Mario Fischer

SCREAMING FROG VERSION 20 – WOW!

Der Screaming Frog gehört sicherlich zu den bekanntesten und beliebtesten SEO-Tools weltweit. Seine Kernkompetenz war bisher, Websites zu crawlen und alle Inhalte so aufzubereiten, dass Fehler und Optimierungspotenzial sichtbar werden. Dann kam die Möglichkeit hinzu, über Datenschnittstellen (API) die gefundenen URLs mit weiteren Metriken anzureichern, zum Beispiel direkt aus der Google Search Console oder Google Analytics. Mit der neuen Version 20 ist dem Anbieter ein heftiger Paukenschlag gelungen. Neben einigen anderen neuen Funktionen gibt es jetzt – auch für Nichtprogrammierer – die Möglichkeit, KI-Schnittstellen direkt beim Crawl anzuzapfen. Sie möchten ChatGPT für jede URL einer Domain eine Frage (Prompt) stellen und die Antwort direkt in den Daten speichern lassen? Kein Problem. Und das Beste: Das kann wirklich jeder!

Neben der Möglichkeit, ein eigenes JavaScript einbinden zu können, hat sich auch beim Thema Mobile-Usability einiges getan. Für die Analyse von Begriffen lassen sich jetzt recht einfach sogenannte N-Gramm-Auswertungen erzeugen. Aber auch für das Umweltgewissen hat sich etwas verbessert. Für jede URL wird ab sofort auch der CO_2-Fußabdruck errechnet und ausgegeben.

Mit ChatGPT und Co. crawlen

Die Einbindung von Scripts und bereits fertigen Vorlagen macht es sehr einfach, einem Crawl nun mit einer Vielzahl von nützlichen Informationen anzureichern. Alles, was

TIPP

Seit über zwölf Jahren sind Tipps und Anwendungsbeispiele für den Screaming Frog in der Website Boosting zu finden. Gehen Sie einfach auf die Rubrik „Artikel", geben Sie Screaming Frog als Suchbegriff ein und klicken Sie gegebenenfalls noch „Online verfügbar" an. Hier finden Sie alles zum Einstieg, aber auch für fortgeschrittene Anwender sind zahlreiche Artikel vorhanden.

» www.websiteboosting.com/artikel.html

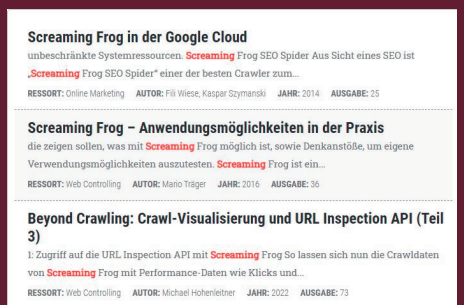

Foto: JanPietruszka / gettyimages.de

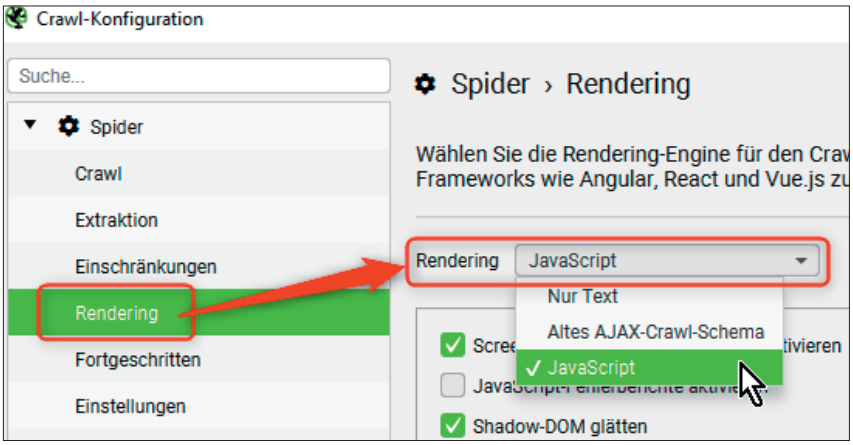

Abb. 1: Vorbereitungsarbeiten – JavaScript aktivieren

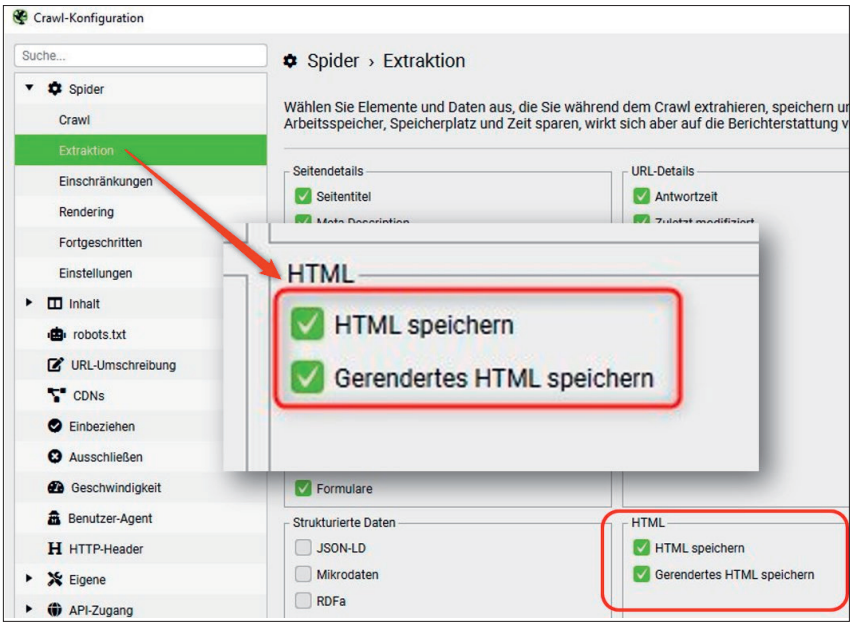

Abb. 2: Das gerenderte HTML muss auch gespeichert werden.

in ChatGPT oder Google Gemini per Prompt zusammen mit einer URL einzeln per Hand abgefragt werden kann, lässt sich jetzt automatisiert über ein Set oder auch alle Websites einer Domain oder Liste erledigen. Und nicht nur ein einzelner Prompt lässt sich nutzen – bis zu zehn solcher Prompts können parallel hinterlegt werden. Die Möglichkeiten, Fragen bezogen auf eine URL zu stellen, sind extrem vielfältig. Passt das Bild über der Falz zu dem Content? Gib mir eine kurze Zusammenfassung in zwei Sätzen über den Inhalt der Seite! Gibt es Rechtschreibfehler auf der Seite und welche? Oder noch extremer: Fasse mir das Thema dieser Seite mit nur einem Wort zusammen. Letzteres ermöglicht, zu prüfen, ob die thematische Architektur der Domain gut und trennscharf ausgerichtet ist (vergleiche dazu die Ergebnisse aus dem Google-Leak).

Ein weiteres nützliches Beispiel wäre „Prüfe, wie gut der Title der URL mit dem Content der Seite übereinstimmt, und mache einen Vorschlag für einen besseren Title. Formuliere den Title so, dass er zum Klicken anregt!“.

Um beispielsweise ChatGPT Antworten beim Crawl zu entlocken, gehen Sie wie folgt vor: Zunächst muss bei „Crawl-Konfiguration“ unter Spider und dann Rendering in dem Pull-down-Menü „JavaScript“ aktiviert werden (Abbildung 1).

Weiterhin muss in der Konfiguration unter „Extraktion“ rechts unten „HTML speichern“ und „Gerendertes HTML speichern“ aktiviert werden (Abbildung 2). Das war es auch schon mit den Einstellungen.

Bevor der Crawl gestartet wird, hinterlegt man dann die entsprechenden Scripts, die ChatGPT oder Google Gemini nutzen. Dazu gibt es eine bereits mitgelieferte Bibliothek, die den Einstieg sehr erleichtert. Unter Crawl-Konfiguration unten beim Punkt „Eigenes JavaScript“ (Abbildung 3, Ziffer 1) ruft man die Maske zur Eingabe der einzelnen Scripts auf. Über die Schaltfläche „Aus Bibliothek hinzufügen“ (Abbildung 3, Ziffer 2) findet man dann alle hinterlegten Scripts zur Übernahme. Oben (Abbildung 3, Ziffer 3) findet man die Möglichkeit, zwischen der Systembibliothek „System“ und „Benutzer“ zu wählen. Unter Benutzer findet man später eigene Scripts, die man nach einer individuellen Anpassung mit eigenem Namen abgelegt hat. Ein Klick auf den Namen für ein Script befördert es in eine der zehn Zeilen (Abbildung 3, Ziffer 4). Die Scripts sind mit sprechenden Namen und dem jeweiligen KI-Modell benannt, sodass die Auswahl einfach ist.

