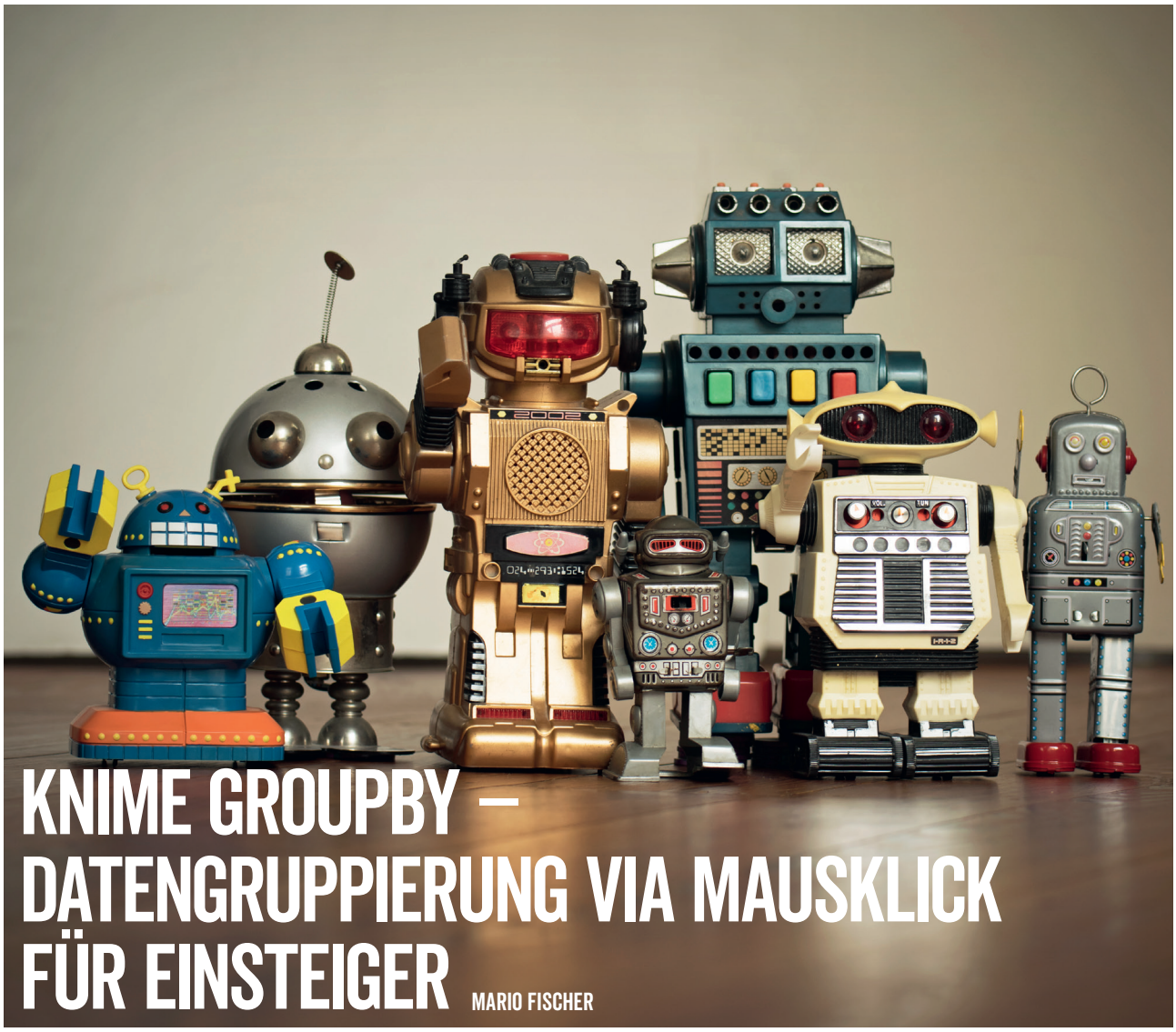


KNIME GROUPBY – DATENGRUPPIERUNG VIA MAUSKLICK FÜR EINSTEIGER

MARIO FISCHER

Open for Innovation KNIME

In den vergangenen Ausgaben haben Sie viele direkte und fertige Anwendungsbeispiele in dem kostenlosen Datentool KNIME kennengelernt. In dieser Ausgabe wollen wir Ihnen zeigen, welche Power Sie mit nur einer einzigen Node nutzen können und dass es sich nur deswegen vielleicht für noch unentschlossene Einsteiger lohnt, sich das Tool auf den Rechner zu holen. Es handelt sich dabei um die Node „GroupBy“. In unserem kleinen Tutorial lernen Sie alles darüber, wie man auch größeren oder komplexeren Datensätzen ihre Informationsschätze entlocken kann. Und das alles nur mit Mausklicks – kein Code, keine Formeln. Wie klingt das für Sie?

Wer mit ernsthaftem Interesse eine Website betreibt, stößt an vielen Stellen auf exportierbare Daten. Teils sind diese bereits aggregiert, teilweise liegen sie im Rohformat vor. Als CSV, als XLS oder in einem anderen Format. Will oder muss man im Alltag mit solchen Daten arbeiten, greift man oft einfach zu Excel und stellt zum Beispiel bei CSV-Dateien je nach Tool/Quelle fest, dass es beim Öffnen bereits alles zerschießt. Mit der Zeit lernt man, solche Probleme zu umgehen und zu fixen. Datenauswertungen in komplexen Tabellen werden dann meist mit Pivottabellen gemacht. Meist nach Schema F und jede Woche/Monat immer wieder. Das ist nicht nur fehleranfällig, sondern kostet auch völlig unnötig Zeit. Zum Glück gibt es einfache Alternativen.

