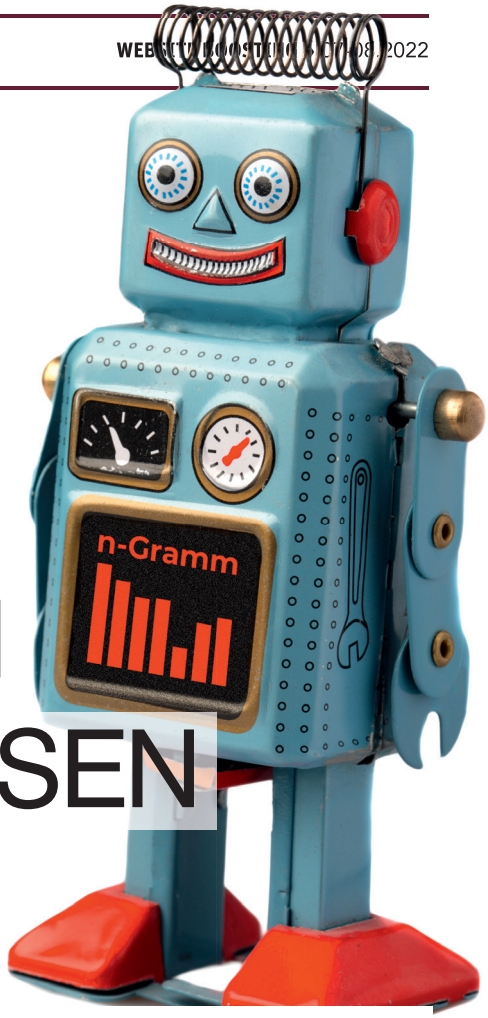
Michael Göpfert

BEYOND CRAWLING: INSIGHTS DURCH N-GRAMM-ANALYSEN

(TEIL 4)



Die Analyse von großen Dokumenten und Keywordlisten ist oft mühselig und echter Erkenntnisgewinn schwierig. Eine Analyse von N-Grammen kann dabei eine völlig neue Perspektive eröffnen, Zusammenhänge aufzeigen und Potenziale veranschaulichen – sowohl bei der Wettbewerbsanalyse als auch bei der eigenen Website und der Keyword-Recherche. Michael Göpfert zeigt Ihnen, wie Sie mit dem kostenlosen und wirklich von jedermann bedienbaren Tool KNIME recht einfach Antworten auf solche Problemstellungen erhalten.

Wer wurde nicht schon einmal mit der Frage „Was machen unsere Wettbewerber eigentlich anders als wir?“ konfrontiert? Zumindest was die inhaltliche Aufstellung der Marktbegleiter angeht, lässt sich diese Frage recht leicht beantworten.

Durch eine N-Gramm-Analyse lassen sich einfach und schnell quantitative Rückschlüsse über den Inhalt einer Website (oder anderer Textdokumente bzw. Listen) ziehen.

Beispielsweise aus Crawl-Daten von Screaming Frog. Sie eignen sich hervorragend, um mithilfe einer N-Gramm-Analyse zu entdecken, welche Begriffe auf einer Website besonders häufig vorkommen.

Dieses Verfahren ist relativ simpel, kann aber spannende Einblicke geben:

- Wie steht es um das Sortiment eines Online-Shops?
- Zu welchen Themen wird besonders häufig Content veröffentlicht?
- Werden Themen, zu denen es viele Inhalte gibt, in der Seitenarchitektur ausreichend berücksichtigt?

Wie bereits angedeutet, ist dieses Verfahren keineswegs auf die Analyse von Crawldaten beschränkt, sondern kann auch ganz andere Fragestellungen beantworten:

- Welche Keywords werden häufig zusammen gesucht?
- Für welche Begriffe bekommt meine Website besonders viele Impressionen?
- Welche Wortkombinationen kosten mich Werbebudget, konvertieren aber nicht?