Wählen Sie zum Ausprobieren zum Beispiel „(ChatGPT) Intent of Page“

GUT ZU WISSEN

Ist sehr viel Content auf Websites, kann es passieren, dass man eine Fehlermeldung als Ergebnis bekommt. Das hat in der Regel damit zu tun, dass Google Gemini derzeit nur 3.071 Eingabetokens akzeptiert, während OpenAI mit ChatGPT 8.191 solcher Tokens akzeptiert. Die Google-KI wirft also viel früher die Füße in die Luft als ChatGPT. Umgekehrt liegen die Antworten von Gemini natürlich näher an dem, was die Suchmaschine für die Ranking-Algorithmen vermutet. Natürlich deckt sich das nicht eins zu eins, aber da die KIs aus dem gleichen Haus kommen, liegt die Vermutung einer höheren Übereinstimmung natürlich nahe.

ABFRAGEN KOSTEN GELD

Beachten Sie, dass jede Abfrage bei den gängigen KI-Tools per API sogenannte Tokens verbraucht und in Rechnung gestellt wird. Informationen dazu und entsprechende Auswertungen für Abfragen findet man auf den Seiten der Anbieter. Die Kosten sind allerdings vergleichsweise wirklich extrem niedrig und liegen in der Regel im Centbereich, sofern man nicht exzessiv und ständig crawlt. Zudem kann man das Budget sicherheitshalber auch deckeln, damit bestimmte Beträge nicht überschritten werden. Sofern Sie die APIs nur für die eigene Website nutzen und vernünftige Prompts verwenden, bleiben die monatlichen Kosten tatsächlich nicht selten unter einem Euro. Und selbst wenn der Betrag ansteigt, bekommt man Informationen auf keinem Weg günstiger und so schnell – vorausgesetzt, es steckt ein echtes und nutzbares Businessinteresse dahinter.

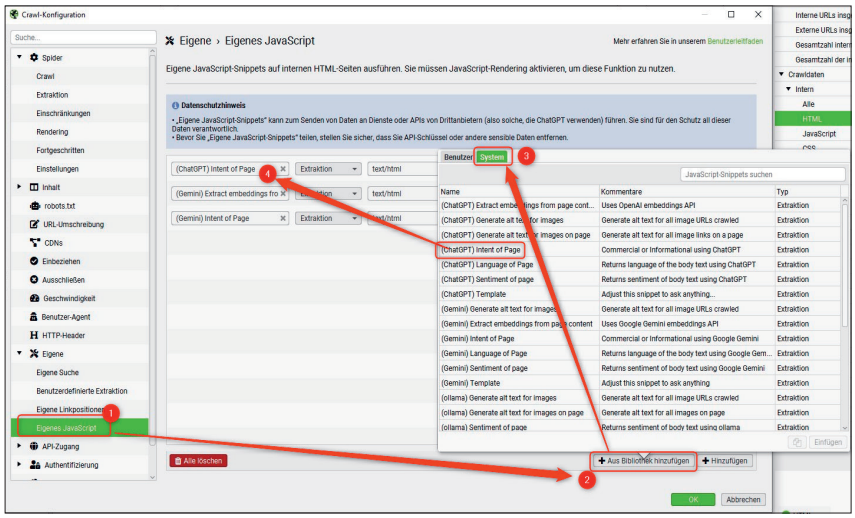

Abb. 3: Das Tool hält praktischerweise bereits ein fertiges Set von Abfragen bereit.

aus und öffnen Sie das so hinzugefügte Script per Button „JS“ am rechten Rand der Zeile mit dem Script-Eintrag. Anschließend öffnet sich der Script-Editor, wie in Abbildung 4 zu sehen ist. Damit man die Datenschnittstellen (API) auch außerhalb des Webinterfaces der KI-Anbieter abrufen kann, braucht man einen sogenannten API-Key und die jeweilige Bezahlversion des Tools. Für ChatGPT loggt man sich dazu auf der Website ein und klickt auf den angelegten Accountnamen beziehungsweise das Logo rechts oben. Unter „Your Profile“ in den Settings gibt es einen Punkt „User API keys“. Dort lässt sich via „Create new secret key“ ein solcher Schlüssel beziehungsweise eine längere Buchstaben-Zahlenfolge erzeugen. Den kopiert man wie in Abbildung 4 gezeigt in die Zeile mit „const OPENAI_API_KEY“ zwischen die beiden einfachen Anführungszeichen. Einen Key für Gemini von Google gibt es unter *ai.google.dev/gemini-api/docs/api-key?hl=de*.

Bevor man einen kompletten Crawl startet, testet man das Script am besten vorab. Das geht ebenfalls recht einfach. Schreiben Sie eine gültige URL in das Feld rechts unten (Abbildung 4) und klicken Sie auf „Test“. Das Script wird jetzt abgearbeitet und nach kurzer Zeit erscheint im rechten Fenster „JavaScript-Tester“ das Ergebnis. In diesem Fall nur die kurze Antwort auf die Frage nach dem Intent der URL: Informational (in Abbildung 4 gelb markiert). Testen Sie am besten immer alle verwendeten Scripts vorab, das erspart unnötige Crawls, sofern etwas nicht richtig funktionieren sollte. Ließe man einen Crawl jetzt laufen, fände man in den Ergebnistabellen des Screaming Frog eine weitere Spalte, in der für jede URL die Einschätzung für die Ausrichtung jeder URL abgelegt wurde.

Unter der Zeile, in die man den API-Key einträgt, findet man den Eintrag „const question“. Dort steht die Frage, die an das KI-Tool übergeben wird – genau genommen der Prompt. Diesen Prompt können Sie jederzeit überschreiben, wie in Abbildung 5, Ziffer 1 gezeigt wird. Im Beispiel wurde „Für welches Thema steht dieser Text? Gib bitte nur ein Wort als Ergebnis zurück: “ hinterlegt. Der Test für www.websiteboosting.com brachte als Ergebnis „Online-Marketing“ (gelb markiert, Abbildung 5, Ziffer 2). Möchte man dieses nun eigene Script später nochmals verwenden, speichert man es am besten ab (Abbildung 5, Ziffer 3) und vergibt einen sprechenden Namen (Abbildung 5, Ziffer 4). So kann man es jederzeit erneut aufrufen.

Ein weiteres Beispiel: Haben alle Ihre wichtigen Bilder auch wirklich sprechende Alt-Texte auf allen Webseiten? Nein? Der Aufwand, das nachträglich zu editieren, ist zu hoch? Jetzt gibt es keine Ausrede mehr. Der Prompt „Erstelle mir einen kurzen, aussagekräftigen Text für jedes Bild, das noch keinen Alt-Text hinterlegt hat“ wäre ein guter Ausgangspunkt, um sich den optimalen Alt-Texten anzunähern.

TIPP

Es lohnt sich durchaus, die beiden KIs von OpenAI (ChatGPT) und Google (Gemini) mit den gleichen Fragen gegeneinander antreten zu lassen. Die Ergebnisse sind nämlich bei Weitem nicht immer deckungsgleich! Für SEO-Fragen lohnt es sich vielleicht eher, die maschinellen Einschätzungen von Google zu verwenden, da die Suchmaschine natürlich bei ihren automatisierten Tools ähnlicher „urteilt“ als eine völlig anders funktionierende KI wie die von OpenAI.

B	C
(ChatGPT) Intent of Page 1	(Gemini) Intent of Page 1
Commercial	Informational
Informational	Informational
Commercial	Commercial
Commercial	Informational
Commercial	Commercial
Informational	Informational
Commercial	Commercial

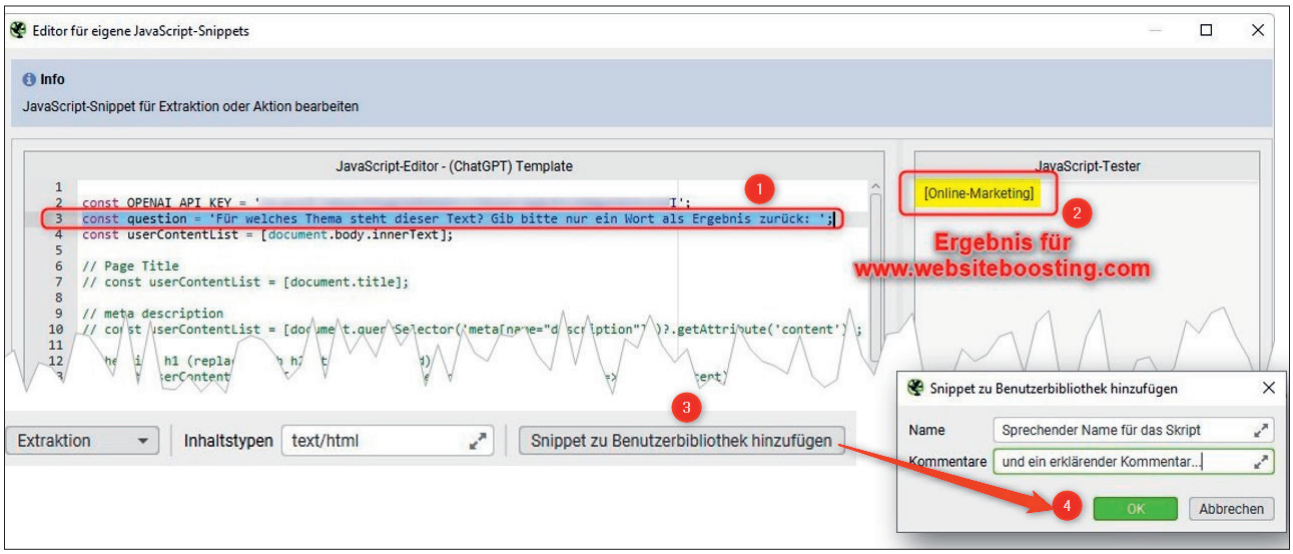

Abb. 5: Eigene Prompts hinterlegen? Einfach die Vorlage überschreiben!

