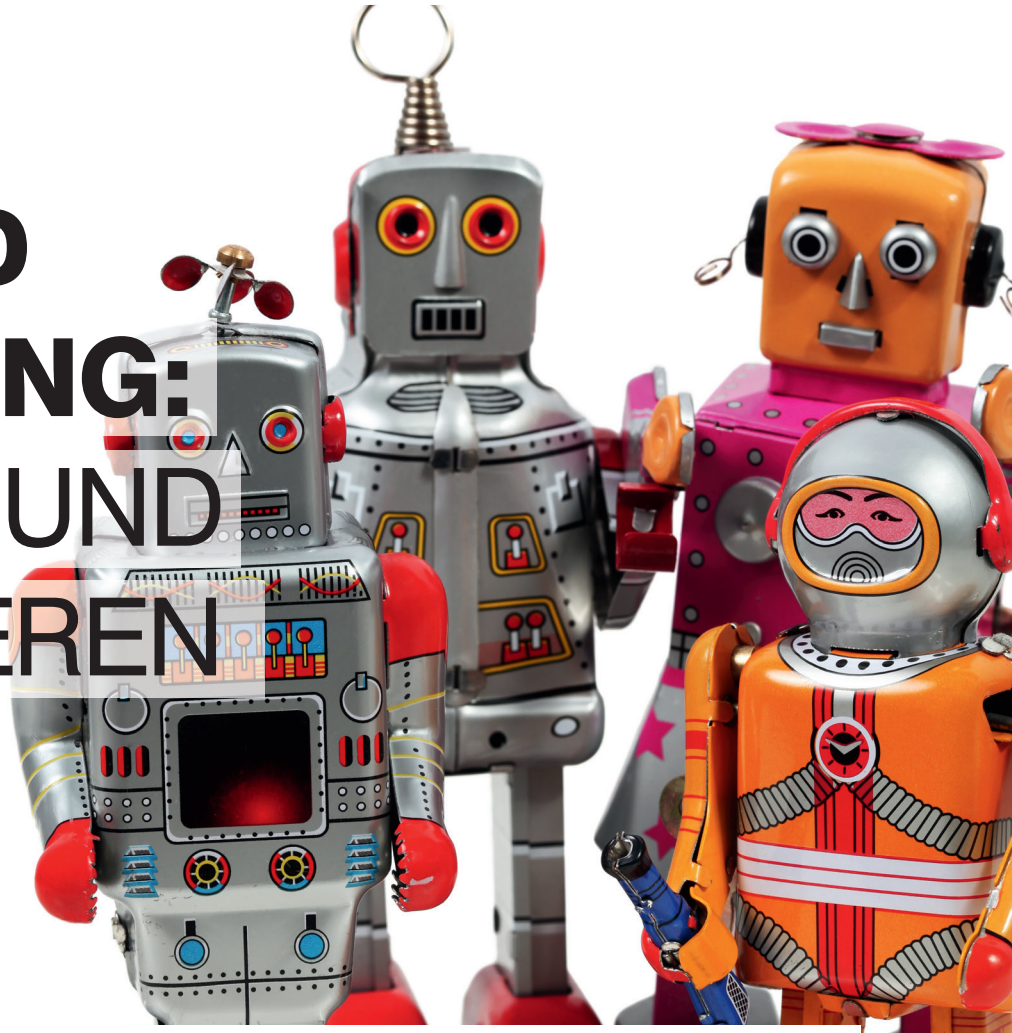
Michael Göpfert

BEYOND CRAWLING: FILTERN UND GRUPPIEREN

(TEIL 2)



Nachdem in der vergangenen Ausgabe ein Screaming-Frog-Crawl um eine zusätzliche Spalte, die den Namen des jeweiligen Seiten-Templates beinhaltet, ergänzt wurde, geht es nun darum, den Crawl mit dem Datenanalyse-Tool KNIME zu filtern, um erste Erkenntnisse über den Zustand der Seite zu gewinnen.
Datenexperte Michael Göpfert zeigt, wie man das in Eigenregie mit einfachen Mitteln unter Zuhilfenahme des kostenlosen Tools KNIME bewerkstelligen kann. Sie kennen KNIME noch nicht? Auch hier gilt wie immer: Mit dem Arbeiten an einer Problemlösung kommt man relativ schnell und einfach neuen (und wie hier mächtigen) Tools ein Stück näher und kann deren Potenzial für andere Problemlösungen recht gut einschätzen.