DER AUTOR

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

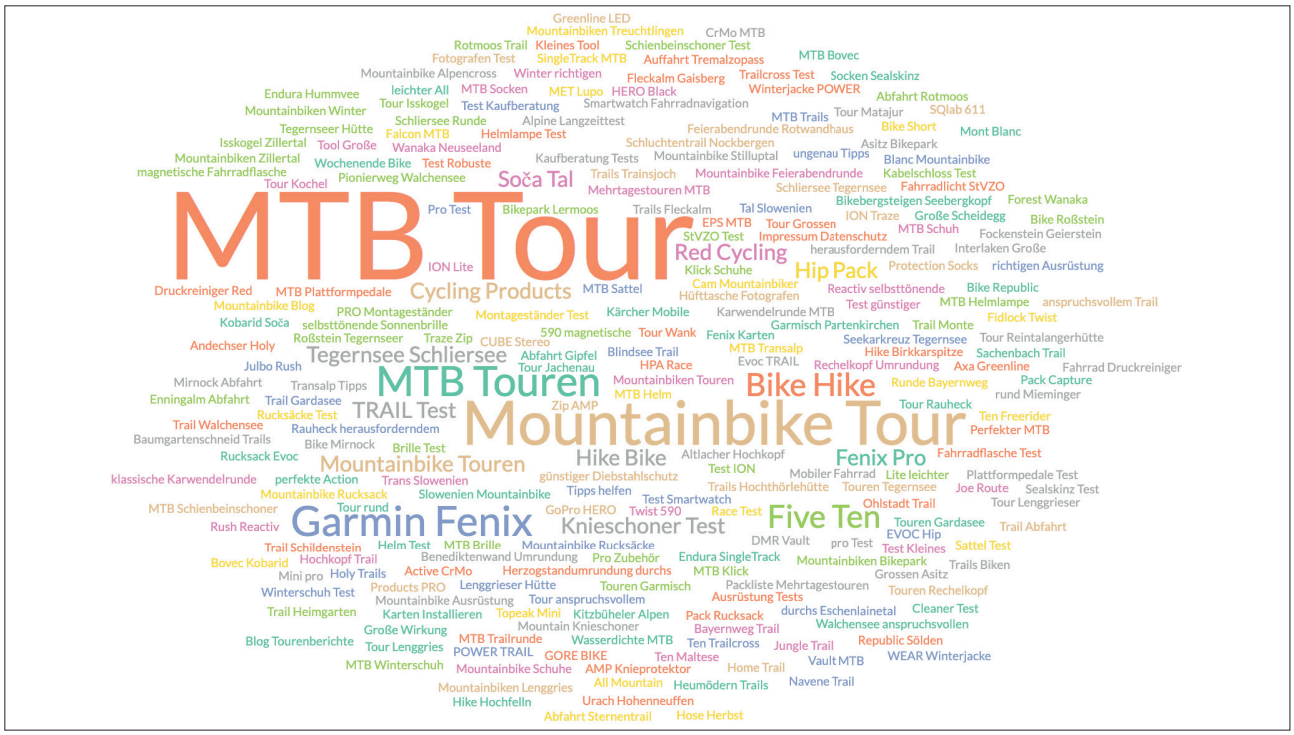


Abb.1: N-Gramm Tag Cloud aus Crawldaten

Was ist überhaupt ein N-Gramm?

Der Begriff „N-Gramm“ hat seinen Ursprung in der Linguistik und beschreibt eine Abfolge von N aufeinanderfolgenden Fragmenten. Das N steht dabei für die Anzahl an Fragmenten. Fragmente sind in diesem Fall Wörter bzw. Terme.

Je nachdem, wie viele Fragmente ein N-Gramm enthält, ändert sich seine Bezeichnung: Ein N-Gramm mit zwei aufeinanderfolgenden Fragmenten wird als Bigramm bezeichnet, drei aufeinanderfolgende Fragmente werden Trigramm genannt. Ist nur ein einziges Fragment enthalten, wird von einem Monogramm gesprochen.

Wie in Tabelle 1 zu sehen ist kann der Satz „Website Boosting ist ein hochwertiges Magazin“ folgende N-Gramme erzeugen.

Spannend wird es, wenn das zu analysierende Dokument mehr als einen Satz enthält und bei der Analyse auch die Häufigkeit der N-Gramme ermittelt wird.

Bei den Monogrammen führt dies zu einer Termfrequenz-Analyse, die vielen SEOs bekannt vorkommen könnte. Dabei wird ermittelt, wie häufig jeder Term in einem Dokument vorkommt.

Eine B-Gramm-Analyse hingegen ermittelt, wie häufig Zweiwortkombinationen in einem Dokument vorkommen. So lässt sich ermitteln, welche Begriffe häufig in Kombination miteinander verwendet werden.

Eine Monogramm-Analyse aller Hauptüberschriften von Produktseiten eines Online-Shops könnte beispielsweise ergeben, dass der Begriff „Sneaker“ am häufigsten in allen Headlines vorkommt. Eine Bigramm-Analyse könnte zeigen, dass „Nike Sneaker“ die häufigste Zweiwortkombination in den Produktnamen des Shops ist. Daraus kann abgeleitet werden, dass Sneaker wohl den größten Teil des Sortiments ausmachen und Nike davon die meisten Produkte stellt.

Alternativ kann auf diese Weise auch analysiert werden, wo die inhaltlichen Schwerpunkte einer Website liegen, sowohl bei der eigenen als auch beim Wettbewerb.

