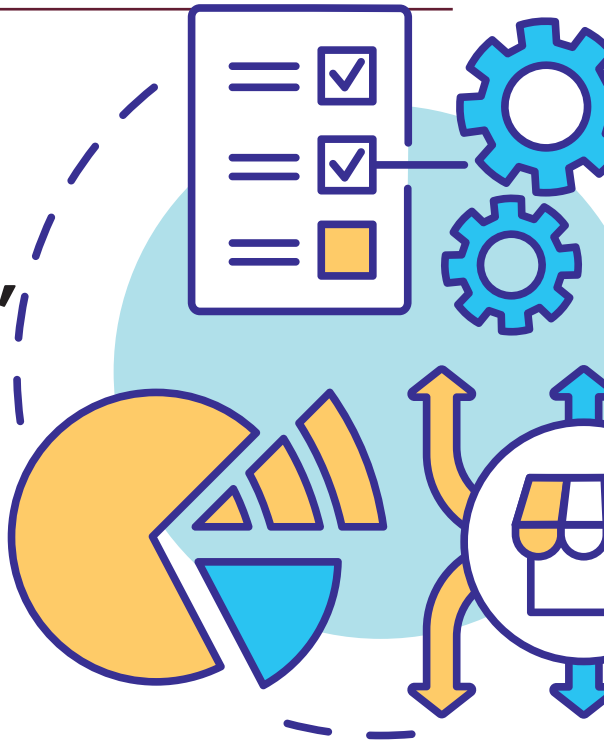
Sarah Zeus

Filtern, Clustern, Segmentieren: Rule-Nodes in KNIME für Einsteiger



Für die Anwendung von Regeln bei komplexen Datenanalysen stellt KNIME eine leicht verständliche, benutzerfreundliche und effiziente Alternative zu Excel dar, dessen Anwendung vor allem bei wiederkehrenden Analysen zeitintensiv sein kann. Im folgenden Beitrag, der sich besonders für Einsteiger eignet, erklärt Sarah Zeus Schritt für Schritt anhand einfacher Anwendungsfälle zum Filtern oder Clustern von Daten, wie verschiedene Rule-Nodes in KNIME Ihre Arbeit erleichtern können.

Die Rule-Nodes: Gemeinsamkeiten und Unterschiede

Beleuchtet werden in diesem Beitrag zwei Rule-Nodes in KNIME: der Rule-based Row Filter und die Rule Engine. Obwohl sich die Nodes für unterschiedliche Zwecke eignen, bietet sich die gemeinsame Betrachtung an, da ihre Konfigurationsmöglichkeiten sehr ähnlich sind.

So ermöglichen beide Nodes die Anwendung von Regeln auf eine Tabelle, wobei der Rule-based Row Filter – wie der Name sagt – ein Filter-Knoten ist, der Zeilen aus einer Tabelle filtert, die den gewählten Kriterien entsprechen oder nicht entsprechen. Wenn man es also mit großen Datenmengen zu tun hat und nur bestimmte Teile für die weitere Analyse extrahieren möchte, nutzt man am besten diesen Knoten: zum Beispiel, wenn in einer Keyword-Liste nur nach Keywords gefiltert werden soll, die einen bestimmten Begriff beinhalten, oder wenn aus einer URL-Liste nur URLs mit einem bestimmten Status-Code benötigt werden. Mehrere Kriterien sind dabei kombinierbar.

Auch mithilfe der Rule-Engine wird jede Zeile auf Regeln mit ein oder mehreren Bedingungen überprüft. Treffen die Kriterien zu, wird hier nicht gefiltert, sondern es wird entweder ein Ergebniswert in einer neuen Spalte hinzugefügt oder ein Wert in einer bestehenden Spalte gemäß der definierten Regel ersetzt. Der Rule-Engine-Knoten kann also eingesetzt werden, wenn Werte in einer Tabelle geändert, ersetzt oder kategorisiert werden sollen. Er ist besonders sinnvoll, wenn eine Reihe von Bedingungen auf die Daten angewendet werden sollen. Angenommen, man möchte Keywords nach bestimmten Begriffen oder nach Brand- und generischen Keywords clustern, URLs nach bestimmten Kriterien segmentieren oder nach Traffic-Daten klassifizieren, dann eignet sich hierfür die Rule-Engine. Auch hier sind mehrere Bedingungen kombinierbar.