Seit der Version 16 bietet Screaming Frog sehr gute Bordmittel, um einen Crawl zu filtern. Warum also trotzdem KNIME nutzen? KNIME bietet gegenüber Filterungen in Excel oder Screaming Frog ein paar spannende Vorteile:

- **Reproduzierbarkeit:** Jeder Workflow muss nur einmal erstellt werden und kann immer wieder genutzt werden.
- **Übersichtlichkeit:** Jedem Workflow können durch Annotationen Beschreibungen hinzugefügt werden, sodass stets deutlich ist, welche Funktionen der Workflow durchführt.
- **Erweiterung und Manipulation:** Die Daten können nicht nur gefiltert, sondern auch verändert oder ergänzt werden.
- **Visualisierung:** Jeder Workflow kann an jeder beliebigen Stelle eine Visualisierung erzeugen.
- **Export:** Ebenso können die Ergebnisse jederzeit exportiert werden.

HINWEIS

Falls noch kein eigener KNIME-Workflow existiert, kann der hier gezeigte Workflow als Vorlage heruntergeladen und in KNIME geöffnet werden: *https://kni.me/w/x17G6dX8IOy5v-Fe*. Anschließend muss lediglich ein fertiger Screaming-Frog-Crawl exportiert und in der CSV Reader Node in KNIME geöffnet werden, um den Workflow zu nutzen.

DER AUTOR

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

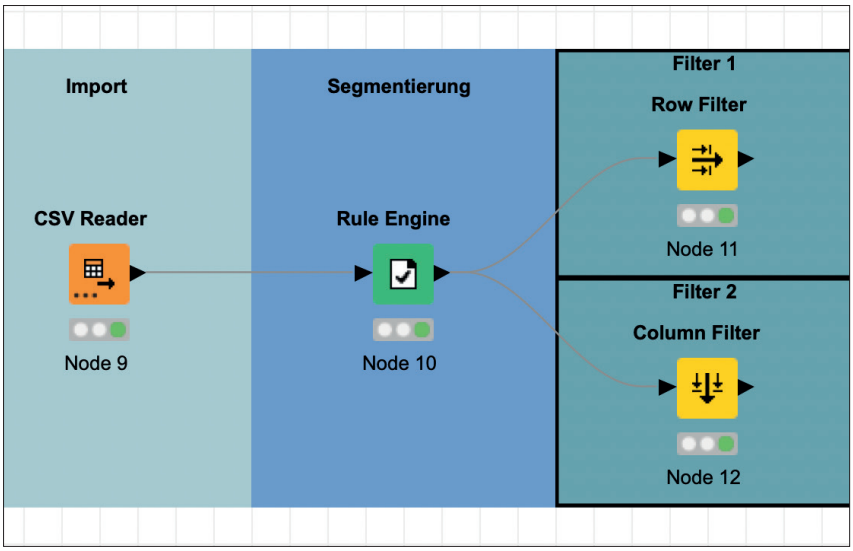
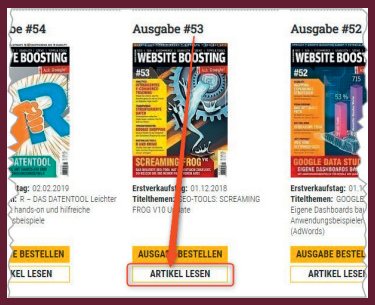


Abb. 1: Aufteilen des Workflows in zwei Teile, die nach verschiedenen Kriterien gefiltert werden

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*.



Auch lässt sich so das Problem umgehen, dass umfangreiche Crawls, bei denen viele Daten erhoben werden, schnell unübersichtlich werden: Durch die Filter in KNIME lässt sich ein Crawl in unterschiedliche Teilbereiche (bestimmte Spalten oder Zeilen) aufteilen, die getrennt voneinander ausgewertet werden.