TIPP

Falls Sie viele und/oder komplexe KI-Abfragen in einem Crawl machen und viele URLs abfragen, kann es notwendig werden, die Crawl-Geschwindigkeit des Screaming Frog zu reduzieren, damit sich die KIs im Hintergrund nicht verschlucken beziehungsweise die Anfragen nicht zu schnell aufeinanderfolgen und einen Überlauf erzeugen. Eine Anpassung lässt sich unter Crawl-Konfiguration „Geschwindigkeit" vornehmen. Dort lässt sich auch die gleichzeitige Anzahl („Max. Threads") an URLs reduzieren.

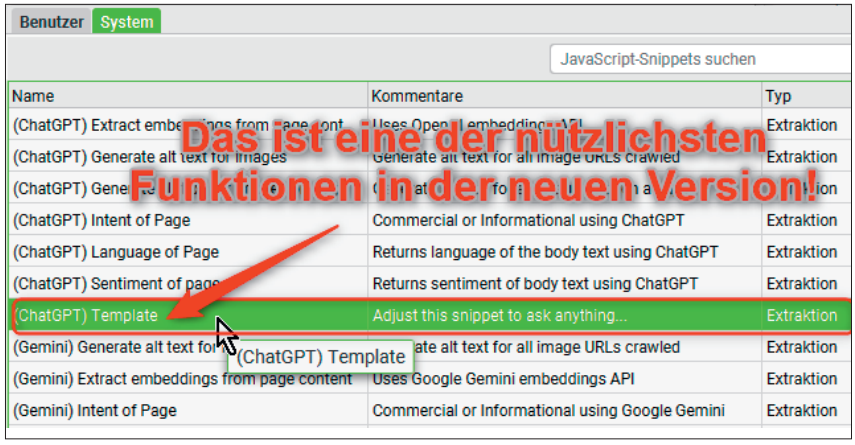

Abb. 6: Nutzen Sie die leere Vorlage für eigene Abfragen

Mittlerweile gibt es sogar schon nützliche Scripts, die in Blogs oder Foren von SEO-Experten zur Verfügung gestellt werden. Die lassen sich ganz einfach per Copy-and-paste in die eigene Bibliothek aufnehmen und so entsteht im Lauf der Zeit eine nützliche Sammlung. Die besten Scripts bindet der Anbieter von Screaming Frog mit jeder neuen Unterversion direkt ein, sodass die Systembibliothek ständig erweitert wird. Es lohnt sich also, nach jedem Update hier nachzusehen, was neu verfügbar ist.

Wer in JavaScript fit ist, kann auch eigene Scripts schreiben und einbinden. So lässt sich prinzipiell unter anderem auch jede Aktion (Klicks, Anmeldungen, Scrollen, Mouseover etc.) auf einer URL ausführen. Einige solcher Aktions-Scripts sind bereits hinterlegt. Das macht auch Informationen einer Seite zugänglich, die bisher beim Crawlen nicht erfasst werden konnten, weil sie zum Beispiel nur bei einem Mouseover dynamisch angezeigt werden.

Wortspiele: N-Gramm-Analysen

Mit N-Grammen zerlegt man Begriffe in Fragmente und hilft, Metriken zu Wortverwendungen zu erstellen. Im Screaming Frog werden sie zur Unterscheidung von Ein- oder Mehrwortphrasen benutzt. Damit lassen sich nach einem Crawl Worthäufigkeiten und Muster analysieren. Eine solche Analyse auf Wort- beziehungsweise Begriffsebene kann sehr nützlich zur Modellierung oder Prüfung einer thematischen Content-Struktur auf semantische Relevanz einzelner Seiten sein. Aber auch einfache Textanalysen für die klassische On-Page-Optimierung lassen sich so durchführen.

Welches wichtige Wort fehlt im Title einer URL oder in verweisenden Linkankern? Welche gleichen Ankertexte sind mit verschiedenen Zielseiten verlinkt? Über Abgleiche mithilfe von N-Grammen lassen sich aber auch Keyword-Lücken finden, indem man die von der Google Search Console abfragbaren Suchworte einbezieht. Verwenden unterschiedliche Seiten bestimmte Keywords zu ähnlich und erzeugen damit die sogenannte Keyword-Kannibalisierung? Welche Keywords verwenden die Mitbewerber im Vergleich zu den eigenen?

CO_2-Fußabdruck und Bewertung

Wer ein schlechtes Umweltgewissen hat, kann sich mit der neuen Version für jede gecrawlte URL den hochgerechneten CO_2-Ausstoß in Milligramm sowie eine einfache Bewertung mit Buchstaben ausgeben lassen. Ideal wäre es, überall den Buchstaben A ausgewiesen zu bekommen. Die meisten Websites werden jedoch wahrscheinlich noch immer deutlich mehr Seiten mit einem E oder F ausgewiesen bekommen. Wer konsequent an der Verringerung des Seitenumfangs und der verlinkten Bilder arbeitet, kann sich über diese Daten recht einfach Vorher-nachher-Abbildungen für Präsentationen erstellen. Summiert man die einzelnen Milligramm über die jährlichen Impressions jeder URL auf, entstehen schnell viele Kilogramm oder gar Tonnen, die man eingespart hat. Tue Gutes und rede darüber. Mit den entsprechenden Daten geht das nun leichter.

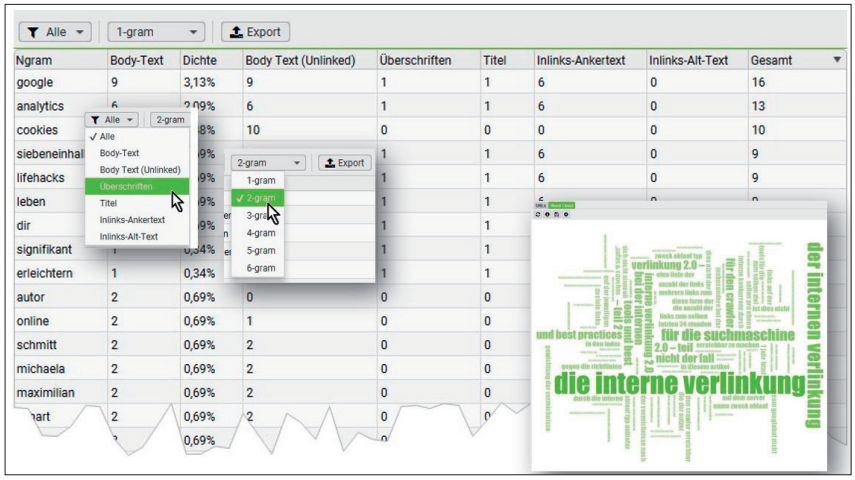

Abb. 7: N-Gramm-Analysen bringen gut nutzbare Optimierungsansätze auf Begriffsebene.

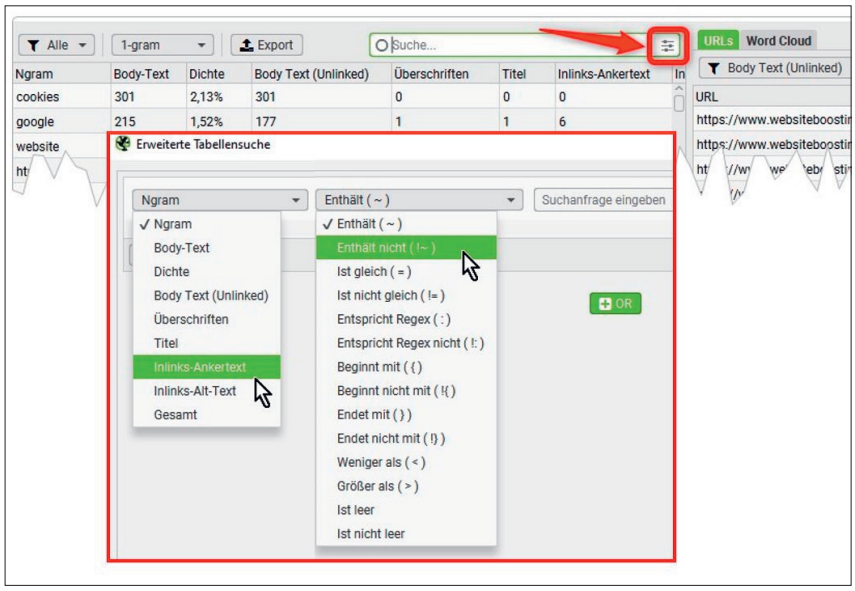

Abb. 8: Etwas versteckt im Suchfeld findet man umfassende Such- und Filtermöglichkeiten für Begriffe und Begriffsteile. Hier lassen sich auch komplexe Abfragen zusammenstellen (zum Beispiel über RegEx).