Die Funktionen dieser Knoten können natürlich auch in Excel umgesetzt werden, zum Beispiel über Filter oder Funktionen wie „IF“ und „AND“. Für komplexere Anwendungsfälle, große Datenmengen oder wenn die Analyse wiederholt durchgeführt und deswegen zumindest teilweise automatisiert werden soll, ist KNIME oft die einfacher anwendbare, flexiblere und effizientere Wahl.

DER AUTOR

Sarah Zeus ist Online Marketing Consultant bei The Boutique Agency, einer Digital-Marketing-Agentur aus München. Sie berät Kunden in den Bereichen SEO und SEA. Seit einigen Jahren bevorzugt sie KNIME als Tool zur Automatisierung regelmäßiger Analysen.

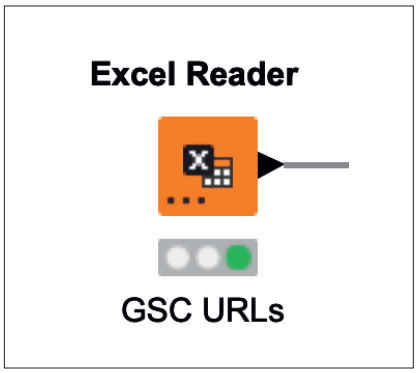


Abb. 1: Durch Hinzufügen der Datei öffnet sich eine Excel-Reader-Node

Row ID	Page	Clicks	Impressions	CTR	Position
Row0	https://www.beispiel.de/shop/hunde	10002	66058	0.151	8.4
Row1	https://www.beispiel.de/shop/hunde/hundefutter	2686	113163	0.024	9.5
Row2	https://www.beispiel.de/shop/hunde/hundefutter/produkt-a	2511	25812	0.097	11.4
Row3	https://www.beispiel.de/shop/hunde/hundefutter/produkt-b	2346	42239	0.056	22.1
Row4	https://www.beispiel.de/shop/hunde/hundefutter/produkt-c	2169	47453	0.046	10.2
Row5	https://www.beispiel.de/shop/hunde/hundefutter/produkt-d	2033	82577	0.025	11.1
Row6	https://www.beispiel.de/shop/hunde/hundefutter/produkt-e	1659	39615	0.042	8.7
Row7	https://www.beispiel.de/shop/hunde/hundefutter/produkt-f	1604	28871	0.056	10.5
Row8	https://www.beispiel.de/shop/hunde/hundefutter/produkt-g	1479	35494	0.042	10.3
Row9	https://www.beispiel.de/shop/hunde/hundefutter/produkt-h	1300	44277	0.029	12.7
Row10	https://www.beispiel.de/shop/hunde/hundefutter/produkt-i	1196	20177	0.059	18.1
Row11	https://www.beispiel.de/shop/hunde/hundefutter/produkt-j	1092	21169	0.052	29.4
Row12	https://www.beispiel.de/shop/hunde/hundefutter/produkt-k	1025	17313	0.059	38.4
Row13	https://www.beispiel.de/shop/hunde/hundespielzeug/produkt-l	993	12667	0.078	10.3
Row14	https://www.beispiel.de/shop/hunde/hundespielzeug/produkt-m	972	26756	0.036	20.7
Row15	https://www.beispiel.de/shop/hunde/hundespielzeug/produkt-n	942	14280	0.066	14.7
Row16	https://www.beispiel.de/shop/hunde/hundespielzeug/produkt-o	938	26639	0.035	18.4