Die in den vorderen Nodes (beispielsweise bei der Segmentierung) getroffenen Einstellungen vererben sich dabei auf alle folgenden Nodes. Deshalb ist es sinnvoll, Einstellungen, die die komplette Analyse betreffen, ganz am Anfang des Workflows zu platzieren. Die Segmentierung (siehe Ausgabe 71) beispielsweise wurde direkt nach dem Import platziert, um in allen weiteren Nodes darauf zugreifen zu können. Aber keine Sorge: Zusätzliche Nodes lassen sich ganz einfach per Drag & Drop zwischen zwei bestehende Nodes einfügen. So kann der Workflow zu jeder Zeit durch Hinzufügen oder Entfernen von Nodes angepasst werden.

Tipp: Die Nodes sind standardmäßig durchnummeriert. Durch einen Doppelklick auf die Nummer lässt sich dort aber auch ein kurzer Text eintragen, der beschreibt, welche Funktion die Node erfüllt.

Mit dem Row Filter Zeileninhalte filtern

Die Node „Row Filter“ ist eine der wichtigsten, wenn es um das Arbeiten mit großen Tabellen geht. Sie kann gut mit der Filter-Funktion in Excel verglichen werden. Der Row Filter in KNIME bietet dabei aber deutlich mehr Möglichkeiten. So ist es möglich, entweder nach einem Attribut (= dem Inhalt in einer bestimmten Spalte), nach Zeilennummern oder nach bestimmten IDs der Zeile zu filtern.

Zusätzlich kann bestimmt werden, ob die Zeilen nach dem Filtern in der Tabelle bestehen bleiben oder ausgeschlossen werden sollen. Der nützlichste Bereich der Node ist zweifelsohne das Filtern nach Attributen in einer beliebigen Spalte.

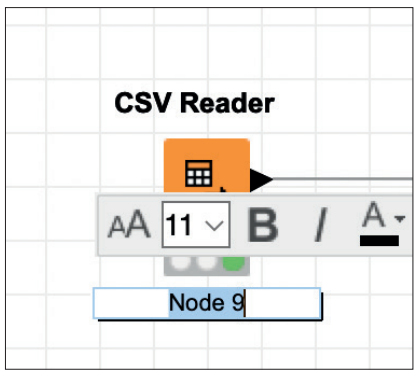


Abb. 2: Jeder Node kann eine kurze Beschreibung hinzugefügt werden

Um die Node zu konfigurieren, reicht ein Doppelklick, um das Konfigurationsfenster zu öffnen. Nun lassen sich verschiedene Einstellungen vornehmen. Im Bereich der Attributfilterung muss zunächst über das Drop-down die Spalte gewählt werden, die gefiltert werden soll. Anschließend können die sogenannten „Matching Criteria“ bestimmt werden, die erfüllt werden müssen, damit der Filter greift. Das funktioniert recht ähnlich wie in Excel. Der per Default ausgewählte Punkt „use pattern matching“ funktioniert wie der „Ist gleich“-Filter in Excel. In KNIME haben wir zusätzlich die Möglichkeit, mit Wildcards (entspricht „Enthält“-Filter in Excel) oder regulären Ausdrücken zu arbeiten.

Einfaches Beispiel: Kurz nach dem Jahreswechsel ist es sinnvoll, einen Blick in die Descriptions zu werfen, um zu sehen, ob dort nicht noch irgendwo die alte Jahreszahl auftaucht. Dazu lässt sich die Spalte „Meta Description 1“ aus dem Screaming-Frog-Crawl auf das Attribut „enthält die Zahl 2021“ filtern. Das lässt sich über eine Wildcard in Form von *2021* lösen (Abb. 3). So reicht es, wenn die Zahl 2021 an irgendeiner Stelle in der Description vorkommt, damit der Filter greift. Ebenfalls nützlich ist die Option „use

range checking“: Sie ermöglicht es, Spalten mit numerischen Werten zu filtern. Dazu stehen die beiden Felder „lower bound“ (der niedrigste Wert) und „upper bound“ (der höchste Wert) zur Verfügung. Es muss allerdings nur eines der beiden Felder befüllt werden, um den Filter zu nutzen.