Wortvektoren: Embeddings

Den theoretischen Hintergrund von sogenannten Embeddings an dieser Stelle zu erklären, würde bei Weitem den Rahmen sprengen. Wer das Thema in der Website Boosting aufmerksamer verfolgt hat, wird sich über die Möglichkeit der einfachen Generierung von Embeddings freuen. Grob gesagt werden die Wörter einer URL in mathematische Vektoren umgewandelt. Über die sogenannte Cosinus-Ähnlichkeit lassen sich dann sowohl mehrere Seiten thematisch vergleichen als auch Sucheingaben bei Suchmaschinen mit URLs auf eine möglichst hohe Übereinstimmung prüfen. Je ähnlicher solche (komplexen) Wortvektoren sind, desto ähnlicher ist der Inhalt. Um an die begehrten Vektoren beziehungsweise Embeddings zu kommen, musste man bisher programmieren können. Jetzt geht via Screaming Frog eine einfache Abfrage über ein vorgegebenes Script (Abbildung 11). Prinzipiell lassen sich Embeddings auch via ChatGPT erzeugen, für SEO-Zwecke bei Google empfiehlt es sich allerdings, besser Gemini zu nutzen (siehe Erklärung am Rand). Möchte man Keywords beziehungs-

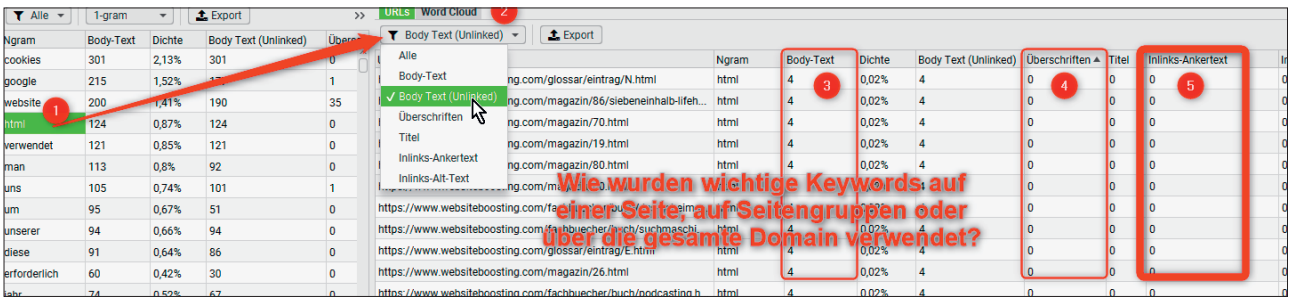

Abb. 9: Wo werden wichtige Begriffe verwendet? Wo gibt es Keyword-Lücken?

Übertragen	Total Transferred	CO_2 (mg)	CO2-Bewertung
14.0 KB	2.4 MB	878,116	F
12.6 KB	1.7 MB	646,705	D
17.0 KB	1.0 MB	386,003	C
17.5 KB	1.9 MB	705,107	E
13.2 KB	703.1 KB	255,235	B
13.9 KB	1.6 MB	603,281	D
13.8 KB	1.0 MB	382,969	C
14.4 KB	2.2 MB	815,270	E
13.0 KB	1.3 MB	490,389	C
17.6 KB	1.7 MB	632,981	D

Abb. 10: Pro URL wird der CO_2 Ausstoß hochgerechnet und bewertet

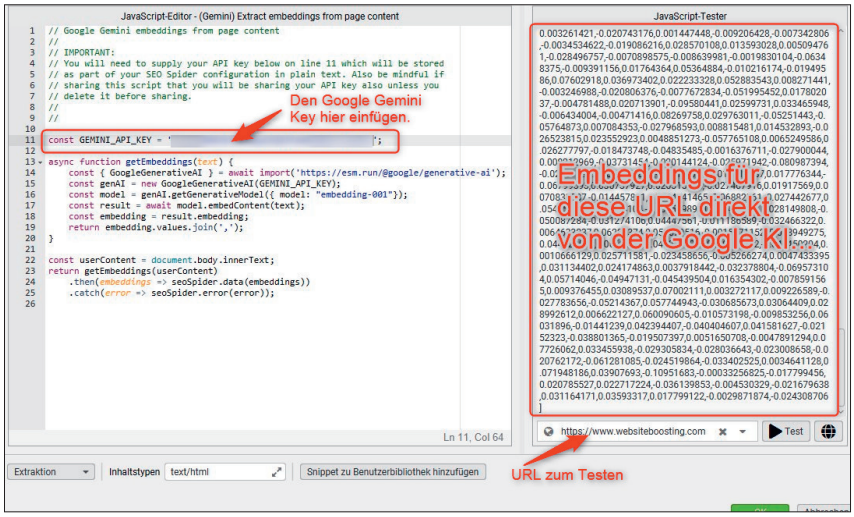

Abb. 11: Embeddings – Content-Vektorisierung über Google Gemini

weise Suchphrasen per Embeddings mit den Embeddings von Seitencontent auf Ähnlichkeit prüfen, sollte man in jedem Fall das gleiche System (Google oder OpenAI) für die Keywords und die URLs verwenden, da solche Systeme unterschiedlich arbeiten. Ein Vergleich ist nur sinnvoll, wenn das gleiche Vektorisierungsmodell genutzt wurde.

Mobile Fitness

Zur einfachen Prüfung von Problemen bei der Darstellung von Websites auf Smartphones gibt es in der neuen Version einen neuen Tab, in dem man die von Google Lighthouse gelieferten Metriken anzeigen lassen kann. Angezeigt werden Fehler wie „Viewport Not Set“, „Target Size“, „Content Not Sized Correctly“, „Illegible Font Size“ oder „Mobile Alternate Link“. Die vorgefertigten Berichte wurden entsprechend ergänzt. Um die Daten von Google zu übertragen, muss man über die Konfiguration beziehungsweise den API-Zugang unter „PageSpeed Insights“ bei dem Pull-down-Menü „Quelle“ „Remote“ auswählen und einen Schlüssel für den Zugriff eintragen. Den Schlüssel erhält man von Google kostenlos unter einfach.st/insightkey.

Alternativ kann man unter „Quelle“ auch „Lokal“ auswählen. Dann verbindet sich der Screaming Frog mit dem Chrome-Browser auf dem Computer und nutzt dessen Schnittstelle. Dies verlangsamt den Prozess allerdings deutlich. Eine gute Anleitung für

WEBSITE BOOSTING RELOADED

Zum Nachlesen zur Bedeutung von Wortvektoren und Embeddings:

Ein Vektor sagt mehr als tausend Wörter

AlphaGo, RankBrain und Google AI – Machine-Learning und künstliche Intelligenz sind für Suchmaschinenoptimierer inzwischen geflügelte Worte. Was allerdings Machine-Learning ist und wie Suchmaschinen es einsetzen, ist wenig geläufig. Grund genug für Stefan

Stefan Fischerländer, Ausgabe 45: einfach.st/vektor1

Googles BERT-Update – Wie der NLP-Algorithmus die Suche verändert

Alle SEOs reden aktuell über BERT, doch kaum jemand kann sich erklären, wieso Google vom größten Update seit fünf Jahren und

Kai Spriestersbach, Ausgabe 60: einfach.st/bert1

ENTITÄTEN-BASIERTE SUCHE:
SO FUNKTIONIERT DER GOOGLE KNOWLEDGE GRAPH

Die Einführung des Knowledge Graph im Jahr 2012 und des Hummingbird-Algorithmus im Jahr 2013 waren neben Panda

Olaf Kopp, Ausgabe 57: einfach.st/kgraph1

Mobile-Audits auf Englisch findet man direkt auf der Website von Screaming Frog unter *einfach.st/frogmobile*.

