# Soekia 2.0 - Handreichung für Lehrpersonen

Informationen im Internet zu finden ist ein vermeintliches Kinderspiel! Wer sich für Informationen zur Entstehung des Ozonlochs interessiert, gibt auf Google die Suchanfrage "Ursache Entstehung Ozonloch" ein und erhält in Sekundenbruchteilen Tausende von Treffern. Ein anderer Benutzer gibt die leicht anderen Suchbegriffe "Ursachen Entstehung des Ozonlochs" ein und erhält viel weniger und teils ganz andere Treffer. Während für uns Menschen verschiedene Wortformen wie “Ursache” und “Ursachen” oder Deklinationsformen wie “Ozonloch”, “Ozonlochs” oder “Ozonloches” semantisch gleichwertig sind, interpretieren Suchmaschinen diese Worte oft als unterschiedliche Begriffe und liefern uns verschiedene Antworten. Im Alltag können uns so durchaus wichtige Informationen vorenthalten werden, etwa wenn die Suchmaschine zu den Anfragen “Ferienhaus Toskana”, “Ferienhäuser Toskana”, “Ferienhäuschen Toskana” oder “Toskana Ferienhaus” verschiedene Vorschläge liefert. Gravierender können die Folgen bei Recherchen zu wissenschaftlichen Themen sein, wenn man relevante Informationen übersieht.

Die Informationsbeschaffung im Internet beschränkt sich deshalb keineswegs auf das Stellen einfacher Suchanfragen bei einer Suchmaschine. Effiziente und effektive Informationsbeschaffung ist eine anspruchsvolle Aufgabe, die ein fundiertes Verständnis für die Funktionsweise von Suchmaschinen voraussetzt. Wie eine Suchmaschine arbeitet, bleibt aber den Benutzern weitgehend verborgen. Die wichtigen Komponenten einer Suchmaschine sind das Erfassen von Webseiten durch Webroboter (Crawling, Spidering), das Erstellen einer effizienten Datenstruktur für die Suche (Indexierung, Index), das Finden zu einer Benutzeranfrage passender Dokumente (Matching) und die Präsentation der gefunden Dokumente in einer guten Reihenfolge (Rangierung). Zusammengefasst ist der grundsätzliche Ablauf in Abbildung 1.