Auch hier gibt es Parallelen zu Excel, und zwar zum Filter „größer als“ oder „kleiner als“.

Um den Crawl auf URLs zu filtern, die eine Weiterleitung sind, kann der lower bound auf 301 und der upper bound auf 308 für die Spalte „Status Code“ gesetzt werden. Weiterleitungen erzeugen immer einen Status-Code 301, 302, 303, 307 oder 308, deshalb kann hier 301 als kleinster und 308 als größter Wert gesetzt werden, um alle möglichen Redirects abzudecken (Abb. 4).

Umgekehrt ließen sich so auch alle Seiten ausschließen, die einen Status-Code ausgeben, der größer als 200 ist. Dazu müsste lediglich „Exclude rows by attribute“ gewählt und als „lower bound“ 201 eingetragen werden (Abb. 5). Damit wären alle URLs mit einem Status-Code (beispielsweise Fehlerseiten mit 40x, Weiterleitungen mit 30x oder Serverfehler mit 50x), der eine Indexierung verhindert, herausgefiltert.

Spalten Filtern

Um den Crawl übersichtlich und die Auswertung performant zu halten, ist es sinnvoll, nur die Spalten zu verwenden, die für die aktuelle Fragestellung wichtig sind. Zu Beginn des Workflows (in Ausgabe 71) wurde dem Crawl die Spalte „Seitentyp“ hinzugefügt. Mit der Node „Column Filter“ kann nun das Gegenteil gemacht werden: Durch das Filtern auf die aktuell benötigten Spalten bleibt die Auswertung übersichtlich.