Automatisierung von N-Gramm-Analysen mit KNIME

Je größer der Datensatz ist, für den die N-Gramm-Analyse erstellt werden soll, desto aussagekräftiger sind die Erkenntnisse. Deshalb ist es sinnvoll, eine Automatisierung mit der kostenlosen Datenanalyse-Software KNIME vorzunehmen, um auch große Daten-

Monogramm	website	boosting	ist	ein	hochwertiges	magazin
Bigramm	website boosting	boosting ist	ist ein	ein hochwertiges	hochwertiges magazin	
Trigramm	website boosting ist	boosting ist ein	ist ein hoch-wertiges	ein hochwertiges magazin		

Tabelle 1

mengen performant zu analysieren und den Workflow bei Bedarf immer wieder nutzen zu können.

Außerdem bietet KNIME die Möglichkeit, die Daten vorab zu bereinigen, damit keine Füllwörter oder Satzzeichen die Analyse verzerren.

Voraussetzung ist also, dass KNIME bereits auf dem eigenen Rechner installiert ist. Anschließend kann der fertige Workflow unter *https://kni.me/w/TMp2hQZ79IfYdhc8* heruntergeladen, in KNIME geöffnet und genutzt werden. Alternativ kann der Workflow anhand der folgenden Beschreibung selbst erstellt und an die eigenen Bedürfnisse angepasst werden.

Installation von KNIME-Erweiterungen

Um die volle Funktionsvielfalt von Textverarbeitung in KNIME nutzen zu können, muss die Erweiterung „KNIME Textprocessing“ installiert werden. Diese kann unter *https://kni.me/e/PH_ptBLLdL1Mich2* aufgerufen werden und von dort per Drag and Drop auf die KNIME-Oberfläche am eigenen PC gezogen werden. Anschließend beginnt KNIME mit der Installation der Erweiterung.

Wie in den KNIME-Workflows in letzten Ausgaben wird auch dieses Mal ein Crawl aus Screaming Frog als Datenquelle für die Analyse verwendet. In diesem Fall der Export der H1.

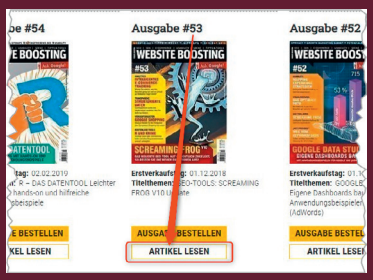
Nach dem Import der Crawl-Daten mit der Node „CSV Reader“ muss die Spalte, in der die Headlines enthalten sind, in ein Dokument umgewandelt werden. Die Prüfung, wie häufig ein N-Gramm vorkommt, soll schließlich über alle Headlines laufen und nicht nur über eine einzelne Zelle. Außerdem setzen viele Textprocessing Nodes voraus, dass es sich bei den zu analysierenden Daten um ein Dokument handelt.

Die Umwandlung in ein Dokument erfolgt über die Node „Strings To Document“. Nachdem diese im Workflow platziert wurde und mit dem CSV Reader verbunden wurde, muss diese konfiguriert werden. Ein Doppelklick auf die Node öffnet das Konfigurationsmenü. Hier muss lediglich im Bereich Title und Text die Spalte, die analysiert werden soll (in diesem Fall „H1-1“), ausgewählt werden. Anschließend kann die Node ausgeführt werden.

Mit Ausführen der Node wird eine neue Spalte namens „Document“ erzeugt, in der ebenfalls die Inhalte der H1 stehen. Diese Spalte ist nun die Basis der weiteren Analyse.

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*.



Datenbereinigung

Um die Qualität der Analyse möglichst hochzuhalten, ist es sinnvoll, die Daten noch etwas zu bereinigen. Mit der Node „N Chars Filter“ können im Tab „Filter options“ alle Begriffe herausgefiltert werden, die weniger als eine bestimmte Anzahl an Zeichen enthalten. So können bei Bedarf einfach Füllwörter wie „zu“ oder „ab“ herausgefiltert werden. Doch Vorsicht: Dabei können auch wichtige Begriffe verloren gehen. Wird der Filter auf den Wert 2 gesetzt, werden alle Begriffe, die weni-