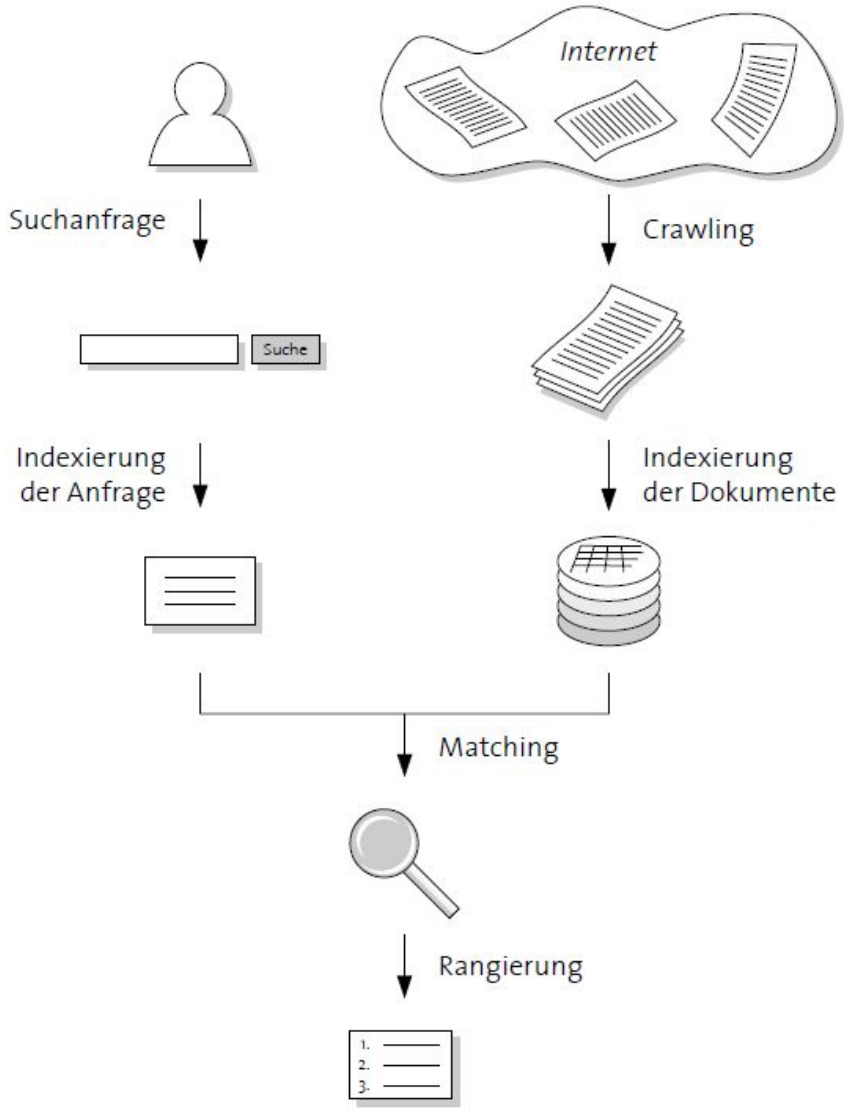



Abb. 1: Komponenten einer Suchmaschine
(Abbildung aus http://swisseduc.ch/informatik/internet/internet_recherche/ )

Die *didaktische Suchmaschine Soekia* ermöglicht einen Blick hinter die Kulissen von Suchmaschinen. Soekia ist speziell für den Unterricht konzipiert worden und kann nicht als Suchmaschine im Internet genutzt werden. Eine Internet-Suchmaschine durchsucht ständig die ihr zugänglichen Dokumente im Internet (Webseiten, PDF-Dateien usw.). Diese wichtige Komponente einer Suchmaschine kann mit Soekia nicht gezeigt werden. Soekia durchsucht nur ein “Experimentier-Internet” in Form einer kleinen, manuell zusammengestellten und damit überschaubaren Dokumentenkollektion. Als Benutzer von Soekia stellt man selbst einige Dokumente in einer Dokumentenkollektion zusammen und untersucht dann, wie diese Dokumente indexiert und zu Suchanfragen rangiert werden. Da man sehr einfach weitere Dokumente hinzufügen, entfernen oder abändern kann, ermöglicht Soekia eine ganze Reihe von kleinen Experimenten. So kann etwa untersucht werden, welche Auswirkung auf die Rangierung die Häufigkeit des Vorkommens eines Suchbegriffs in einem Dokument hat oder welchen Einfluss Wortnormalisierung auf den Umfang des Indexes und damit auf die Ausbeute und Präzision einer Recherche haben.

Die beiden Begriffe Ausbeute und Präzision sind dabei zentral für die Informationssuche. In Abbildung 2 sind verschiedene Suchen dargestellt. Zu einer Suchanfrage gibt es eine Menge tatsächlich relevanter Dokumente (rot) die jedoch nicht alle durch die Suche erschlossen werden (gefundene Dokumente). Eine Suche mit hoher Präzision, aber geringer Ausbeute verpasst zum Beispiel einen Grossteil der relevanten Dokument, liefert aber wenig irrelevante Dokumente. Sowohl hohe Präzision als auch Ausbeute zu erzielen ist in der Praxis jedoch sehr schwierig. Beginnt eine Recherche mit einem einzigen Suchbegriff liefert die Suchmaschine zunächst eine hohe Ausbeute mit geringer Präzision. Durch das Hinzufügen weiterer Suchbegriffe kann die Präzision erhöht, aber die Ausbeute reduziert werden. Die Wunschvorstellung wird in der Abbildung durch das Diagramm rechts unten illustriert: Man finden alle relevanten Dokumente und keine irrelevanten Dokumente.