Abb. 2: Vorschau auf die Tabelle

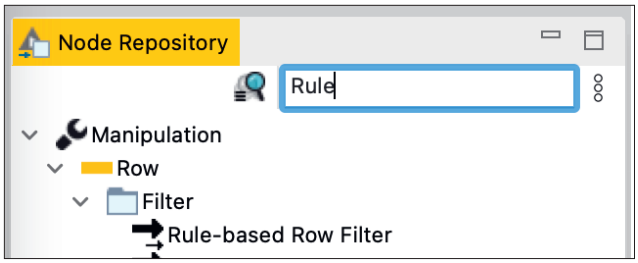


Abb. 3: Node im Node Repository suchen

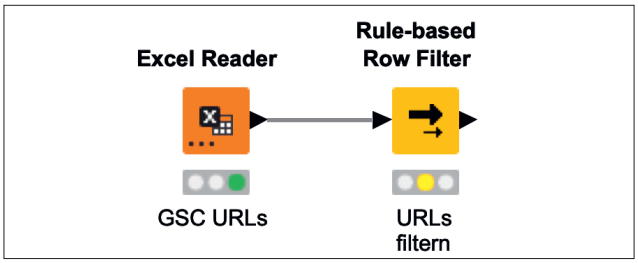


Abb. 4: Excel-Reader und Rule-based Row Filter miteinander verbinden

Konfiguration des Rule-based Row Filters

Ein Beispiel: Es liegt eine Liste mit URLs vor bzw. ein Export aus der Google Search Console, aus der für die weitere Analyse nur URLs mit bestimmten Merkmalen relevant sind. Wie beschrieben kann man hierfür die Filter-Node nutzen. Bevor wir einen Blick auf die Konfiguration werfen, müssen Daten importiert werden. Dafür zieht man die entsprechende Datei in die Arbeitsfläche von KNIME, woraufhin sich eine geeignete Reader-Node mit Vorschau auf die Tabelle öffnet - siehe Abbildung 2.

Im Anschluss an den Excel-Reader-Knoten fügen wir einen Rule-based Row Filter hinzu. Dafür sucht man im Node Repository (links unten) nach der entsprechenden Node, zieht sie in die Arbeitsfläche und zieht eine Verbindung zwischen den beiden Nodes (Abbildung 3 und 4). Nach Ausführung des ersten Knotens ist der Filter bereit zur Konfiguration.

Mit Doppelklick auf den Filter-Knoten öffnet sich ein Dialog. Oben rechts in der Column List werden alle Spalten aus der Tabelle aufgelistet (Abbildung 5, Ziffer 1), mit denen Regeln gebildet werden können. Daneben findet man eine Auswahl an möglichen Funktionen (Abbildung 5, Ziffer 2), im Feld „Expression“ (Abbildung 5, Ziffer 3) werden die Regeln definiert.

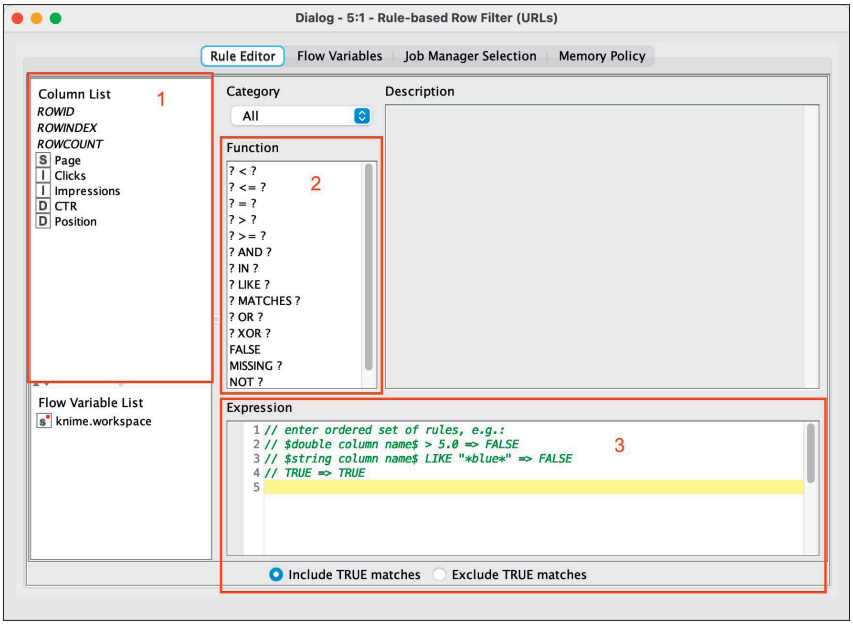


Abb. 5: Konfigurationsmöglichkeiten des Rule-based Row Filters

Jede Regel wird in eine neue Zeile geschrieben. Eine Regel besteht aus einer Bedingung, auf die die Zeilen der Tabelle geprüft werden, und einem Ergebnis, das entweder „TRUE“ oder „FALSE“ lauten kann. Unter dem Eingabefeld wählt man nach Definition der Regeln aus, ob man Zeilen, auf die eine Regel mit dem Ergebnis „TRUE“ zutrifft, in die Ergebnistabelle aufnehmen oder sie ausschließen möchte.