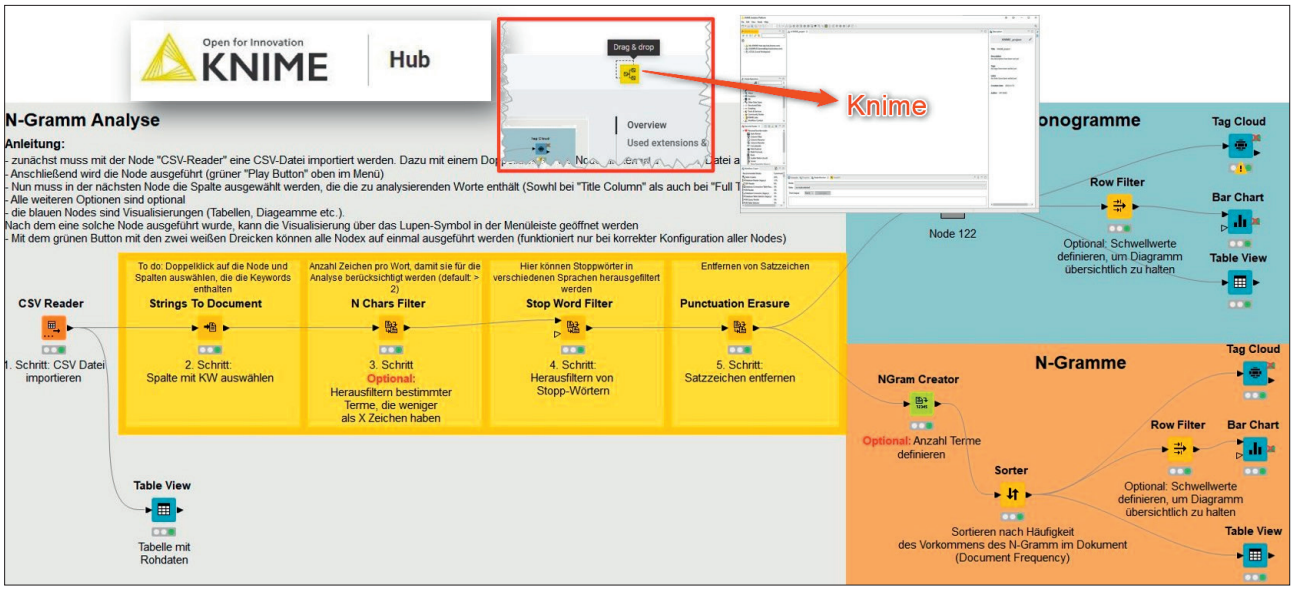


Abb. 2: Den kompletten Workflow kann man sich bei Bedarf einfach vom KNIME-Hub (URL im Beitrag) per Drag and Drop in die Arbeitsfläche von KNIME ziehen.