Fazit

Die Möglichkeit, eigene Scripts und somit auch KIs für das Crawling einbinden zu können, stellt sicher einen Quantensprung für den Screaming Frog dar. Ab sofort ist man nicht mehr darauf angewiesen, die durchaus umfangreichen fest implementierten Funktionen zu nutzen, sondern kann beim Abruf von Seiten eigene Daten und Antworten auf Fragen integrieren. Auch Inhalte, die eine Benutzeraktion erfordert haben, lassen sich nun erfassen. Neben dieser Innovation erscheinen die vielen anderen Neuerungen in Version 20 fast nebensächlich – auch wenn sie sicherlich eines zweiten Blicks wert sind.

Eine Jahreslizenz für die neue Version liegt derzeit übrigens bei 239 Euro. Wer bereits eine gültige Lizenz hat, bekommt das Update kostenlos.

Rebecca Schwarz

KNIME für E-Commerce-SEO: Out-of-Stock-Produkte in redaktionellen Inhalten identifizieren

Alle Betreiber von Online-Shops kennen es: Produkte sind hin und wieder nicht mehr verfügbar. Während das auf Übersichtsseiten in Shop-Umgebungen in der Regel recht schnell identifizierbar und auswertbar ist, kann die Suche nach dieser Art von Produkt in redaktionell erstellten Inhalten sehr mühsam sein.
Um Nutzer, die über Ratgeber- und Magazininhalte auf Online-Shops einsteigen, nicht zu frustrieren und ihnen nur verfügbare Produkte zu verlinken, gibt es eine einfache Lösung: die schnelle Analyse mit KNIME, ob in redaktionellen Inhalten Out-of-Stock-Produkte verlinkt sind. Wie das funktioniert und welche Voraussetzungen notwendig sind, wird im nachfolgenden Beitrag Schritt für Schritt beschrieben.

Die Identifizierung von Verlinkungen auf Out-of-Stock-Produkte eines Shops kann sehr mühsam sein und führt in einer händischen Prüfung schnell zur Verzweiflung. Doch es gibt eine einfache und effiziente Lösung dafür: die Analyse mit KNIME. Schritt für Schritt wird hier gezeigt, wie sich diese lästigen Verlinkungen identifizieren lassen. Der Workflow kann dazu individuell auf Seitenbereiche eingeschränkt werden, um nur Verlinkungen auf Produkte zu analysieren, die in redaktionell erstellten Bereichen liegen. Die folgenden Fragestellungen können so gelöst werden:

- Werden in bestimmten Seitenbereichen Produkte verlinkt, die nicht verfügbar sind?
- Mit welchen Ankertexten und an welcher Stelle auf einer Seite sind die Verlinkungen zu finden?
- Wie häufig werden nicht verfügbare Produkte verlinkt?

Warum wird KNIME verwendet?

Mit der kostenfreien Open-Source-Software KNIME lassen sich viele SEO-Analysen durchführen. Der Vorteil ist, dass zu jedem Zeitpunkt überwacht werden kann, wie sich die Daten verändern. Denn jeder Workflow besteht aus einer Aneinanderreihung von Knoten, durch die bereitgestellte Daten fließen. Im Gegensatz zu Excel sind einmal erstellte Workflows auch jederzeit wiederholbar. Denn liegen Daten immer wieder in gleicher Form vor, können diese jederzeit durch bereits erstellte Workflows fließen. Das macht das Tool vor allem für Regeltätigkeiten in der SEO zu einem mächtigen Werkzeug! Zusätzlich ist die Software für Anfänger geeignet und erfordert keine Programmierkenntnisse!

DIE AUTORIN

Rebecca Schwarz ist SEO-Consultant bei der get:traction GmbH. Ihr Arbeitsalltag dreht sich um die Konzeption von SEO-Strategien und die Unterstützung von Kunden in der redaktionellen SEO. Um größere Datenmengen effizient zu verarbeiten und um bei wiederkehrenden SEO-Tasks Zeit zu sparen, nutzt sie die Open-Source-Software KNIME.

ERLÄUTERUNG FÜR DIE KONFIGURATION IM SCREAMING FROG:

Zur Konfiguration der Custom Extraction ist Folgendes notwendig: Zunächst wird über *Configuration » Custom » Custom Extraction* angesteuert (Abbildung 1, Ziffer 1). Dann wird über Add ein neuer Eintrag angelegt (Abbildung 1, Ziffer 2). Damit der KNIME-Workflow später nicht mehr angepasst werden muss, kann jetzt der folgende Eintrag übernommen werden (Abbildung 1, Ziffer 3):
1. Feld: Produkt-Verfügbarkeit
2. Feld: Regex
3. Feld: (InStock|OutOfStock)

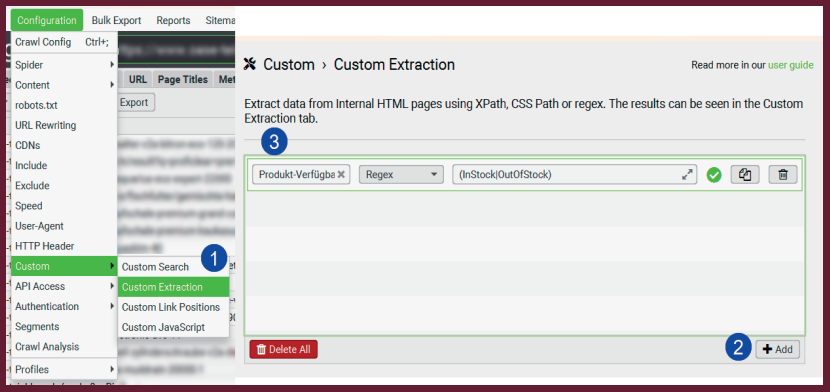

Abb. 1: Konfiguration der Custom Extraction im Screaming Frog

Prämisse zur Durchführung der Analyse

Auch wenn Online-Shops immer wieder ihre Eigenheiten haben und Fehlerquellen unter Umständen sehr individuell sein können, sind sie im Allgemeinen ähnlich gestrickt. Für die vorgestellte Analyse werden deshalb die folgenden Annahmen getroffen:

» Alle Produktseiten des Online-Shops sind mit semantischen Daten im Vokabular von schema.org ausgezeichnet und stehen im Format JSON-LD zur Verfügung.

» Die semantische Auszeichnung der Produktdetailseiten entspricht den Vorgaben von Google und enthält den Tag availability mit der Ausprägung ist verfügbar = *https://schema.org/InStock* oder ist nicht verfügbar = *https://schema.org/OutOfStock*.

Für den ersten Schritt der Analyse wird zunächst das Crawling-Tool Screaming Frog benötigt, um an die Informationen der semantischen Auszeichnung zu gelangen. Hierzu wird die gesamte Website mit einer sogenannten Custom Extraction Anweisung gecrawlt. Ziel ist es so, den Status der Verfügbarkeit aus den Produktinformationen zu extrahieren. Enthält der Quellcode OutOfStock oder InStock, wird dies extrahiert. Unter der Annahme, dass alle Produktdetailseiten einen Status der Verfügbarkeit enthalten, lässt sich so entweder OutOfStock oder InStock extrahieren.

Ist die Custom Extraction-Anweisung konfiguriert, wird der Crawl für die entsprechende Domain gestartet. Nach erfolgreichem Durchlauf wird der Crawl exportiert. Für die weitere Analyse sind die folgenden Exports notwendig:

1. *internal_all-Export:* Der Export kann einfach über den Export-Button als CSV-Datei exportiert werden.
2. *all_inlinks-Export:* Dieser Export wird im Screaming Frog über den Reiter *Bulk-Export ⟶ Links ⟶ All Inlinks* ausgewählt und wird ebenfalls als CSV-Datei exportiert.

Nun sind alle Vorbereitungen getroffen und die Arbeit mit dem Workflow in KNIME kann beginnen.

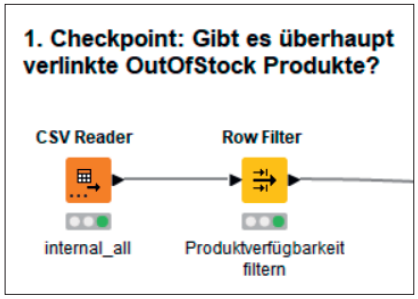

Abb. 2: Einlesen und Filtern des internal_all-Exports

Checkpoint eins: Existieren im Shop nicht verfügbare Produkte?

Ein vermutlich seltener Fall, aber im ersten Schritt des Workflows wird geprüft, ob im Shop überhaupt Produkte vorliegen, die mit OutOfStock gekennzeichnet sind. Dazu wird der CSV-Export internal_all über den Knoten CSV READER eingelesen. Daran angebunden wird ein ROW FILTER (Abbildung 2). Dieser Knoten filtert im

INFO

Für KNIME-Einsteiger: Die einzelnen Knoten werden aus dem Node Repository in die grafische Oberfläche gezogen und dann nach und nach über die Pfeilsymbole miteinander verbunden. Damit die Knoten Daten an den nächsten Knoten übertragen können, müssen die Knoten mit Rechtsklick ⟶ „EXECUTE" ausgeführt werden. Wechselt die Ampel-Anzeige unterhalb des Knotens auf Grün, können die Daten problemlos weitergegeben werden. Leuchtet nach Ausführung eines Knotens ein rotes Warnsymbol auf, gibt es Probleme mit der Ausführung und der Workflow stoppt. Um zu erfahren, wodurch das Problem verursacht wird, kann einfach über das Symbol gehovert werden, um die Fehlermeldung angezeigt zu bekommen.