Beginnt eine Zeile mit „//“, handelt es sich um einen Kommentar, der nicht als Regel interpretiert wird. Die Regeln werden von oben nach unten verarbeitet. Hat man also mehrere Regeln defi-

niert und die erste Regel trifft zu, wird das Ergebnis nicht überschrieben, auch wenn eine darauffolgende Regel ebenfalls zutreffen sollte.

Für das Beispiel benötigen wir die Spalte „Page“. Per Doppelklick auf die Spalte in der Colum List wird sie ins Expression-Feld eingefügt, durch das Zeichen $ vor und nach dem Spaltennamen wird angegeben, dass es sich um eine Spalte handelt. Wir wollen in die Ergebnistabelle nur URLs aufnehmen, die „hundefutter“ oder „katzenfutter“ beinhalten. Dafür bildet man zum Beispiel eine Regel mit MATCHES in Kombination mit einer einfachen RegEx wie in Abbildung 6. Der Punkt steht in dem regulären Ausdruck für ein beliebiges Zeichen, das Sternchen steht für null oder mehrere Wiederholungen des vorangegangenen Zeichens.

Über „Include TRUE matches“ beschränkt man das Ergebnis auf die Zeilen, auf die die Regel zutrifft. Im Anschluss kann die Node ausgeführt werden. Das Ergebnis lässt sich dann per Rechtsklick auf die Node und Klick auf „Filtered“ prüfen.

Natürlich können die Regeln auch mehrere Bedingungen beinhalten, wie im Beispiel in Abbildung 8, wo mithilfe von „AND“ weitere Bedingungen zur Regel hinzugefügt wurden.

Übrigens gibt es noch eine ähnliche nützliche Rule-Node: den Rule-based Row Splitter. Anstatt die Zeilen zu filtern, werden sie in zwei separate Tabellen aufgespalten: eine Tabelle mit den Daten, auf die die definierten Kriterien zutreffen, und eine Tabelle mit dem Rest. Die Konfiguration funktioniert genau wie oben beschrieben. Der Unterschied ist nur, dass die Node zwei Ausgänge hat – und nicht nur einen.

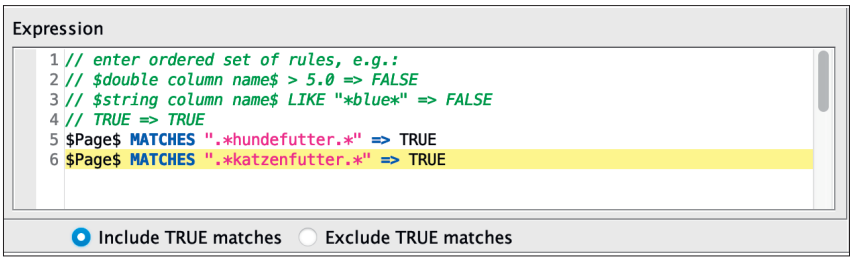
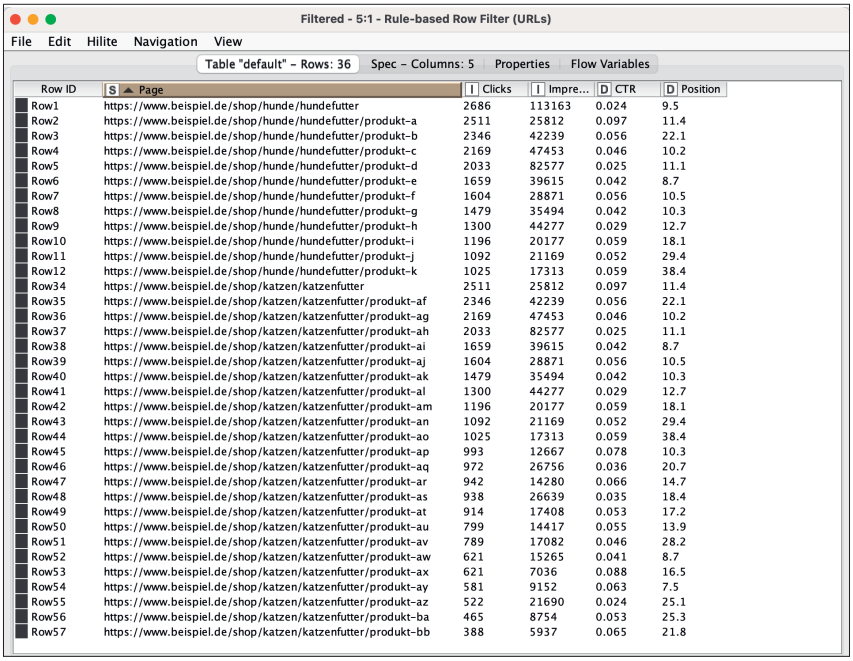