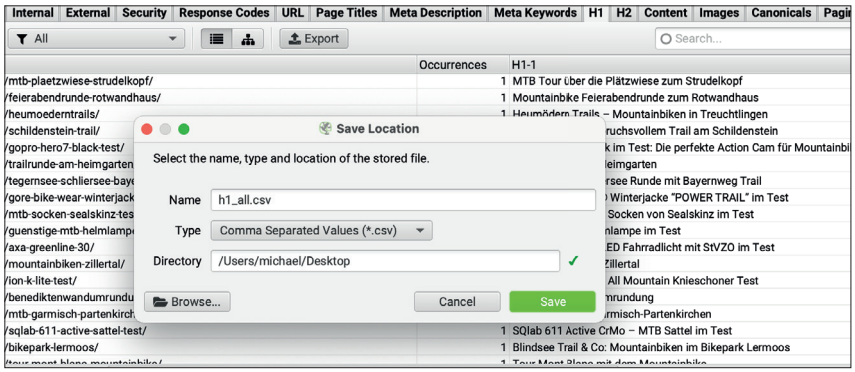


Abb. 3: Export der H1-Überschriften aus Screaming Frog

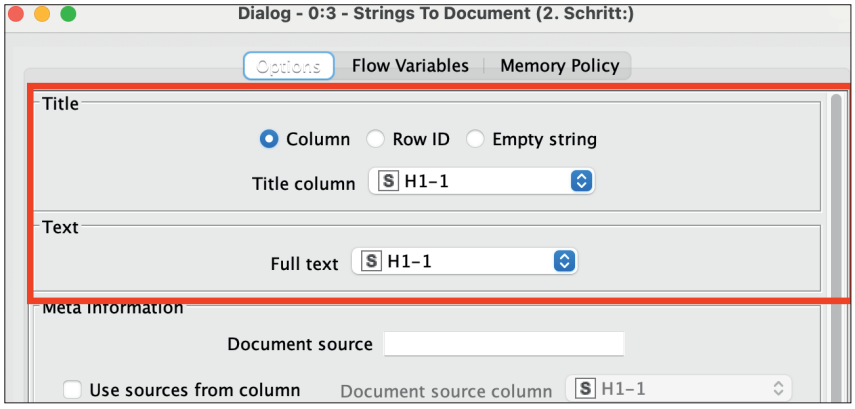


Abb. 4: Umwandlung der Strings in ein Dokument

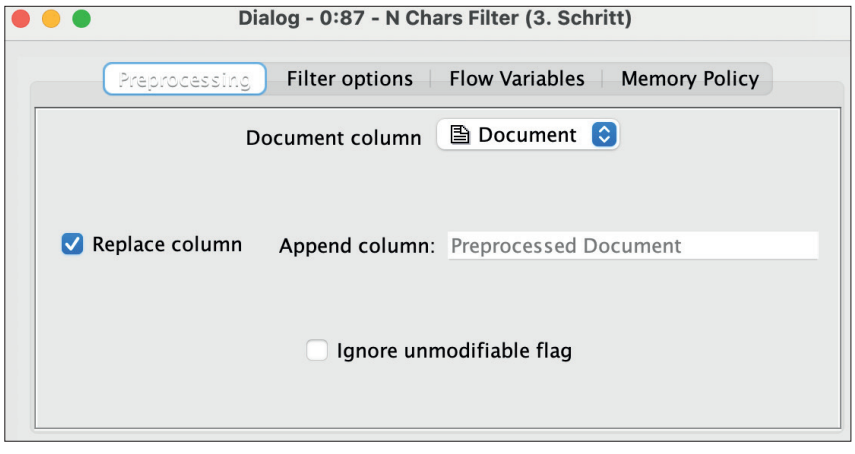


Abb. 5: Ersetzen der Spalte „Document“

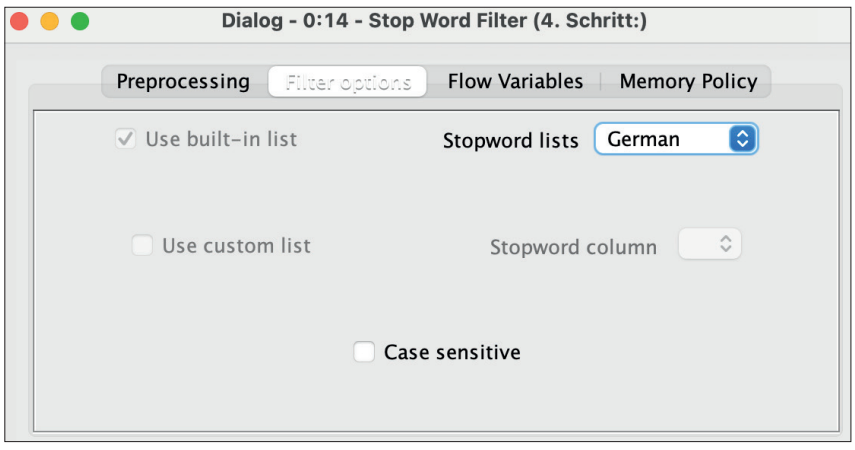


Abb. 6: Herausfiltern von Stoppwörtern

ger als zwei Zeichen haben, herausgefiltert. Bei einem Unterwäschehändler, der BHs verkauft, wäre das keine gute Option.

Wichtig ist auch, dass der Haken bei „Replace Column“ gesetzt wird, damit wird die Spalte, die mit der „String to Document“ Node erzeugt wurde, überschrieben, und die nächste Node kann auf diese Daten zugreifen.

Im nächsten Schritt werden mit der Node „Stop Word Filter“ Füllwörter entfernt. Praktischerweise bedient sich die Node an vorgefertigten Listen mit Stoppwörtern, die unter anderem auch in Deutsch zur Verfügung stehen.

Nun zeigen sich schon deutliche Unterschiede zwischen den Ursprungsdaten in der Spalte „H1-1“ und den bereinigten Daten in der Spalte „Document“. So wurde aus der Überschrift „MTB-Tour am Walchensee mit dem anspruchsvollen Snakeline Trail“ nun „MTB-Tour Walchensee anspruchsvollen Snakeline Trail“.

Im letzten Schritt der Datenbereinigung werden mit der Node „Punctuation Erasure“ sämtliche Satzzeichen entfernt.

Auf Basis der bereinigten Liste wird der Workflow nun aufgesplittet in einen Teil, der für das Zählen der Monogramme zuständig ist, und einen, der sich um die restlichen N-Gramme kümmert (Abb. 6).

Monogramme

Um die Monogramme zu erzeugen, muss die Spalte „Documents“ mit der Node „Bag Of Words Creator“ in eine Bag of Words umgewandelt werden. Das führt dazu, dass jede Zelle in der Spalte „Documents“ in ihre Terme zerlegt wird und jeder Term eine eigene Zeile erzeugt (Abb. 8).