Falls Sie KNIME noch nicht installiert haben, holen Sie dies einfach nach. Auf *www.knime.com* finden Sie direkt einen Down-

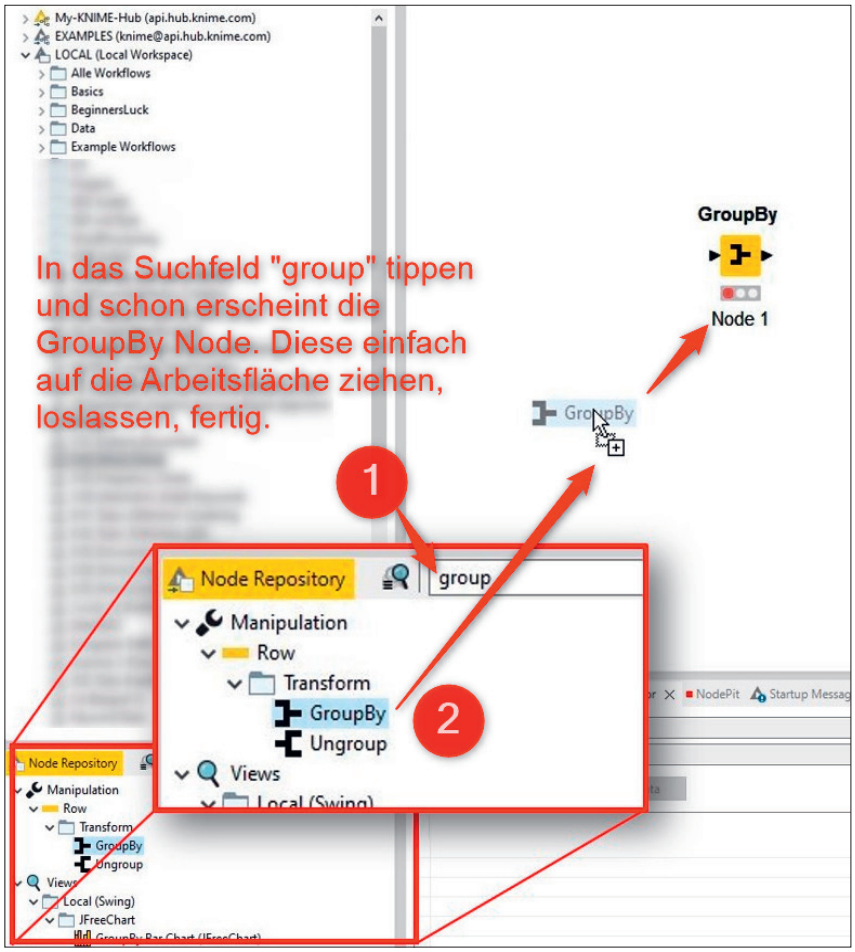


Abb. 1: Suchen Sie im Repository einfach nach der „GroupBy“-Node und ziehen Sie diese auf die Arbeitsfläche.

data_date	site_url	query	is_anonymized_query	country	search_type	device	impressions	clicks	sum_top_position
24.04.2023	https://www.websiteboosting.com/	youtube statistiken	FALSCH	abw	WEB	DESKTOP	17	33	10
24.04.2023	https://www.websiteboosting.com/	seo shop	FALSCH	abw	WEB	DESKTOP	4	259	3
24.04.2023	https://www.websiteboosting.com/	robot txt no follow	FALSCH	abw	WEB	DESKTOP	3	7	3
24.04.2023	https://www.websiteboosting.com/	website analyse google	FALSCH	ago	WEB	DESKTOP	2	5	4
24.04.2023	https://www.websiteboosting.com/	fachbücher	FALSCH	ago	WEB	DESKTOP	2	33	9
24.04.2023	https://www.websiteboosting.com/	meta robots noindex	FALSCH	ago	WEB	MOBILE	2	20	9
24.04.2023	https://www.websiteboosting.com/	seitenquelltext anzeigen	FALSCH	ago	WEB	DESKTOP	2	13	11
24.04.2023	https://www.websiteboosting.com/	suchmaschine test	FALSCH	ago	WEB	DESKTOP	2	6	11
24.04.2023	https://www.websiteboosting.com/	drupal cms	FALSCH	ago	WEB	MOBILE	2	7	12
24.04.2023	https://www.websiteboosting.com/	bing unternehmen eintrage	FALSCH	alb	WEB	MOBILE	2	24	12
24.04.2023	https://www.websiteboosting.com/	ankertexte	FALSCH	alb	WEB	DESKTOP	2	4	9
24.04.2023	https://www.websiteboosting.com/		WAHR	alb	WEB	MOBILE	2	9	9
24.04.2023	https://www.websiteboosting.com/	joomla frankfurt	FALSCH	alb	WEB	MOBILE	2	2	11
24.04.2023	https://www.websiteboosting.com/	seo reporting	FALSCH	alb	WEB	DESKTOP	1	2	30
24.04.2023	https://www.websiteboosting.com/	usability test	FALSCH	alb	WEB	DESKTOP	1	4	30
24.04.2023	https://www.websiteboosting.com/	alternativen zu screamingfr	FALSCH	alb	WEB	DESKTOP	1	6	31
24.04.2023	https://www.websiteboosting.com/	gute angebote schreiben m	FALSCH	alb	WEB	DESKTOP	1	3	32
24.04.2023	https://www.websiteboosting.com/	webseite analysieren	FALSCH	alb	WEB	DESKTOP	1	2	36
24.04.2023	https://www.websiteboosting.com/	mitbewerber url	FALSCH	alb	WEB	MOBILE	1	0	40
24.04.2023	https://www.websiteboosting.com/	web anaylitics	FALSCH	alb	WEB	DESKTOP	1	0	44
24.04.2023	https://www.websiteboosting.com/	tag-[illegible] ent sy ten	FALSCH	[illegible]	[illegible]	MOBIL	1	0	47
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	DESKT	[illegible]	0	49