TIPP

Um herauszufinden, wie die Daten zum aktuellen Zeitpunkt im Workflow tabellarisch aussehen, kann jederzeit mit Rechtsklick auf den zuvor ausgeführten Knoten geklickt werden und über „File Table" die Tabellenvorschau angezeigt werden.

Allgemeinen Zeilen eines Datensets anhand eines definierten Werts.

Die Konfiguration des Knotens ROW FILTER läuft wie folgt ab: Im ersten Feld wird die Spalte „Produkt-Verfügbarkeit 1" ausgewählt, die gefiltert werden soll (Abbildung 3, Ziffer 1). Im zweiten Feld wird dann der Wert für die Filterung eingegeben (Abbildung 3, Ziffer 2). Auch die Formulierung von Wildcards und regulären Ausdrücken ist möglich. Damit die Filterung einwandfrei funktioniert, wird eine Formulierung eingegeben und anschließend mit dem Häkchen ausgewählt, um welche Art von Pattern es sich handelt. In diesem Fall wird auf den genauen Wert (= case sensitive match) „OutOfStock" gefiltert, da der Wert durch die Custom Extraction im Screaming Frog so in der entsprechenden Spalte vorzufinden ist, wenn ein Produkt nicht verfügbar ist.

Der ROW FILTER wird nun über Rechtsklick → „EXECUTE" ausgeführt. Die Ergebnistabelle enthält nun nur noch Zeilen, die in der Spalte „Produkt-Verfügbarkeit 1" den Wert „OutOfStock" enthalten. Ist die Tabelle leer? Dann befinden sich im Shop keine Produkte, die den Status „OutOfStock" aufweisen. Zeigt die Tabelle URLs an, geht es nun in den zweiten Checkpoint.

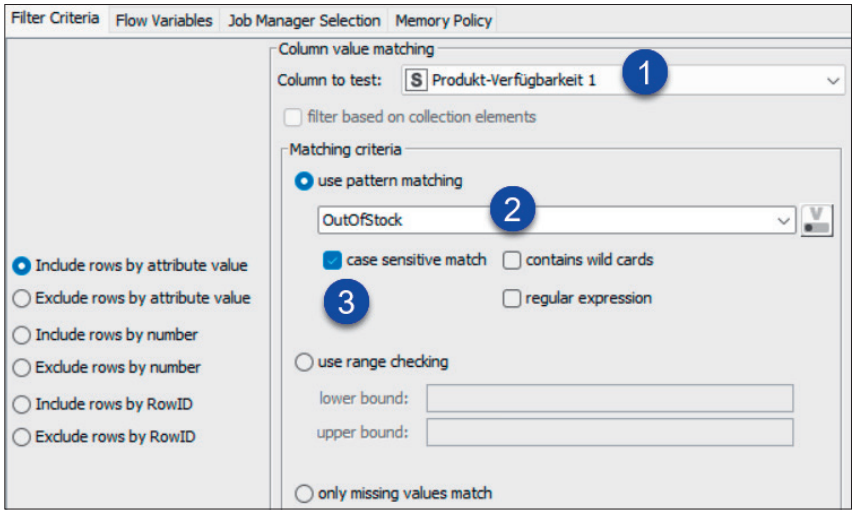

Abb. 3: Konfiguration des Knotens ROW FILTER

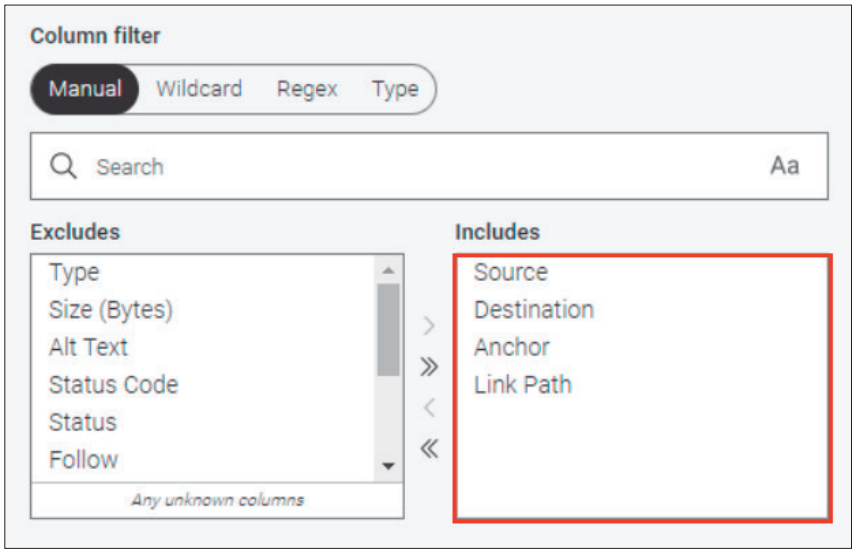

Abb. 4: Konfiguration des Knotens COLUMN FILTER

Checkpoint zwei: Auf welchen Seiten werden nicht verfügbare Produkte verlinkt?

Für die Analyse, wo auf der Website Out-of-Stock-Produkte verlinkt sind, muss die interne Verlinkung betrachtet werden. Dazu wird nun der all_inlinks-Export aus dem Screaming Frog verwendet. Die folgenden Spalten mit der Standardbenennung des Exports sind relevant für die Analyse:

- Spalte „Source": Sie zeigt die Quelle, aus der eine URL verlinkt wird.
- Spalte „Destination": Sie zeigt das Ziel, auf das von der „Source" aus verlinkt wird.
- Spalte „Anchor": Sie zeigt den Ankertext, mit dem die Ziel-URL verlinkt wird.
- Spalte „Link Path": Sie zeigt mithilfe eines XPaths an, an welcher Stelle der Seite sich der Link befindet.

Zur Arbeit mit diesen Informationen wird der all_inlinks-Export zunächst via CSV READER in KNIME eingelesen und wie gewohnt ausgeführt. Um nun nur mit den oben genannten Spalten weiterzuarbeiten, werden mit dem Knoten COLUMN FILTER nur die vier oben genannten Spalten über Include beibehalten (Abbildung 4).

In nächsten Schritt wird nun ausgewählt, in welchem Seitenbereich die Verlinkung von Out-of-Stock-Produkten analysiert werden soll. Dazu wird ein weiterer ROW FILTER definiert. In diesem wird die Spalte „Source" auf ein bestimmtes URL-Muster hin gefiltert. Soll beispielsweise geschaut werden, wo Out-of-Stock-Produkte im Verzeichnis /ratgeber/ verlinkt sind, wird im matching criteria „*ratgeber*" (Abbildung 5, Ziffer 2) mit der Auswahl „contains wild cards" (Abbildung 5, Ziffer 3)