Abb. 6: Erstellung von zwei Regeln zum Filtern der Tabelle



Row ID	Page	Clicks	Impre...	CTR	Position
Row1	https://www.beispiel.de/shop/hunde/hundefutter	2686	113163	0.024	9.5
Row2	https://www.beispiel.de/shop/hunde/hundefutter/produkt-a	2511	25812	0.097	11.4
Row3	https://www.beispiel.de/shop/hunde/hundefutter/produkt-b	2346	42239	0.056	22.1
Row4	https://www.beispiel.de/shop/hunde/hundefutter/produkt-c	2169	47453	0.046	10.2
Row5	https://www.beispiel.de/shop/hunde/hundefutter/produkt-d	2033	82577	0.025	11.1
Row6	https://www.beispiel.de/shop/hunde/hundefutter/produkt-e	1659	39615	0.042	8.7
Row7	https://www.beispiel.de/shop/hunde/hundefutter/produkt-f	1604	28871	0.056	10.5
Row8	https://www.beispiel.de/shop/hunde/hundefutter/produkt-g	1479	35494	0.042	10.3
Row9	https://www.beispiel.de/shop/hunde/hundefutter/produkt-h	1300	44277	0.029	12.7
Row10	https://www.beispiel.de/shop/hunde/hundefutter/produkt-i	1196	20177	0.059	18.1
Row11	https://www.beispiel.de/shop/hunde/hundefutter/produkt-j	1092	21169	0.052	29.4
Row12	https://www.beispiel.de/shop/hunde/hundefutter/produkt-k	1025	17313	0.059	38.4
Row34	https://www.beispiel.de/shop/katzen/katzenfutter	2511	25812	0.097	11.4
Row35	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-af	2346	42239	0.056	22.1
Row36	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ag	2169	47453	0.046	10.2
Row37	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ah	2033	82577	0.025	11.1
Row38	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ai	1659	39615	0.042	8.7
Row39	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-aj	1604	28871	0.056	10.5
Row40	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ak	1479	35494	0.042	10.3
Row41	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-al	1300	44277	0.029	12.7
Row42	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-am	1196	20177	0.059	18.1
Row43	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-an	1092	21169	0.052	29.4
Row44	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ao	1025	17313	0.059	38.4
Row45	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ap	993	12667	0.078	10.3
Row46	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-aq	972	26756	0.036	20.7
Row47	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ar	942	14280	0.066	14.7
Row48	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-as	938	26639	0.035	18.4
Row49	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-at	914	17408	0.053	17.2
Row50	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-au	799	14417	0.055	13.9
Row51	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-av	789	17082	0.046	28.2
Row52	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-aw	621	15265	0.041	8.7
Row53	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ax	621	7036	0.088	16.5
Row54	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ay	581	9152	0.063	7.5
Row55	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-az	522	21690	0.024	25.1
Row56	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-ba	465	8754	0.053	25.3
Row57	https://www.beispiel.de/shop/katzen/katzenfutter/produkt-bb	388	5937	0.065	21.8

Abb. 7: Vorschau auf das Ergebnis des Filterns

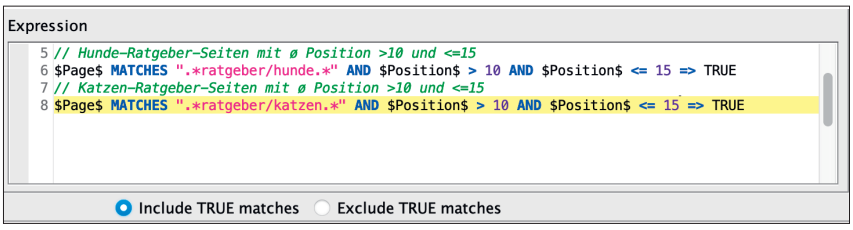


Abb. 8: Filtern der Tabelle nach Hunde- und Katzenratgeberseiten mit durchschnittlichen Positionen › 10 und <= 15