Abb. 2: Immer wenn in Spalten mehrere gleiche Daten auftauchen, ist die „GroupBy“-Node nützlich (beispielsweise Daten aus der Search Console).

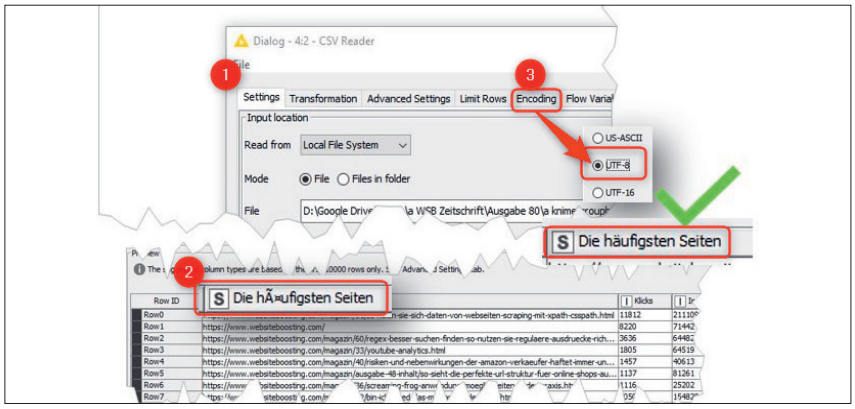
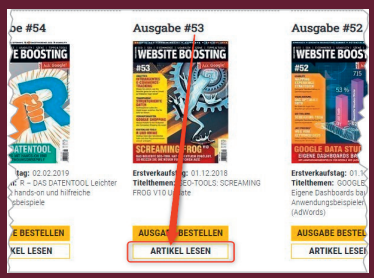


Abb. 3: Das Einlesen von Dateien in KNIME geht kinderleicht.

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*.



loadbutton. Die Installation verläuft in der Regel völlig problemlos. In Ausgabe 53 (online kostenlos unter *einfach.st/knime53* nachlesbar) finden Sie dazu bei Bedarf gerade für Einsteiger weitere nützliche Infos und einige Beispiele, was man mit KNIME alles tun kann.

Nach dem Öffnen von KNIME erstellen Sie ganz einfach unter „File / New“ über den Wizard einen neuen Workflow bzw. eine leere Arbeitsfläche. Der Wizard hält nur zwei Möglichkeiten bereit und Sie wählen „New KNIME Workflow“.

Auf der linken Seite nutzen Sie nun das Node Repository, dort finden Sie alle Nodes hinterlegt, die aktuell verfügbar sind. Statt über die Hierarchie zu klicken, geben Sie einfach in das Suchfeld „group“ ein und schon erscheint unten in Echtzeit gefiltert alles, was dies im Namen trägt (Abbildung 1, Ziffer 1). Greifen Sie sich die Node „GroupBy“ und ziehen Sie diese mit der Maus auf die Arbeitsfläche (Ziffer 2). Nach dem Loslassen erscheint sie dort, wie in der Abbildung gezeigt.

Die Node zeigt auf der quer liegenden grauen Ampel rot. Das ist korrekt und normal. Das rote Signal zeigt an, dass noch keine Daten am Eingang der Node vorhanden sind. Den Eingang symbolisiert das kleine schwarze Dreieck auf der linken Seite – das auf der rechten den Ausgang.

Wir brauchen für unsere Node also Daten. Im Prinzip können Sie beliebige Daten verwenden, hier in diesem Tutorial verwenden wir zunächst für einen ersten Eindruck der Funktionsweise der „GroupBy"-Node Daten aus der Google Search Console. Generell gilt, möchten Sie andere Daten nutzen, transferieren Sie das Erklärte einfach entsprechend auf Ihre Datensätze. Abbildung 2 zeigt, wann Sie die „GroupBy"-Node am besten einsetzen können. Nämlich immer dann, wenn in einer oder mehreren Spalten gleiche Einträge vorhanden sind. Diese gilt es, nach unterschiedlichen Fragestellungen zu gruppieren und zu aggregieren. Die wahre Power liegt in der Flexibilität dieser Aggregationen, die sich ganz einfach mit der Maus zusammenklicken lassen.

Öffnet man CSV-Dateien in Excel, ergibt sich häufig ein Problem. Die deutschen Umlaute fehlen und beim Versuch, die Daten sauber in Spalten zu bekommen (über das Menü „Daten"/„Text in Spalten"), wird aus Werten wie 24.8 plötzlich der 24. August. Für Datenprofis eine Kleinigkeit, für Einsteiger oft eine sehr ärgerliche Hürde. Diese Probleme hat man in KNIME übrigens generell nicht.