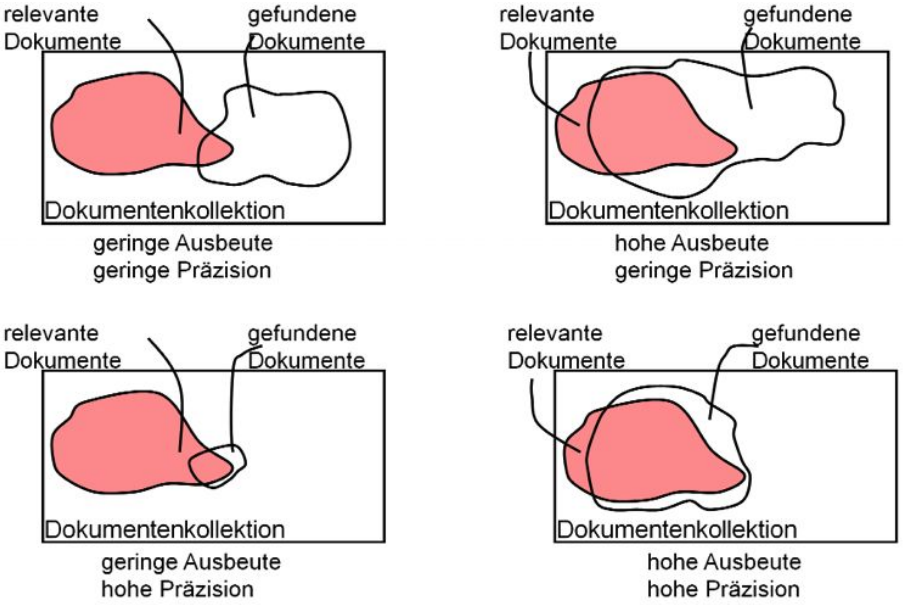



Abb. 2: Ausbeute und Präzision bei der Suche in einer Dokumentenkollektion

# 1. Der Mensch als entscheidender Faktor bei der Suche

Bereits 1971 hielt Tefko Saracevic, einer der Mitbegründer der Wissenschaftsdisziplin "Information Retrieval" fest:

> *The human factor, i.e. variations introduced by human decision-making, seems to be the major factor affecting performance of every and all components of an information retrieval system.*[1]

Aus vielen Untersuchungen zum Benutzerverhalten bei Suchmaschinen ist bekannt, dass die Benutzer (User) die angeboten Möglichkeiten von Suchdiensten nur selten richtig nutzen:

- User verwenden zwar oft Suchmaschinen, planen die Suche aber kaum. Insbesondere legen sich nur wenige User vorgängig Rechenschaft ab, ob es sich bei der Fragestellung um eine offene Frage oder eine geschlossene Frage handelt. Offene Fragen (z.B. Was sind die Ursachen für Hooliganismus?) bedingen eine ausbeuteorientierte Suche und können nicht mit einer einmaligen einfachen Suchanfrage beantwortet werden. Geschlossene Fragen hingegen (z.B. Wie hoch ist der Eiffelturm?) sind präzisionsorientiert und können ohne Weiteres mit einer einfachen Suchanfrage (z.B. "Höhe Eiffelturm") effizient beantwortet werden.
- Die User verwenden einfache Suchanfragen mit in der Regel nur wenigen Suchbegriffen, die zudem oft sehr unspezifisch sind.
- User schauen häufig nur die erste Seite mit den Treffern an.
- User verfeinern nur selten eine erste Suchanfrage mit weiteren Suchbegriffen, welche die Resultate noch mehr eingrenzen.
- User schätzen die Glaubwürdigkeit der Informationen auf dem Web als hoch ein und sind sich nicht bewusst, dass sie oft einen Grossteil der an und für sich relevanten Dokumente übersehen.

Ein grosser Teil dieses unzulänglichen Verhaltens der Benutzer beruht wohl darauf, dass nur die wenigsten eine Vorstellung von der Funktionsweise einer Suchmaschine haben. Die meisten Leute sind der falschen Ansicht, dass Suchmaschinen bei einer Anfrage das Internet in Echtzeit (also "gerade in diesem Moment") durchsuchen. Eine klare Vorstellung vom Aufbau und der Funktionsweise eines Indexes kann hier vielen Missverständnissen vorbeugen.

[1] Saracevic, Tefko: Selected results from an inquiry into testing of information retrieval systems. Journal of the American Society for Information Science, 22(2):126–139, 1971