Row ID	Page	Clicks	Impre...	CTR	Position
Row100	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-t	2169	47453	0.046	10.2
Row132	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-az	2169	47453	0.046	10.2
Row104	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-x	1479	35494	0.042	10.3
Row109	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-ac	993	12667	0.078	10.3
Row136	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-bd	1479	35494	0.042	10.3
Row141	https://www.beispiel.de/ratgeber/katzen/katzenzucht/artikel-bi	993	12667	0.078	10.3
Row103	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-w	1604	28871	0.056	10.5
Row135	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-bc	1604	28871	0.056	10.5
Row101	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-u	2033	82577	0.025	11.1
Row133	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-ba	2033	82577	0.025	11.1
Row98	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-r	2511	25812	0.097	11.4
Row130	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-ax	2511	25812	0.097	11.4
Row95	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-o	200	17964	0.011	11.5
Row127	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-au	200	17964	0.011	11.5
Row92	https://www.beispiel.de/ratgeber/hunde/hunderassen/artikel-l	288	10756	0.027	12.1
Row124	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-ar	288	10756	0.027	12.1
Row105	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-y	1300	44277	0.029	12.7
Row137	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-be	1300	44277	0.029	12.7
Row93	https://www.beispiel.de/ratgeber/hunde/hunderassen/artikel-m	254	9572	0.027	13.6
Row125	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-as	254	9572	0.027	13.6
Row82	https://www.beispiel.de/ratgeber/hunde/hunderassen/artikel-b	799	14417	0.055	13.9
Row114	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-ah	799	14417	0.055	13.9
Row146	https://www.beispiel.de/ratgeber/katzen/katzenzucht/artikel-bn	799	14417	0.055	13.9
Row94	https://www.beispiel.de/ratgeber/hunde/hundezucht/artikel-n	216	16433	0.013	14.5
Row126	https://www.beispiel.de/ratgeber/katzen/katzenrassen/artikel-at	216	16433	0.013	14.5

Abb. 9: Ergebnis der Filterung anhand von Regeln mit kombinierten Kriterien

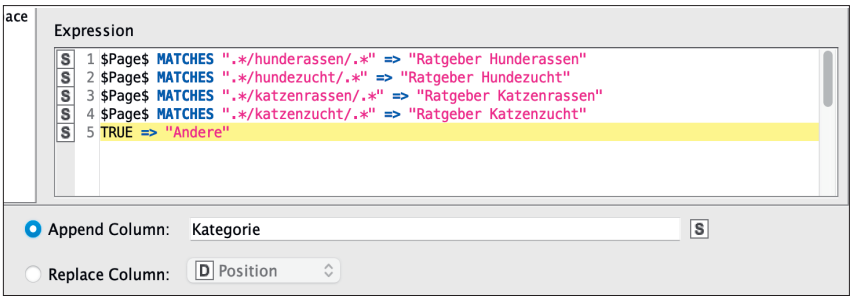


Abb. 11: Regeln zur Klassifikation von URLs

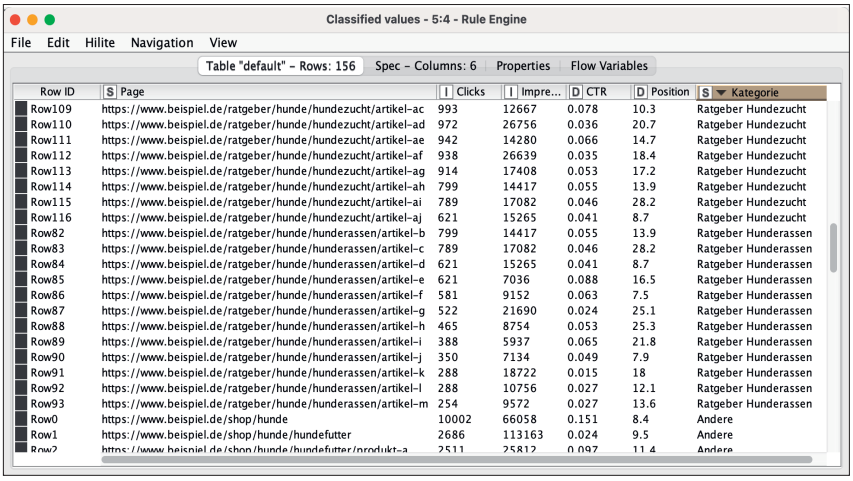


Abb. 12: Prüfung der Ergebnisse. Eine neue Spalte mit den Ergebniswerten wurde der Tabelle hinzugefügt.