Ziehen Sie nun einfach die gewünschte Datei (z. B. CSV oder XSL) direkt aus Ihrem Filesystem auf die Arbeitsfläche von KNIME. Beim Loslassen erscheint sofort der Konfigurationsdialog und Sie sehen unten bereits eine Vorschau auf Ihren Datensatz. Diesen Konfigurationsdialog können Sie später jederzeit mit der rechten Maustaste auf der Node aufrufen. Im ersten Reiter, den Settings (Ziffer 1 in Abbildung 3) müssen Sie normalerweise keine Anpassungen vornehmen. Je nach Datenquelle tritt oft ein Umlautproblem auf (Ziffer 2). Unter dem Reiter „Encoding" (Ziffer 3) klicken Sie auf den Zeichensatz „UTF-8" und das Problem verschwindet im gesamten Daten-

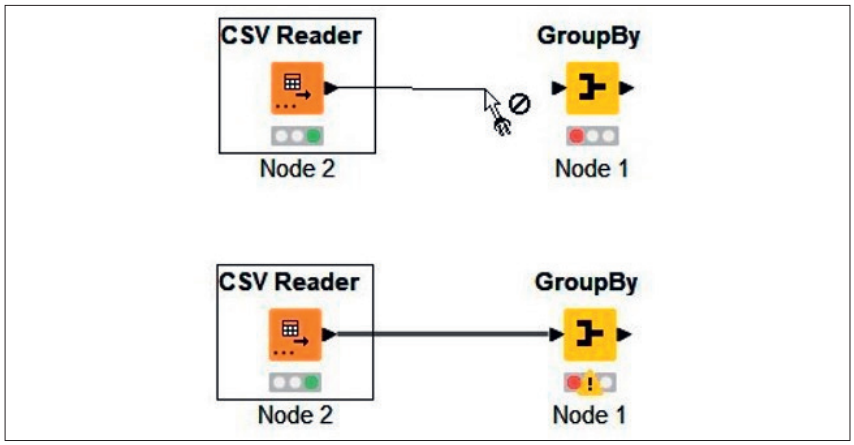


Abb. 4: Datenverbindungen zieht man einfach mit der Maus.

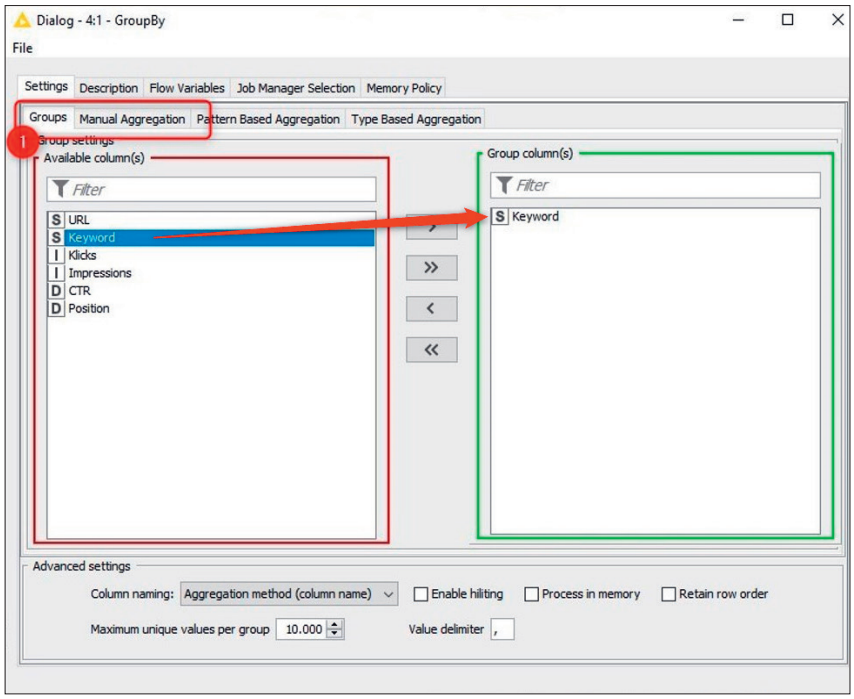


Abb. 5: Schritt 1: Spaltendaten werden einfach per Mausklick auswählt.

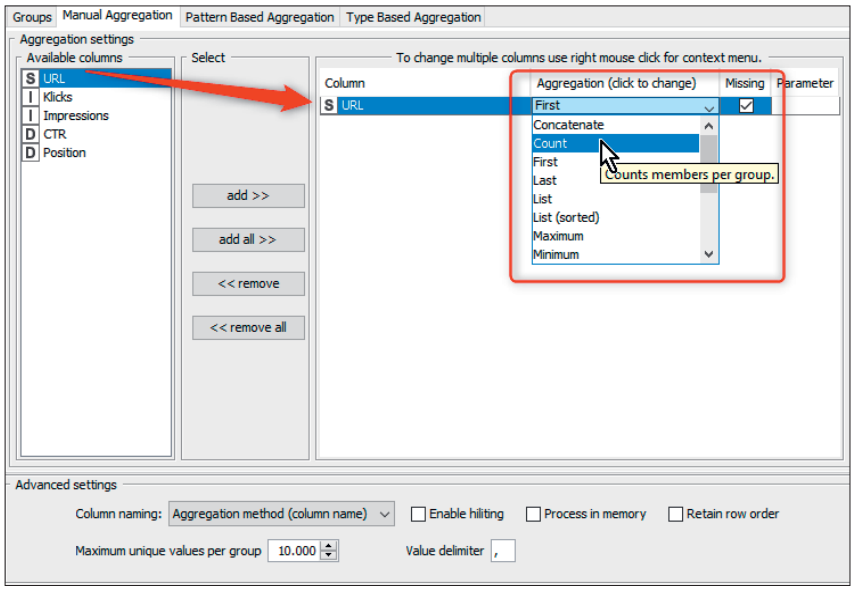


Abb. 6: Schritt 2: Auswahl der zu aggregierenden Datenspalten und wie aggregiert werden soll