## 2. Erfassen und Speichern aller Dokumente einer Kollektion

Es gibt Millionen von Web-Servern, die Hunderte Millionen von Webseiten anbieten. Irgendwann entsteht auf irgendeinem dieser Server eine neue Seite, ein neues Dokument oder es wird eine der bestehenden Seiten geändert oder gelöscht. Diese Änderungen werden den Suchmaschinen aber nicht automatisch mitgeteilt. Es gibt auch keine zentralen Meldestellen, bei der Website-Betreiber ihre Änderungen bekannt machen können. Jede Suchmaschine muss deshalb selbst herausfinden, ob sich die Inhalte im Internet verändert haben. Bedenkt man die Grösse mit Hunderten Millionen von Webseiten, scheint dies ein schier unmögliches Unterfangen. Hier kommt der Web-Roboter (auch Spider oder Crawler genannt) ins Spiel. Der Web-Roboter ist ein Programm, das zur Aufgabe hat, Webseiten zu finden. Dazu nutzt der Roboter die Eigenschaft des World Wide Web aus, dass die Dokumente über Hyperlinks miteinander verbunden sind. Ausgehend von einer Liste von URLs durchsucht der Web-Roboter Webseiten nach weiterführenden Verweisen (Hyperlinks). Die gefundenen Hyperlinks landen ebenfalls in der Tabelle mit den URLs, damit der Web-Roboter über die schon besuchten Seiten Bescheid weiss. Später werden auch die neu eingetragenen Seiten nach weiteren Verweisen untersucht. Auf diese Weise arbeitet sich der Web-Roboter immer weiter in die Tiefen des World Wide Web vor. Früher oder später findet der Web-Roboter somit alle Seiten, die auf irgendeinem Weg von den Startseiten aus erreicht werden können Dabei können die Roboter aber nur jene Webseiten durchsuchen, welche öffentlich, ohne Zugangskontrolle oder weitere Einschränkungen abrufbar sind. Ein Grossteil der Inhalte des Internets wird deshalb von Suchmaschinen gar nicht erfasst.

Den Prozess des Erfassens von Webseiten kann man mit Soekia angesichts der immensen Grösse des Internets nicht zeigen. Erfahrungsgemäss ist dieser Sachverhalt für Schülerinnen und Schüler aber auch leicht nachvollziehbar. Wir alle kennen die Methoden, um ein Labyrinth abzuwandern. Genau so funktioniert ein Web-Roboter. Im Gegensatz zum Labyrinth ist bei Suchmaschinen der Prozess aber nie abgeschlossen. Ist die Suche innerhalb eines Bereichs des Internet abgeschlossen, beginnt sie gleich wieder von vorn, um mögliche neue Änderungen zu entdecken.

Damit die Benutzer nicht selbst HTML-Dokumente zusammenzustellen müssen, stellt Soekia zwei Ausgangskollektionen zur Verfügung. Die Dokumentenkollektion “Ozonschicht” umfasst 12 Dokumente, “Autovermietung” 5 Dokumente. Es kann aber auch mit einer leeren Dokumentenkollektion gestartet werden oder eine eigene Dokumentenkollektion hochgeladen werden (siehe Abb. 3). Die Anzahl der Dokumente ist bewusst klein gehalten und ermöglicht so überschaubare Experimente.