Die Anwendung dieser Node ist denkbar einfach: Alle Spalten, die

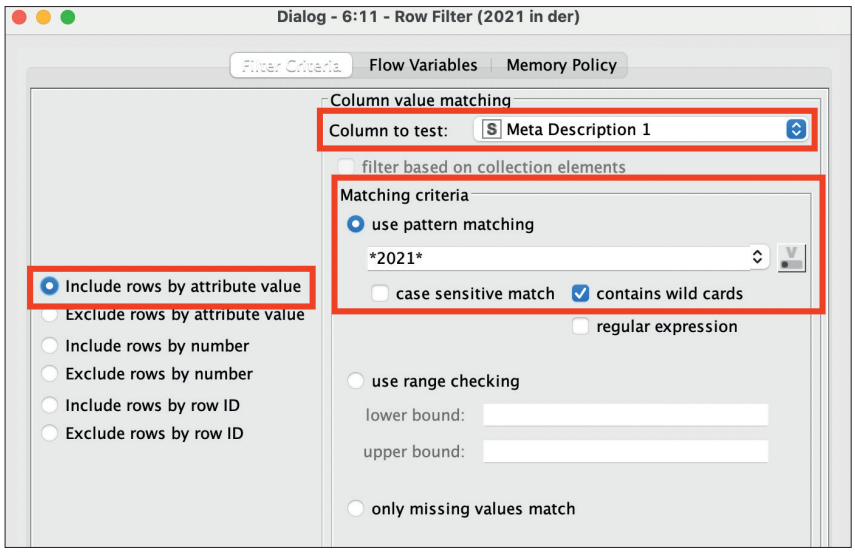


Abb. 3: Filtern mit Wildcards

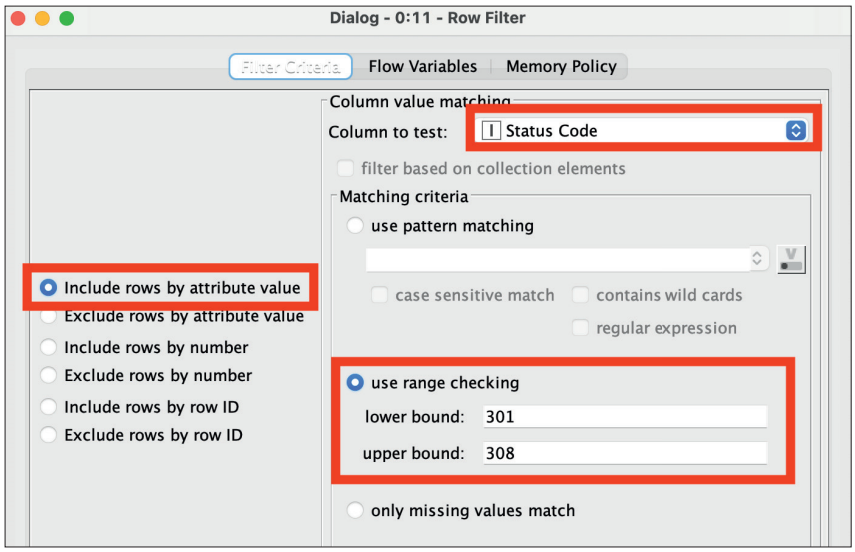


Abb. 4: Filtern nach Status-Codes

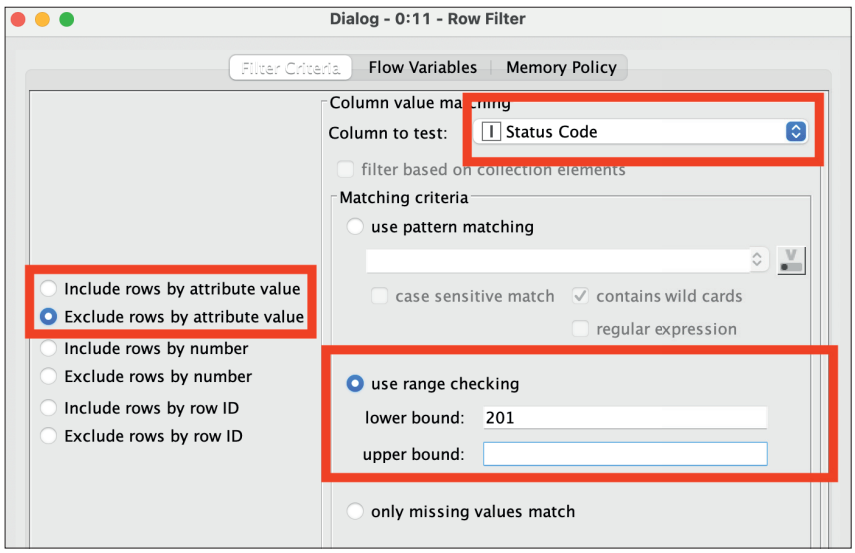


Abb. 5: Ausschließen von Seiten mit einem Status-Code größer als 200

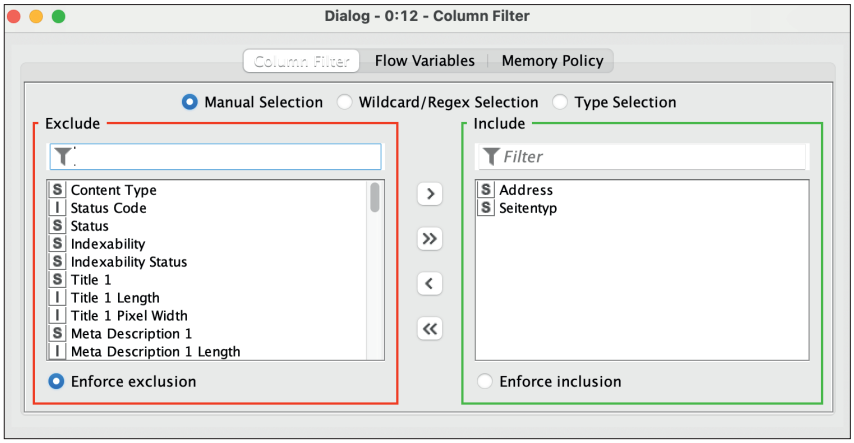


Abb. 6: Der Crawl wird auf die Spalten „Address“ und „Seitentyp“ gefiltert

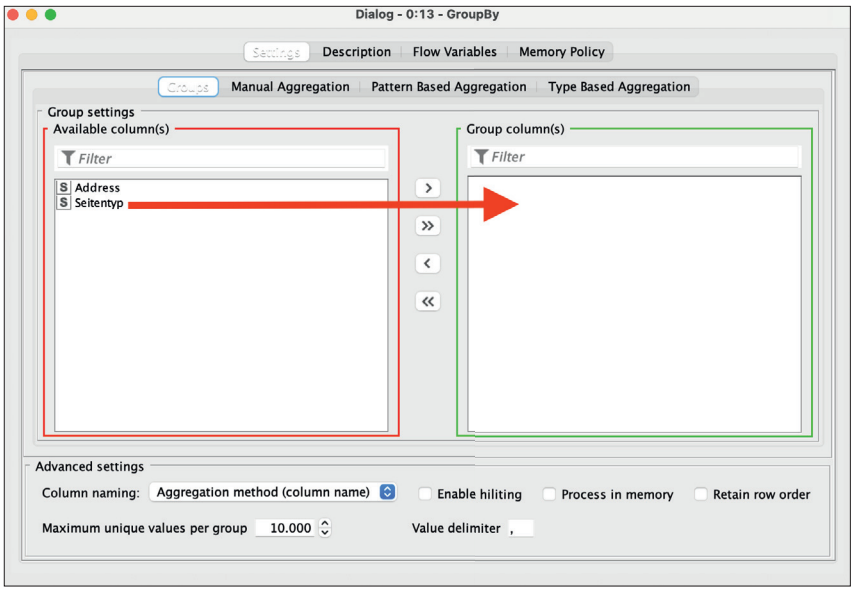


Abb. 7: Gruppieren der Tabelle nach Seitentyp

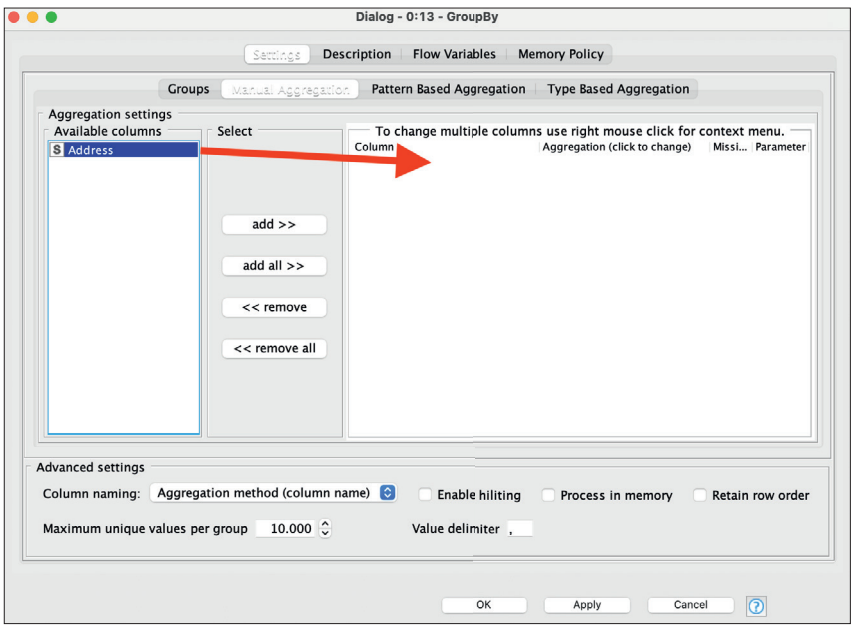


Abb. 8: Aggregation der Spalte „Address“

behalten werden sollen, müssen im rechten Feld stehen, alle, die ausgeschlossen werden sollen, im linken (Abb. 7). Durch die Pfeile lassen sich eine oder alle ausgewählten Spalten in das jeweils andere Feld verschieben.

Gruppieren mit der GroupBy Node

Kein Filter im eigentlichen Sinne ist die Node „GroupBy“. Vielmehr lässt sie sich mit einer Pivot-Tabelle in Excel vergleichen. Nachdem der Crawl auf die Spalten „Address“ und „Seitentyp“ gefiltert wurde, lässt sich mit der GroupBy Node ganz einfach herausfinden, wie viele URLs es pro Seitentyp gibt.