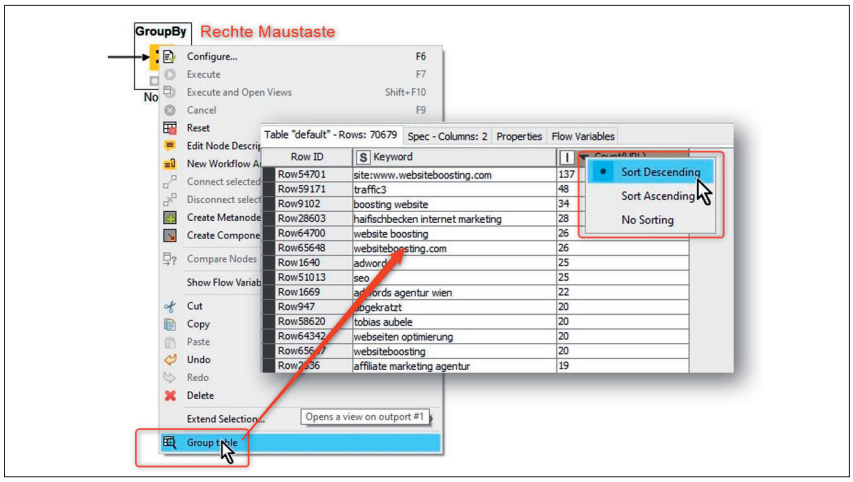


Abb. 7: Wie viele unterschiedliche URLs ranken für ein Keyword?

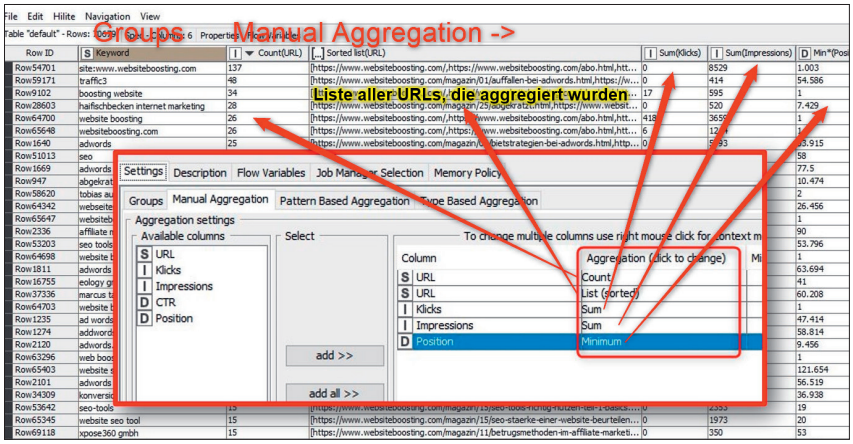


Abb. 8: Je nach Auswahl und Aggregationsmethode werden neue Datenspalten erzeugt.

satz. Eine Fehlerkennung von Daten bzw. eine unerwünschte Umwandlung in andere Formate wie zum Beispiel das Datum tritt bei KNIME nicht auf.

Die Node „CSV Reader“, die automatisch durch das Ziehen der Datei auf die Arbeitsfläche erschienen ist, hat jetzt noch ein gelbes Licht. Das bedeutet, Daten sind da, aber die Node muss noch ausgeführt werden. Das ist einfach. Rechte Maustaste und „Execute“ macht die kleine Ampel grün und zeigt unten im „Node Monitor“ sofort alle Daten der CSV-Datei an. Verändern Sie etwas in der Konfiguration einer Node oder den Daten, die hineinfließen, bekommen Sie das dann wieder gelbe Signal mit „Execute“ wieder grün. Auch der nächste Schritt ist einfach. Man zieht mit der Maus am Ausgang der Node „CSV Reader“ eine Linie zum Eingang der „GroupBy“-Node (Abbildung 4). Die Verbindung schnappt auf den Eingang ein und es erscheint eine Warnung (gelbes Ausrufezeichen). Das ist der Hinweis, dass diese Node noch konfiguriert werden muss. Dazu ruft man per rechte Maustaste den ersten Punkt „Configure“ auf.

Abbildung 5 zeigt den Konfigurationsdialog. Die oberste Reihe der Reiter „Settings“, „Description“ können Sie ignorieren und achten Sie bitte darauf, diese Reihe nicht mit der wichtigen darunter (Ziffer 1) zu verwechseln. Dort wählen Sie aus, nach welchen Datenspalten Sie gruppieren wollen (Groups) und welche der anderen Spaltendaten wie aggregiert werden sollen („Manual Aggregation“). Wählen Sie zur Gruppierung die Spalte URL aus. Ein Doppelklick befördert sie nach rechts und sieht sie als Element vor, nach dem gruppiert werden soll.