Nun können die Terme mit der Node „Term to String“ wieder in normalen Text (Strings) umgewandelt werden.

Um die Terme zu zählen und Duplikate zu entfernen, bietet sich die Node

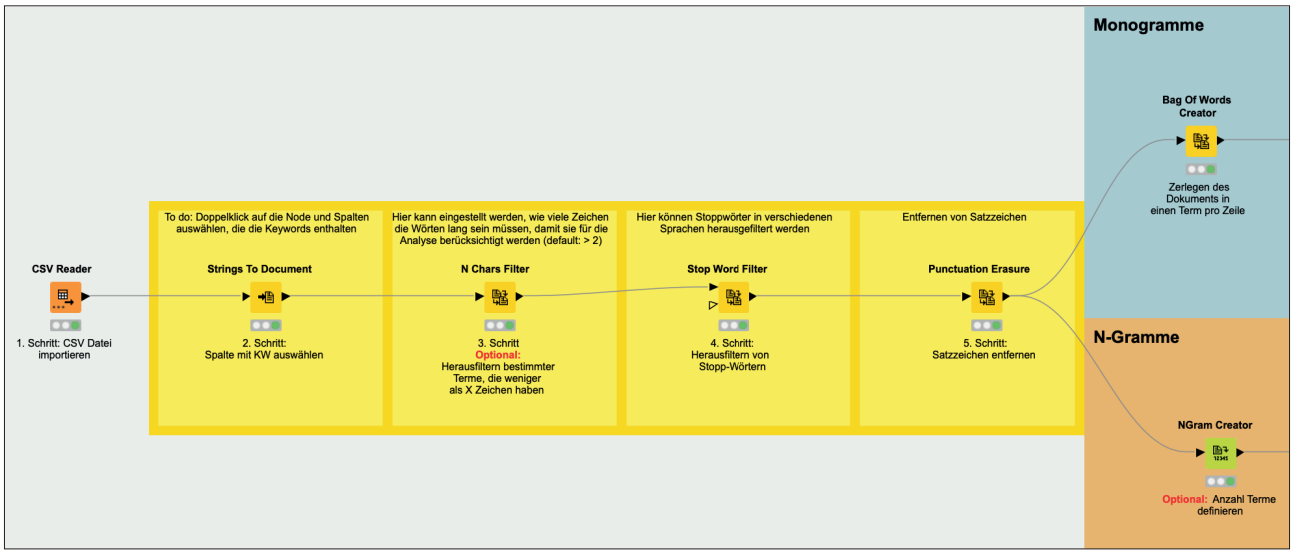


Abb. 7: Aufteilung des Workflows in Monogramme und Bigramme

„GroupBy“ an. Sie wird wie folgt konfiguriert: Im Tab „Groups“ wird festgelegt, nach welcher Spalte gruppiert werden soll, was dazu führt, dass alle anderen Spalten entfernt werden, ähnlich wie in einer Pivot-Tabelle in Excel.

Im Tab „Manual Aggregation“ wird die Spalte „Term“ ausgewählt. Im Bereich Aggregation wird nun „Count“ ausgewählt (Abb. 9). Das sorgt dafür, dass gezählt wird, wie häufig die einzelnen Terme in der Spalte „Term“ vorkommen.

Tipp: Bei sehr großen Datenmengen sollte der Wert im Feld „Maximum unique Values per Group“ deutlich nach oben gesetzt werden, sonst nimmt KNIME nur die ersten 10.000. Außerdem sollte bei den Advanced settings die Einstellung „Keep original name(s)“ getroffen werden, damit die Benamung der Spalten gleich bleibt.

Als Nächstes ist es an der Zeit, die Monogramme nach ihrer Häufigkeit zu sortieren. Dies geschieht über die Node „Sorter“. Hier einfach die Spalte „Term“ auswählen und mit „Descending“ absteigend sortieren.

Damit die Spalten etwas aussagekräftigere Namen bekommen, können diese noch mit der Node „Column Rename“ angepasst werden.