Zunächst muss im Tab „Groups“ festgelegt werden, nach welcher Spalte gruppiert werden soll. Dazu muss die Spalte „Seitentyp“ in das rechte Feld verschoben werden (Abb. 7).

Im nächsten Schritt wird im Tab „Manual Aggregation“ die Spalte „Address“ ausgewählt und durch Doppelklick in das rechte Feld verschoben (Abb. 8.)

Im Bereich „Aggregation“ kann nun gewählt werden, wie die Aggregation erfolgen soll. Dabei können vor allem mathematische Funktionen durchgeführt werden, wenn die zu aggregierende Spalte eine numerische ist. Beispielsweise kann der Durchschnitt oder der Median ermittelt oder die Summe gebildet werden.

Aber auch Spalten, die Text beinhalten, können auf verschiedene Weisen aggregiert werden. So kann mit den Methoden „Count“ bzw. „Unique Count“ ermittelt werden, wie häufig ein Zelleninhalt pro Gruppe (in diesem Fall pro Seitentyp) vorhanden ist (Abb. 10).

Standardmäßig kann jede Gruppe nur 10.000 Einträge haben. Übersteigt eine Gruppe dieses Limit, zeigt sich das in Form eines gelben Dreiecks nach Ausführen der Node. Dann kann

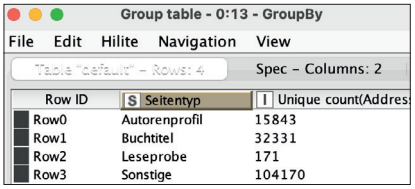


Abb. 9: Die neue gruppierte Tabelle mit der Anzahl an URLs pro Seitentyp

der Wert im Feld „Maximum unique values per Group“ entsprechend erhöht werden (Abb. 10).

Mit der nun erzeugten Tabelle (Abb. 9) können schon erste Erkenntnisse gewonnen werden: Mit der Information, wie viele Seiten es pro Template gibt, können technische Optimierungen anhand der Anzahl an betroffenen Seiten priorisiert werden.

Visualisieren

Zum Abschluss lässt sich die Auswertung auch in Form eines Diagramms visualisieren. Dazu bietet sich die Node „Bar Chart“ an. Mit ihr lässt sich ein einfaches Balkendiagramm erstellen. Nach Hinzufügen der Node muss lediglich in der Konfiguration die „Aggregation Method“ von „Occurence Count“ zu „Sum“ geändert werden (Abb. 11).

Das fertige Chart kann nach dem Ausführen via Rechtsklick und dem Menüpunkt „Interactive View: Grouped Bar Chart“ aufgerufen werden.

Fazit

Durch Filtern und Gruppieren lässt sich auch ein großer und unübersichtlicher Crawl in verdauliche Häppchen aufteilen. Ganz gleich, ob praktische Handlungsempfehlungen wie das Filtern auf alte Jahreszahlen in der Description oder der Blick auf das große Ganze durch Segmentieren und Gruppieren: Das Arbeiten mit Rohdaten erzeugt spannende Insights, die sonst häufig im Verborgenen bleiben. ¶