Nun wechselt man den Reiter zu „Manual Aggregation“ und wählt dort nach dem gleichen Schema die zu aggregierenden Datenspalten aus (Abbildung 6). In diesem Beispiel hier geht es zunächst darum, wie viele unterschiedliche URLs für ein Keyword ranken. Dazu befördert man per Klick „URL“ nach rechts in die Auswahl. Diese erscheint dann mit einem Pull-down-Menü unter „Aggregation (click to change)“. Dort wählt man „Count“ aus, also eine einfache Zählung, wie viele Einträge in der Datenspalte URL sind – gruppiert nach je einem Keyword. Dazu gleich mehr.

Wie man die soeben konfigurierte Node ausführt, ist ja nun schon bekannt. Rechts Maustaste und „Execute“. Sofort erscheint unten im „Node Monitor“ das Ergebnis. Alternativ klicken Sie erneut die rechte Maustaste und wählen Sie ganz unten „Group table“ aus. In der erscheinenden Ansicht kann man mit einem Klick auf die Spaltenüberschriften auf- oder absteigend sortieren (Abbildung 7). Geübte SEOs sehen sofort: Hier ranken viel zu viele URLs für ein Keyword. So sind meist keine Top-Positionen zu erreichen. Im gezeigten Beispiel rankt die Domain mit „SEO“ tatsächlich mit 25 unterschiedlichen URLs.

Nächste Frage. Welche URLs ranken jeweils für ein Keyword? Ab jetzt nutzen wir einfach immer wieder die Konfigurationsansicht mit dem Reiter „Manual Aggregation“. Zusätzlich wollen wir die Anzahl Klicks und Impressions für jedes Keyword (das ja mehrfach mit unterschiedlichen URLs auftauchen kann) summieren. Und wir wollen wissen, was jeweils die beste Ranking-Position für jedes Keyword ist. Dazu klicken wir wie in Abbildung 8 gezeigt nochmals URL nach rechts, wählen aber im Pull-down-Menü „List (sorted)“ aus. Für Klicks und Impressions wählt man „Sum“ aus und für Position „Minimum“.

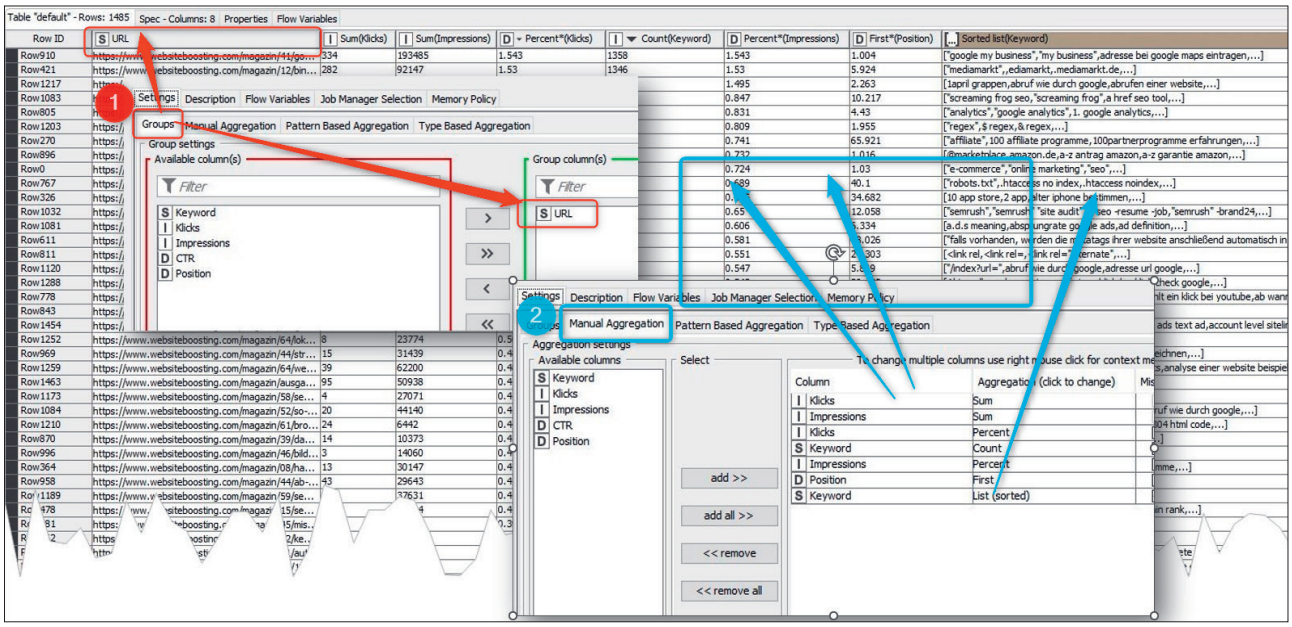


Abb. 9: Eine Auswertung nicht nach Keywords, sondern nach URL? Kein Problem

Abbildung 8 zeigt prinzipiell, wie einfach es ist, Daten entsprechend zu gruppieren bzw. mit unterschiedlichen Methoden zu aggregieren. In der Aggregationsspalte gibt es mehrere wählbare Einträge, von denen hier einige wichtige kurz vorgestellt werden.

» **Count und Unique Count**
Count zählt, wie oft ein Datensatz für die in „Group" gewählte Filterung vorkommt. Im Beispiel wurde nach Keyword gruppiert. Sobald also ein Keyword mehrfach vorkommt, wird dies (über die Auswahl von URL) gezählt. Kann es vorkommen, dass die Kombination „Keyword" und „URL" nicht nur einmal (wie hier) vorkommt, sondern mehrfach und möchte man das aber nur jeweils einmal zählen, wählt man Unique Count als Methode aus.