Document	Term
"MTB Tour Plätzwiese Strudelkopf"	MTB[]
"MTB Tour Plätzwiese Strudelkopf"	Tour[]
"MTB Tour Plätzwiese Strudelkopf"	Plätzwiese[]
"MTB Tour Plätzwiese Strudelkopf"	Strudelkopf[]
"Mountainbike Feierabendrunde Rotwandhaus"	Mountainbike[]
"Mountainbike Feierabendrunde Rotwandhaus"	Feierabendrunde[]
"Mountainbike Feierabendrunde Rotwandhaus"	Rotwandhaus[]
"Heumödern Trails Mountainbiken Treuchtlingen"	Heumödern[]
"Heumödern Trails Mountainbiken Treuchtlingen"	Trails[]
"Heumödern Trails Mountainbiken Treuchtlingen"	Mountainbiken[]
"Heumödern Trails Mountainbiken Treuchtlingen"	Treuchtlingen[]
"MTB Tour anspruchsvollem Trail Schildenstein"	MTB[]
"MTB Tour anspruchsvollem Trail Schildenstein"	Tour[]
"MTB Tour anspruchsvollem Trail Schildenstein"	anspruchsvollem[]
"MTB Tour anspruchsvollem Trail Schildenstein"	Trail[]
"MTB Tour anspruchsvollem Trail Schildenstein"	Schildenstein[]