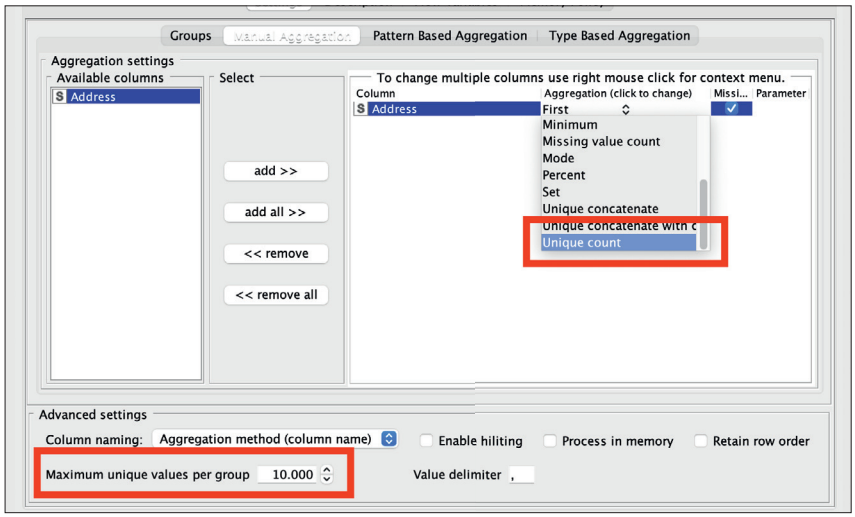


Abb. 10: Mit „Unique Count“ werden die URLs pro Seitentyp gezählt

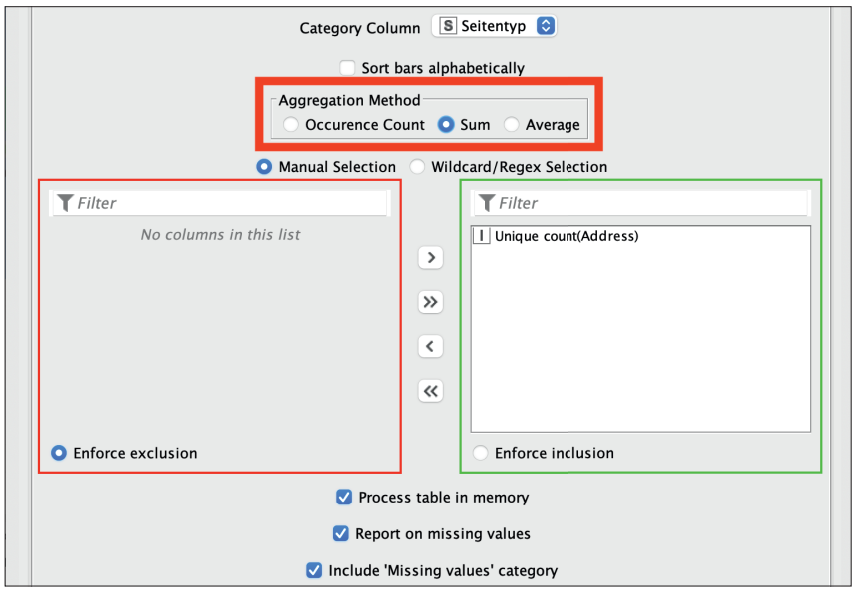


Abb. 11: Die Konfiguration des Bar Chart

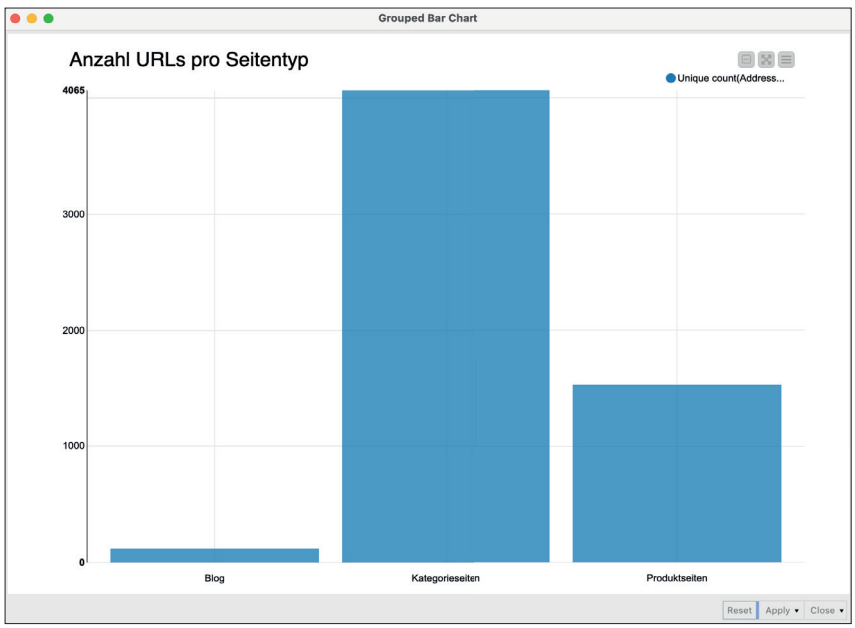


Abb. 12: Das fertige Chart