» **Percent und Percent from Unique Count**
Die Erklärung hier ist denkbar einfach. Es wird der Anteil einer Datenzeile (hier eines Keywords) jeweils am Gesamten in einer Spalte ausgegeben. Würde man zum Beispiel nochmals „Klicks" in die Auswahl nach rechts befördern und wählt dann „Percent" unter Aggregation, wird eine weitere Spalte hinzu-

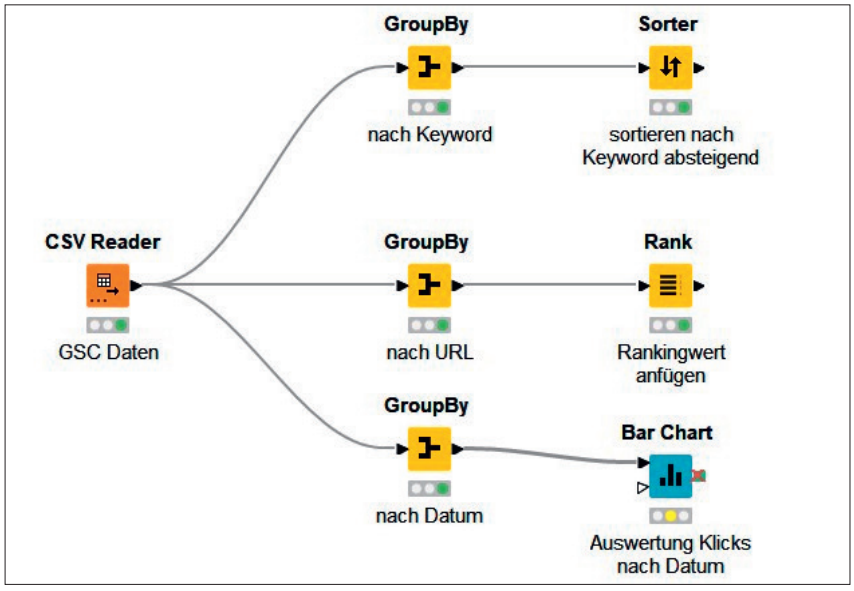
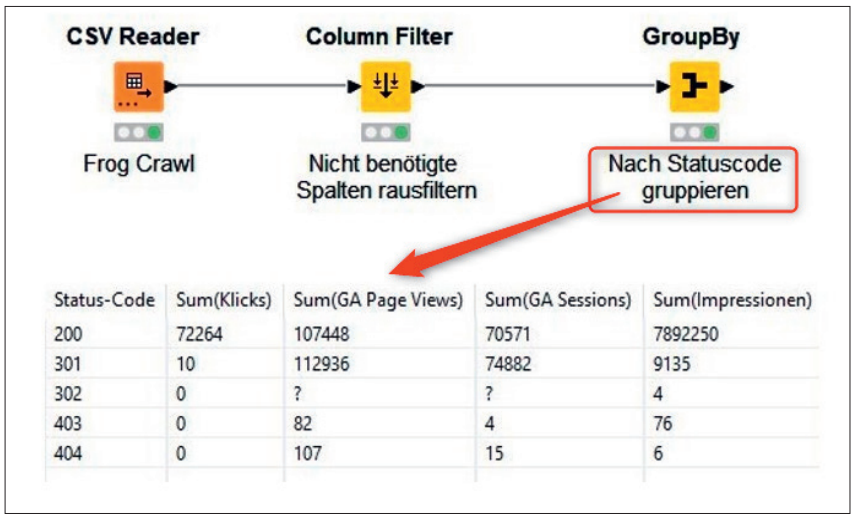
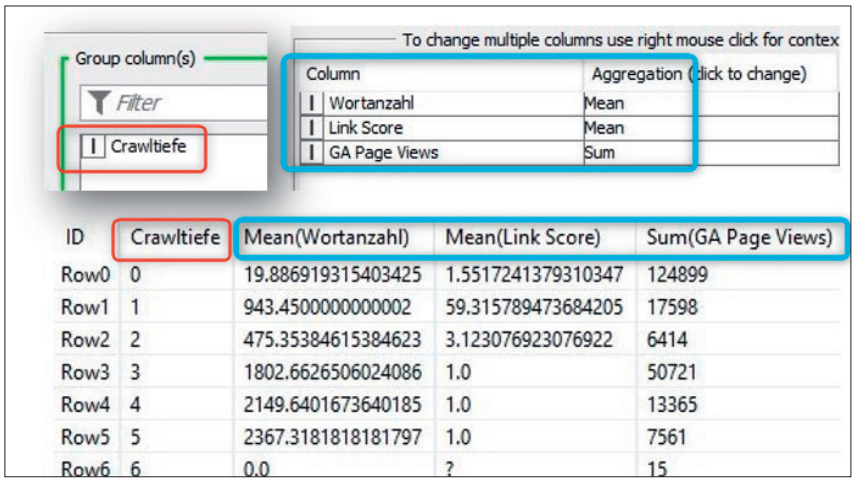


Abb. 10: In einem Workflow sind viele gleichzeitige Auswertungen möglich.