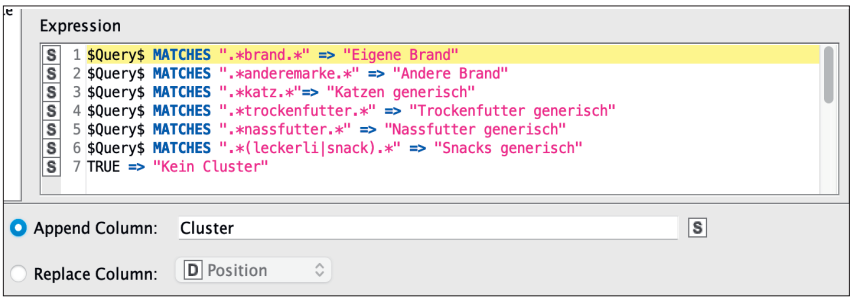


Abb. 13: Regeln zum Clustern von Suchbegriffen

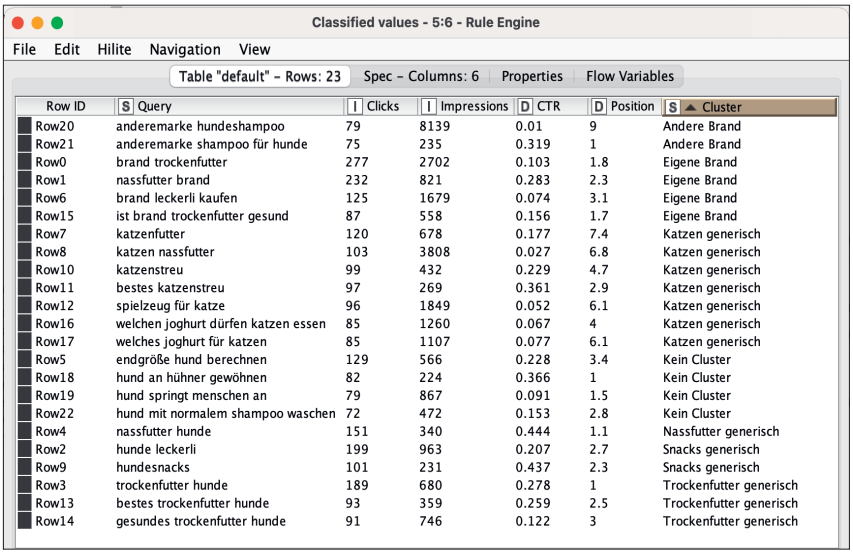


Abb. 14: Ergebnis: Eine neue Spalte mit den definierten Clustern wurde hinzugefügt.

Konfiguration der Rule Engine

Für dieses Beispiel sollen die URLs nicht gefiltert, sondern klassifiziert werden. Die Konfiguration des Rule-Engine-Knotens (Abbildung 10) funktioniert wie oben beschrieben, der Unterschied ist hierbei, dass ein Ergebniswert definiert wird, der dann in einer neuen Spalte ergänzt wird, wenn die Regel zutrifft. Das heißt, alle URLs mit dem Verzeichnis „/hunderassen/“ sollen als „Ratgeber Hunderassen“ klassifiziert werden etc. Mit „TRUE => Andere“ werden alle übrigen URLs, auf die die Regel nicht zutrifft, als „Andere“ klassifiziert (Abbildung 11).

Man hat dann die Möglichkeit, entweder eine neue Spalte hinzuzufügen und diese zu benennen oder eine der bestehenden Spalten durch die definierten Ergebniswerte zu ersetzen.

Auch hier könnte man die Bedingungen in einer Regel wieder beliebig kombinieren.

Ein weiterer Anwendungsfall ist das Clustern von Suchbegriffen. Diese könnte man beispielsweise nach Brand- und generischen Begriffen clustern, die Regeln hierfür könnten aussehen wie in Abbildung 13.

Fazit

Die Möglichkeiten zum Filtern, Clustern und Segmentieren sind genauso zahlreich wie die Bereiche und Anwendungsfälle, für die sich die Rule-Nodes eignen. Wie hilfreich KNIME bei der Analyse ist, wird allerdings erst deutlich, wenn die Rule-Nodes mit anderen Knoten kombiniert werden. Nach dem Clustern von Keywords oder dem Klassifizieren von URLs könnte man beispielsweise eine GroupBy-Node zur Gruppierung der Daten (vgl. Ausgabe 80) oder eine Joiner-Node (vgl. Ausgabe 81) anschließen, um das Ergebnis mit Daten aus einer anderen Tabelle zusammenzuführen.