Abb. 8: Neue Spalte mit einzelnen Termen

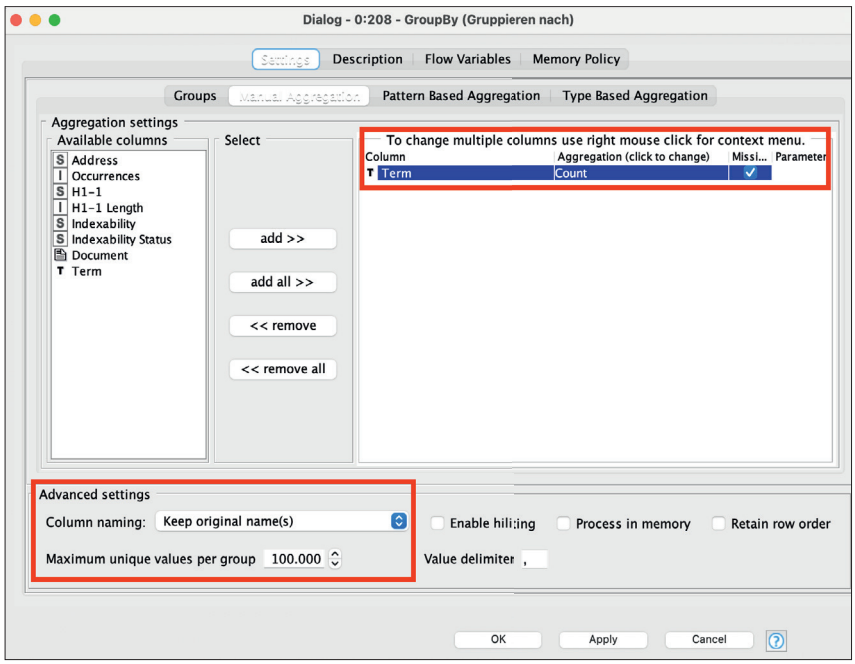


Abb. 9: Neue Spalte mit einzelnen Termen

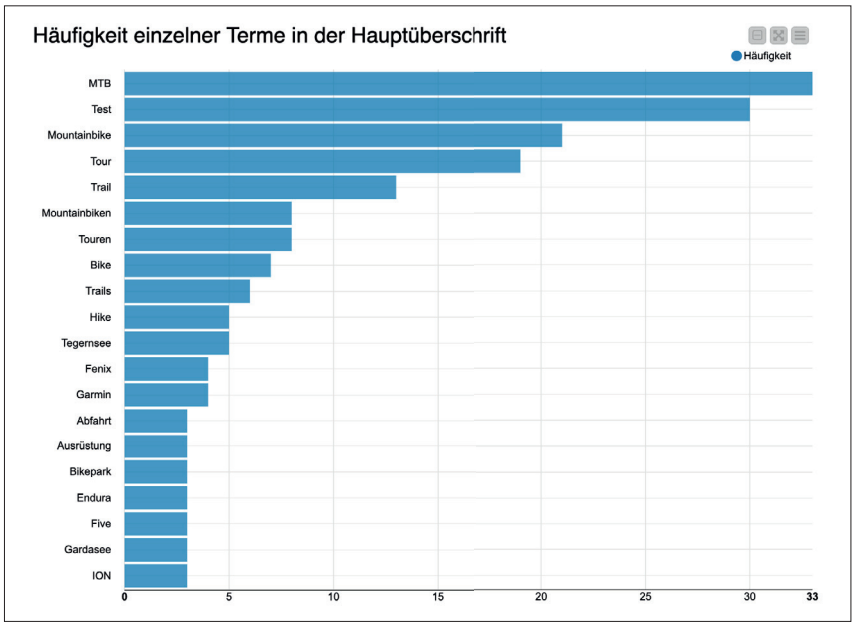


Abb. 10: Balkendiagramm mit Monogrammen

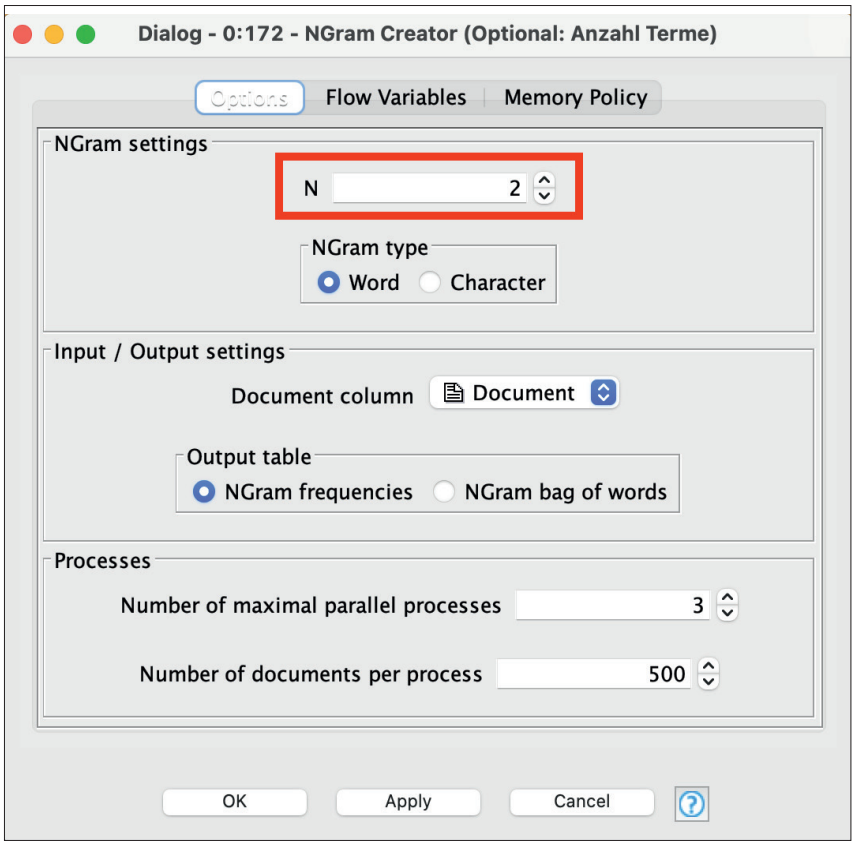


Abb. 11: Konfiguration des NGram Creator

Visualisierung der Monogramme

Bis jetzt liegen die Monogramme lediglich in Tabellenform vor. Soll die Auswertung jedoch präsentiert werden, stehen verschiedene Möglichkeiten zur Visualisierung zur Verfügung.

Am ansprechendsten ist häufig eine Tag Cloud mit der gleichnamigen Node (Abb. 1). Um zu sehen, wie groß die Unterschiede zwischen den einzelnen Monogrammen sind, lässt sich auch ein Bar Chart erstellen. Dazu sollte aber vorher noch mit einem Row-Filter dafür gesorgt werden, dass nicht zu viele Zeilen im Chart abgebildet werden, weil sonst die Übersichtlichkeit leidet.

Als dritte Visualisierung bietet sich die Node „Table View“ an. Sie erstellt eine filterbare Tabelle.

Damit lässt sich schon recht gut erkennen, was die Topthemen auf der gecrawlten Website sind: Mountainbike-Touren und Produkttests.

N-Gramme

Die Erstellung der Bi- und Trigramme ist deutlich einfacher als die der Monogramme. Es muss lediglich die Node „Ngram creator“ mit der letzten Node der Datenbereinigung (in unserem Fall „Punctuation Erasure“) verbunden werden. Anschließend kann in der Konfiguration des Ngram Creators eingestellt werden, aus wie vielen Termen die N-Gramme bestehen sollen.

Nun muss nur noch mit der Sorter-Node die Sortierung angepasst werden, sodass die N-Gramme absteigend nach ihrem Vorkommen im Dokument (Spalte „Document Frequency“) sortiert werden.

Abschließend können Nodes für die Visualisierungen der Monogramme per Copy and Paste für die N-Gramme übernommen werden. Es müssen lediglich die Namen der Spalten, die visualisiert werden sollen, in den Einstellungen der Chart-Nodes angepasst werden.

Fazit

Ist der Workflow einmal erstellt und konfiguriert, bieten N-Gramm-Analysen eine schnelle Möglichkeit, quantitative Einblicke über die Inhalte auf einer Website zur erhalten. Die Flexibilität des Workflows erlaubt es auch, diese Analyse auf andere Datensätze aus CSV-Dateien anzuwenden, beispielweise auf Exporte aus einem Keyword Tool, den Daten der Google Search Console oder Google Ads. ¶