Status-Code	Sum(Klicks)	Sum(GA Page Views)	Sum(GA Sessions)	Sum(Impressionen)
200	72264	107448	70571	7892250
301	10	112936	74882	9135
302	0	?	?	4
403	0	82	4	76
404	0	107	15	6

Abb. 11: Schneller Überblick über einen Screaming-Frog-Crawl mit API-Daten



ID	Crawltiefe	Mean(Wortanzahl)	Mean(Link Score)	Sum(GA Page Views)
Row0	0	19.886919315403425	1.5517241379310347	124899
Row1	1	943.4500000000002	59.315789473684205	17598
Row2	2	475.35384615384623	3.123076923076922	6414
Row3	3	1802.6626506024086	1.0	50721
Row4	4	2149.6401673640185	1.0	13365
Row5	5	2367.3181818181797	1.0	7561
Row6	6	0.0	?	15

Abb. 12: Ein weiteres Beispiel: Wortanzahl, LinkScore (entspricht etwa dem internen PageRank) und Traffic nach Klicktiefe gruppiert

gefügt, die den Prozentanteil eines Keywords an allen Klicks berechnet. Percent from Unique Count berücksichtigt Mehrfachdaten (siehe oben bei Count).

» **Sum**
Hier werden alle Werte für eine Datengruppe aufsummiert.

» **Range**
Dies gibt den Abstand zwischen dem niedrigsten und dem höchsten Wert aus, also das „Delta".

» **First, Last, Minimum, Maximum**
Dies gibt entweder den ersten, den letzten, den niedrigsten oder den höchsten Wert aus, der in einer Tabelle auftaucht. Möchte man zum Beispiel die beste Ranking-Position, wählt man Minimum. Für die schlechteste das Maximum, also den höchsten Wert für ein Keyword.

» **Mean, Median, Modus, Standard Deviation, Variance**
Dies gibt den Mittelwert, den Median, den Modus, die Standardabweichung oder die Varianz der Daten für statistische Auswertungen aus.

» **List, List (sorted), SetList** fügt in eine Zelle alle auftauchenden Werte hintereinander durch Komma getrennt ein, Set macht prinzipiell das Gleiche, aber ohne gegebenenfalls vorhandene Duplikate. Der Sinn einer solchen Aggregationsmethode ist, dass man die Daten später in einem Workflow bei Bedarf wieder „auseinanderziehen" kann. Im vorliegenden Fall wertet man also nicht nur die Anzahl an Mehrfachrankings für ein Keyword aus, sondern behält in den Daten auch alle jeweils betroffenen URLs.

Geht es darum, Metriken und Keywords nach den URLs zu gruppieren, ist auch das mit wenigen Mausklicks erledigt. Dazu wird statt „Keyword" im Reiter „Groups" einfach „URL" in den Container rechts gestellt und unter „Manual Aggregation" werden die entsprechenden Spalten und Aggregationsmethoden ausgewählt. Rechte Maustaste und „Execute" und schon liegen alle Daten wie gewünscht vor (Abbildung 9). Hat man beispielsweise noch ein Datum für jeden Datensatz vorliegen, ist gegebenenfalls auch eine gesonderte Gruppierung nach Datum, Woche oder Monat sinnvoll.

Der hier verwendete Datensatz aus der Search Console ist natürlich nur beispielhaft zu sehen. Die „GroupBy"-Node kann prinzipiell für alle Daten verwendet werden, wo Datensätze in Zeilen mehrfach identisch vorkommen. Der Vorteil einer Verarbeitung/Auswertung zum Beispiel in Excel liegt darin, dass solche Analysen nur einmal erstellt werden müssen und die Umformung/Gruppierung „zerstörungsfrei" funktioniert. Man verwendet einen Eingangsdatensatz und kann über die Datenleitungen diese Daten an viele Nodes übergeben und auch unterschiedliche Darstellungen am Ende verwenden. Abbildung 10 zeigt schematisch eine solche einfache Auswertungslogik. Natürlich können einer Gruppierung weitere Nodes folgen oder auch Charts verknüpft werden. Hat man neue oder aktualisierte Daten, tauscht man einfach nur die Datenbasis (hier in der „CSV Reader"-Node) aus bzw. wählt in den Einstellungen der einlesenden Node eine andere Quelldatei aus. Über „Execute" wird dann automatisch alles neu berechnet.

Abbildung 11 zeigt, wie einfach eine exportierte Datei aus einem Screaming-Frog-Crawl, bei dem per API Call-Daten aus Google Analytics und der Search Console erfasst wurden, nach dem Statuscode gruppiert werden. Denkbar sind zum Beispiel auch Durchschnittswerte nach Wortanzahl, eingehenden Links, der Linkstärke und natürlich allen anderen Metriken, die man für eine Analyse oder einfach nur für einen ersten Eindruck braucht.

Ein letztes Beispiel zeigt Abbildung 12. Dort wurden einige Metriken bzw. deren Mittelwerte nach Klicktiefe (im Frog Crawltiefe genannt) gruppiert. Hier ist leicht zu erkennen, dass die URLs mit viel Text sehr weit hinten in der Sitehierarchie liegen und diese von den eingehenden Links her gesehen wenig Wert übermittelt bekommen.

Fazit

Allein die Mächtigkeit und Einfachheit der „GroupBy"-Node in KNIME rechtfertigt fast schon die Beschäftigung mit dem kostenlosen Tool. Schnell entdeckt man dann weitere Möglichkeiten, die die wahre Power Stück für Stück offenlegen. Aber selbst wenn man sich nur für häufige Auswertungen einige Workflows zusammenstellt und abspeichert, lässt sich enorm viel Zeit sparen.