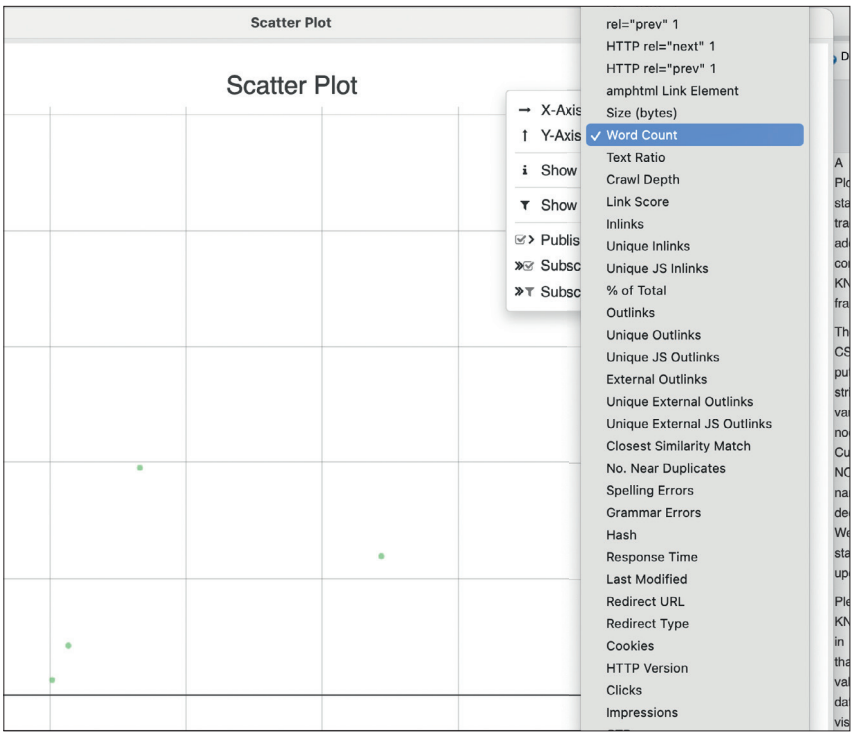



Abb. 11: Einfache Auswahl der Metriken, die für die jeweiligen Achsen verwendet werden sollen

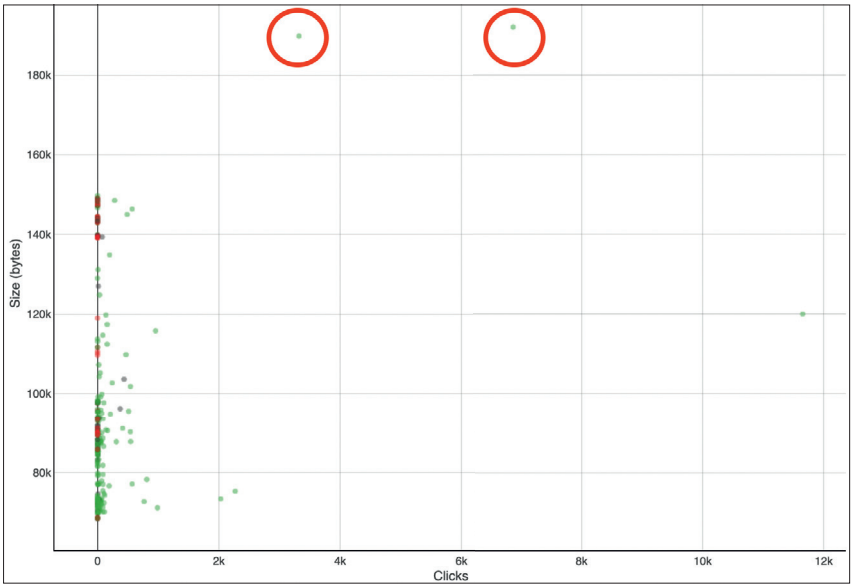



Abb. 12: Ausreißer bei der Dateigröße

die Werte für x- und y-Achse liefern, schnell und einfach gewechselt werden (Abb. 11).

So lassen sich schnell Erkenntnisse gewinnen wie:

» Gibt es viele URLs, die keine Klicks bringen?
» Gibt es bei vielen Seiten Probleme bei der Indexierung?
» Gibt es Seiten mit wenig eingehenden Links, die (zu) wenige Klicks bekommen?
» Gibt es Ausreißer bei der Dokumentgröße (Spalte „Size (bytes)“; Abb. 12)?
» U. v. m.

## Fazit

Streudiagramme in KNIME sind eine einfache und aufschlussreiche grafische Darstellung von Crawl- und Performance-Daten. Durch die Anzeige von zwei (x- und y-Achse) oder drei (Farbe) Variablen lassen sich leicht Anomalien oder Zusammenhänge erkennen, die auf positive oder negative Eigenschaften hindeuten. Da Screaming Frog neben der Search Console API noch die Anbindung zu zahlreichen anderen Schnittstellen wie PageSpeed Insights unterstützt, gibt es hier noch deutlich mehr zu entdecken. Probieren Sie es aus. ¶


# WEBSITE BOOSTING #074 erscheint am 14.6.2022

**Herausgeber & Chefredakteur (verantwortlich):**
Mario Fischer
E-Mail: redaktion@websiteboosting.com

**Autoren dieser Ausgabe:**
Marie Bachmayr, Dr. Martin Bahr, Alexander Beck, Britta Behrens, Dr. Beatrice Eiring, Darius Erdt, Michael Göpfert , Stefan Gottwald, Marco Janck, Thomas Kaiser, Markus Kellermann, Leonard Metzner, Stefan Vorwerk, Sarah Weitnauer

**Anzeigenleitung:**
Markus Lutz
E-Mail: anzeigenleitung@websiteboosting.com

**Art Direction, Layout/Produktion:**
Kai Neugebauer

**Lektorat:**
Bärbel Philipp, textperlen.de,
Ursula Wenke, www.lektorat-wenke.de

**Fotos & Illustrationen:**
Website Boosting / GettyImages

**Druck:**
Vogel Druck und Medienservice GmbH
Leibnizstr. 5, 97204 Höchberg

**Vertrieb:**
PressUp GmbH
Postfach 70 13 11
22013 Hamburg
E-Mail: websiteboosting@pressup.de

**Abonnement:**
Website Boosting Aboservice
PressUp GmbH
Postfach 70 13 11
22013 Hamburg
Tel. 040 / 38 6666 – 342
Fax: 040 / 38 6666 - 299
E-Mail: websiteboosting@pressup.de

Erscheinungsweise: 6 x jährlich
Bezugspreis: Einzelheft: 11,80€
Bezugspreis Inland jährlich 62,00€ inkl. Versand
Bezugspreis Ausland jährlich 70,80€
inkl. Versand

Studenten im Inland erhalten gegen Vorlage einer Immatrikulationsbescheinigung einen Preisvorteil – Details finden Sie auf der Website.

**Verlagsleitung:**
Michael Müßig
Tel: +49 931 / 26 038 04,
verlag@websiteboosting.com

**Anschrift des Verlages**
Hotspot Verlag GmbH
Obere Landwehr 4a, 97204 Höchberg
Tel: + 49 931 / 26 038 04
Fax: +49 931 / 26 038 05
E-Mail: verlag@hotspotverlag.de
www.hotspotverlag.de

**Geschäftsführung:**
Kai Neugebauer

Die Inhaber- und Beteiligungsverhältnisse lauten wie folgt: Gesellschafter zu 100% ist die Webvalue Holding GmbH

ISSN: 2191-6241