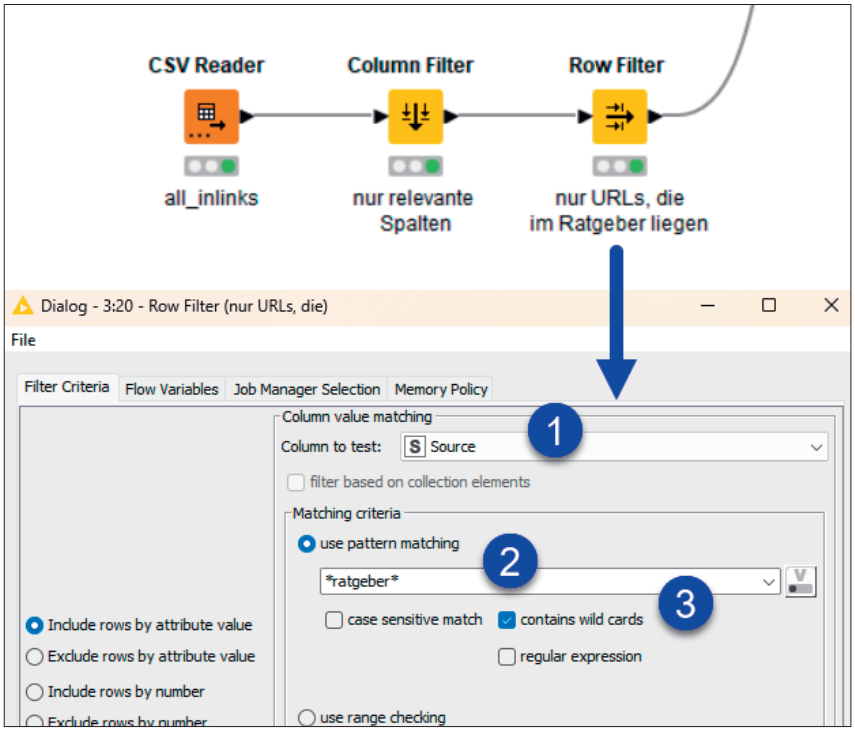

Abb. 5: Konfiguration des Seitenbereichs

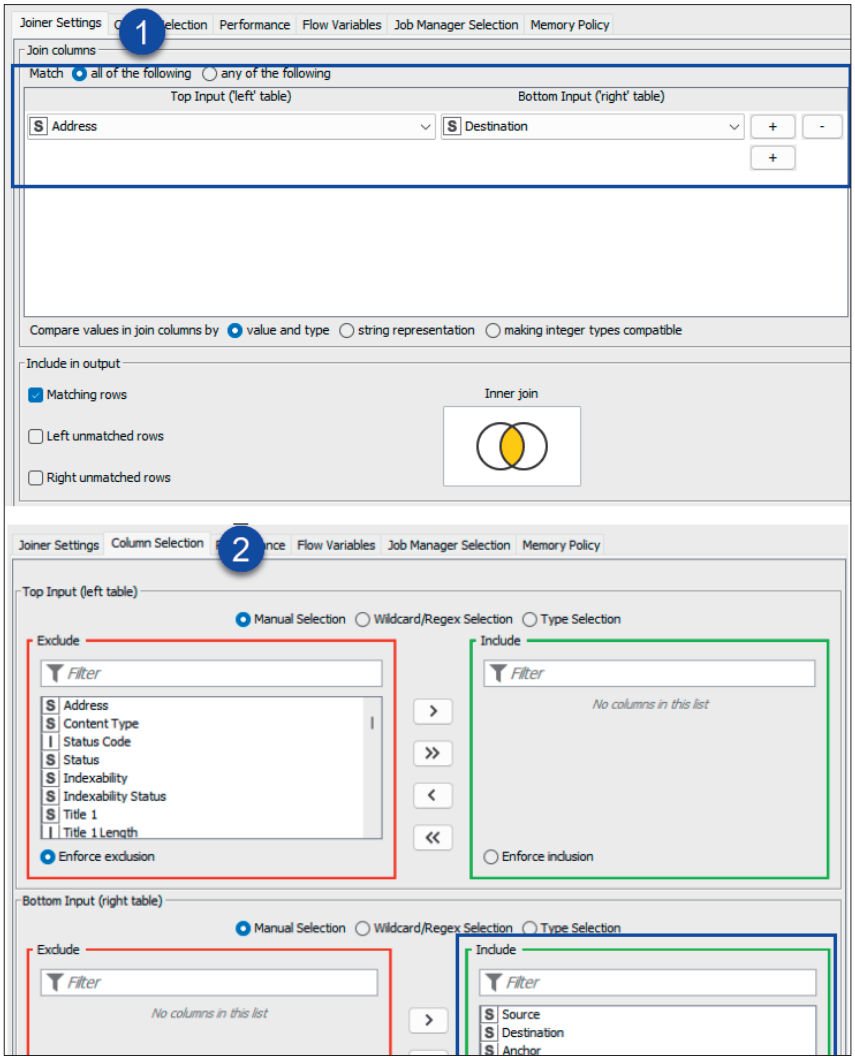

Abb. 7: Konfiguration des Knotens JOINER

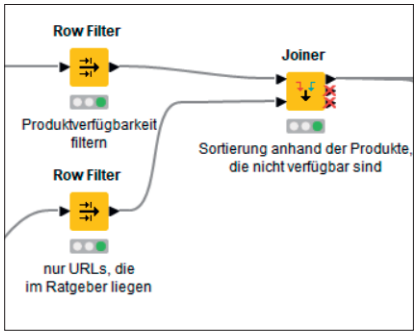

Abb. 6: Zusammenführung der Daten im JOINER

ausgewählt. Je nach Verzeichnisstruktur und Benennung muss das matching criteria entsprechend angepasst werden.

Zusammenführung der Daten

Im nächsten Schritt werden nun die Daten aus Checkpoint eins „Existieren Out-of-Stock-Produkte“ mit den Daten der internen Verlinkung zusammengeführt. Dieser Schritt ist notwendig, um nur noch Daten zu erhalten, die zwei Bedingungen erfüllen:

1. Produkte, die mit „OutOfStock“ gekennzeichnet sind
2. Produkte, die in einem definierten Seitenbereich (hier: im Ratgeber) verlinkt sind

Das Zusammenführen der Daten funktioniert in KNIME mit dem JOINER, dem Äquivalent zum SVERWEIS in Excel. Der JOINER hat zwei Ports an der linken Seite, an die die beiden ROW FILTER angebunden werden (Abbildung 6).

Im JOINER selbst wird nun noch ausgewählt, welche Spalten zum Matching miteinander verglichen werden sollen und welche Art von Matching ausgeführt werden soll. Für den Anwendungsfall wird die Spalte „address“ mit der Spalte „Destination“ verglichen.

Zum besseren Verständnis:

» In den Daten, die am oberen Port eingelesen werden, befinden sich die Produkt-URLs, die die Bezeichnung „OutOfStock“ beinhalten. Die Spalte dieser URLs heißt „address“.

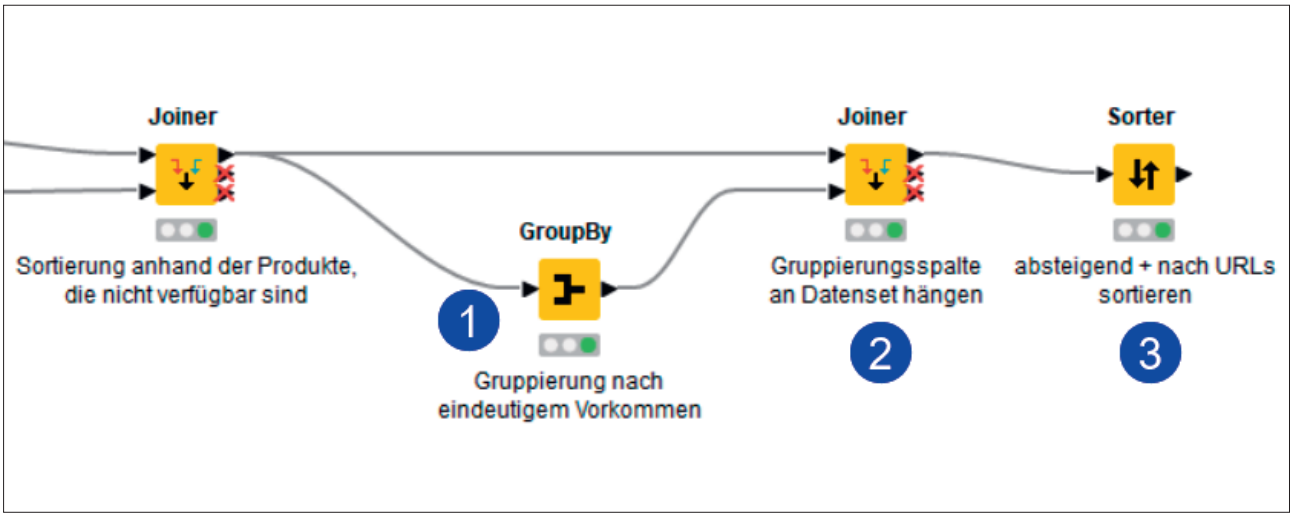

Abb. 8: Ordnung durch drei weitere Knoten schaffen

» Im unteren Port werden die URLs eingelesen, die im Seitenbereich / ratgeber/ verlinkt werden. Die Spalte der URLs heißt „Destination“.

Weil nur mit Daten weitergearbeitet werden soll, die beide Bedingungen – Produkt ist nicht verfügbar und es wird im Ratgeber-Bereich verlinkt – erfüllen, wird im JOINER matching rows ausgewählt, was einem Inner Join entspricht (Abbildung 7, Ziffer 1). Zur besseren Übersicht werden innerhalb des Knotens JOINER im Reiter Column Selection zusätzlich nur die relevanten Spalten übernommen und alle anderen verworfen (Abbildung 7, Ziffer 2). Durch die Ausführung des Knotens und Rechtklick → „*Join result*“ ergibt sich nun bereits das Ergebnis der Auswertung. Nun wird noch Ordnung geschaffen, um später eine bessere Übersicht zu erhalten.

Ordnung schaffen und URLs priorisieren

Zum besseren Überblick über die Mengengerüste der verlinkten Produktdetailseiten wird der Workflow nun noch um drei Knoten erweitert (Abbildung 8).

Zunächst wird mit dem Knoten GROUP BY (Abbildung 8, Ziffer 1) gezählt, wie häufig jede einzelne Produktseite im Seitenbereich verlinkt ist. Die daraus resultierende Ansicht entspricht einer Pivot-Tabelle in Excel.

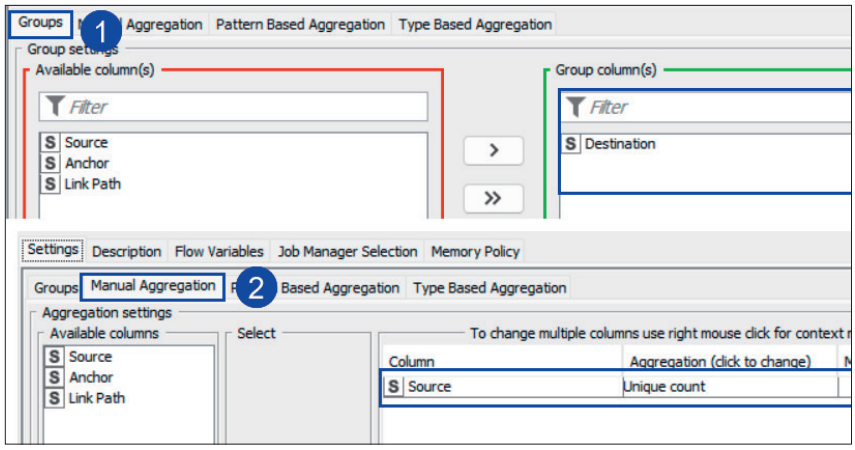

Abb. 9: Konfiguration des Knotens GROUP BY

Row ID	Destination	Unique count(Source)
Row0	https://www.	1
Row1	https://www.	2
Row2	https://www.	1
Row3	https://www.	2
Row4	https://www.	1
Row5	https://www.	1
Row6	https://www.	1
Row7	https://www.	1
Row8	https://www.	1

Abb. 10: Ergebnistabelle des Knotens GROUP BY

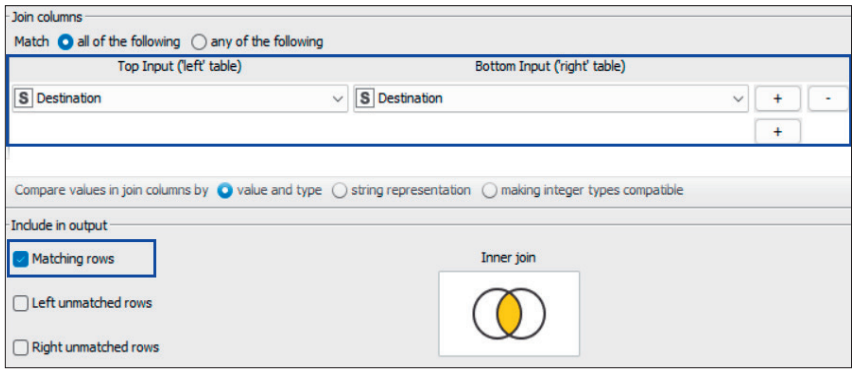

Abb. 11: Konfiguration des zweiten JOINER-Knotens

WICHTIGER HINWEIS

In der Vorschautabelle können die Daten zwar auf- und absteigend sortiert werden, dies ist jedoch nur eine Vorschau und hat keine Auswirkung auf die weitere Verarbeitung der Daten. Für einen späteren Export müssen die Daten immer über den SORTER sortiert werden, damit sich die Veränderung auf das Datenset auswirkt.

Source	Destination	Anchor	Link Path	Unique Count
example.org/ratgeber/artikel2	example.org/produkt/1235	Produkt 1235	//body/div[@id='content']/div[4]/div	2
example.org/ratgeber/artikel3	example.org/produkt/1235	Produkt 1235	//body/div[@id='content']/div[4]/div	2
example.org/ratgeber/artikel1	example.org/produkt/1234	Produkt 1234	//body/div[@id='content']/div[4]/div	1

Abb. 12: Ergebnistabelle der Auswertung

Dazu wird im Reiter groups die Spalte „Destination“ ausgewählt (Abbildung 9, Ziffer 1). Im Reiter Manual aggregation wird nun noch bestimmt, welche Werte zum Zählen verwendet werden sollen und wie gezählt wird. Dazu wird die Spalte „Source“ mit der Aggregation „unique count“ ausgewählt (Abbildung 9, Ziffer 2).

Daraus ergibt sich eine Tabelle, die die eindeutigen Verlinkungsziele und die Anzahl des Vorkommens im Seitenbereich anzeigt (Abbildung 10).

Damit die neu gewonnenen Informationen der Häufigkeit des Vorkommens mit dem vorherigen Datenset verbunden werden können, wird an dieser Stelle ein weiterer Knoten JOINER angebunden (Abbildung 8, Ziffer 2). In diesem zweiten JOINER-Knoten wird dazu jeweils die Spalte „Destination“ gewählt, da diese in beiden Datensätzen vorliegt, und mit einem erneuten Inner Join gematcht.

Der letzte Knoten ist der SORTER, mit dem die Daten absteigend nach Häufigkeit des Vorkommens und nach URL sortiert werden (Abbildung 8, Ziffer 3).

Ergebnis der Analyse

Das Ergebnis der Analyse ist eine Tabelle, die zeigt, welche Produkte (Spalte „Destination“) nicht verfügbar sind und wo diese verlinkt werden (Spalte „Source“). Zusätzlich enthält die Tabelle die Information über den zugehörigen Linktext, den Linkpfad auf der Seite und die Häufigkeit des Vorkommens.

Gerade bei einer größeren Datenmenge kann mit dieser Angabe besser priorisiert werden, welche Produkte am häufigsten verlinkt werden und auf welchen Seiten die Überarbeitung der Verlinkungen somit mehr Impact hat (Abbildung 12).

Fazit

Der Workflow zeigt, wie Analysen schnell und einfach über KNIME durchgeführt werden können, die ohne eine solche Analysesoftware deutlich zeitaufwendiger wären. Die Konfiguration der Knoten kann dabei individuell angepasst und jederzeit auf andere Seitenbereiche angewendet werden. Der fertige Workflow ist wie immer im KNIME-Hub zusammen mit den Workflows der letzten drei Ausgaben unter einfach.st/knime86 verfügbar und ist per einfachem Drag-and-drop in die eigene KNIME-Oberfläche einlesbar.

Und noch ein abschließender Hinweis: Die vorgestellten Workflows haben nicht nur das Ziel, die Anwendungsfälle zu erklären, sondern sollen auch die Methodenkompetenz mit KNIME vermitteln. Wer anfängt, sich mit KNIME zu beschäftigen, merkt schnell, dass ein relativ kleines Set an Knoten ausreicht, um individuelle Anwendungsfälle zu kreieren. In diesem Sinne viel Spaß mit KNIME!

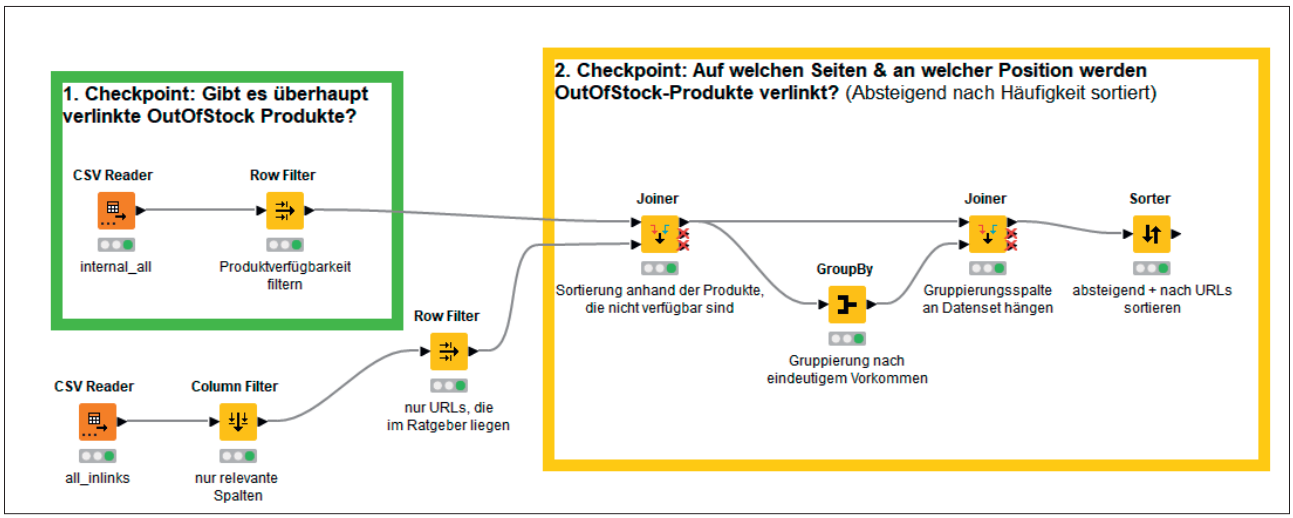

Abb. 13: Gesamter KNIME-Workflow