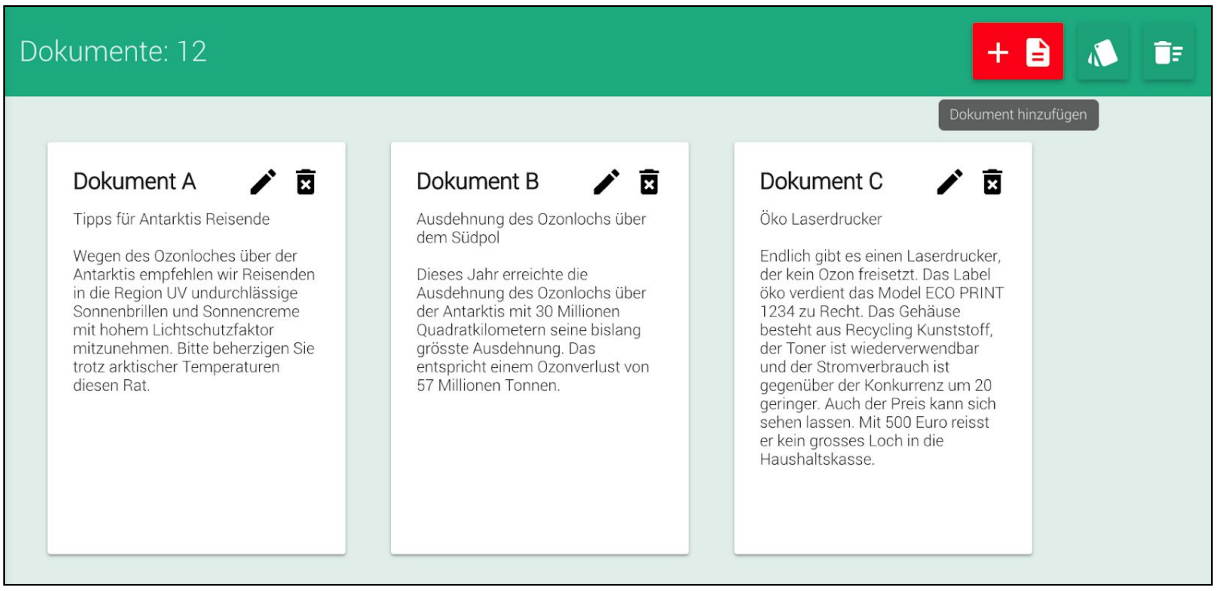



Abb. 3: Dokumentenkollektionen in Soekia 2.0

## 3. Indexierung: Erstellen einer effizienten Datenstruktur für die Suche

Die von einer Suchmaschine erfassten Dokumente bilden die sogenannte Dokumentenkollektion, in welcher die Suchmaschine sucht. Eine Ablage der einzelnen Webseiten und Dokumente in unveränderter Form ist nicht zweckmässig. Die Suchmaschine erstellt deshalb einen Index der Dokumentenkollektion. Der Index entspricht ziemlich genau dem Stichwortverzeichnis am Ende eines Buches und umfasst die in den Dokumenten vorkommenden Begriffe samt einem Verweis auf die entsprechenden Dokumente. Im Unterschied zum Stichwortverzeichnis eines Sachbuches ist der Index einer Suchmaschine natürlich viel umfangreicher. Die Analogie “Stichwortverzeichnis” eignet sich aber im Unterricht gut, um dem Zweck eines Indexes zu thematisieren und zu diskutieren, welche Stichworte warum in das Stichwortverzeichnis aufgenommen wurden (siehe Abb. 4).

Soekia stellt den Index in einer lesbaren, alphabetisch geordneten Form dar und zeigt die gesamte Anzahl der Einträge im Index an. Soekia zeigt auch, wie oft ein Begriff in der Dokumentenkollektion vorkommt (roter Balken) und in wie vielen Dokumenten er auftritt. Mit dieser Index-Darstellung lassen sich zahlreiche Fragen untersuchen. Wie verändert sich der Index beim Hinzufügen von gleichartigen Dokumenten zu einer Dokumentenkollektion, wie bei artfremden Dokumenten?

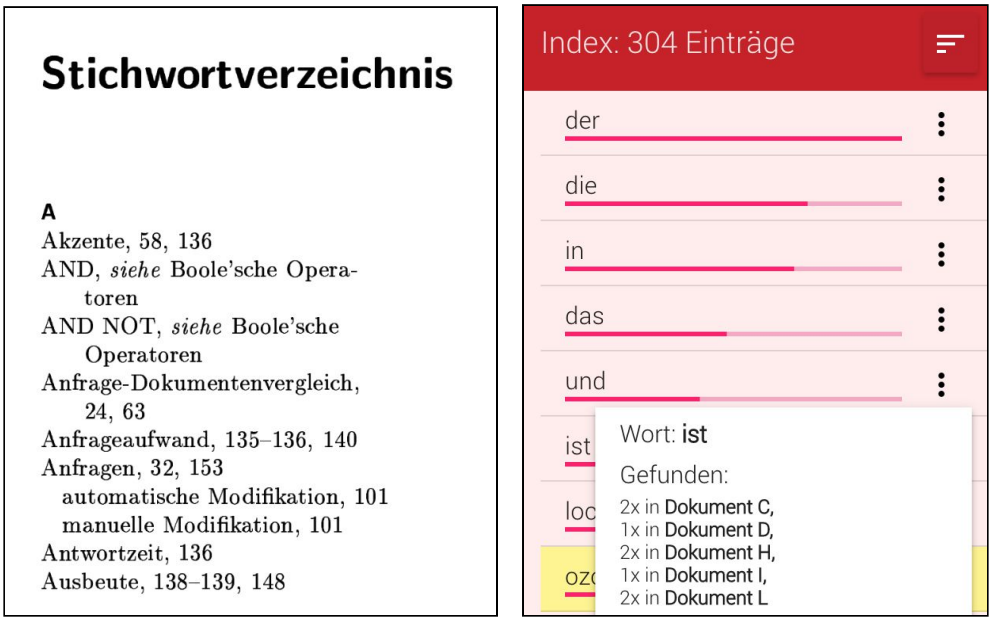



Abb. 4: Stichwortverzeichnis in einem Sachbuch (links) und Index in Soekia (rechts)

Die von Soekia zur Verfügung gestellten Dokumente können einzeln bearbeitet oder auch gelöscht werden. Ebenfalls können neue Dokumente hinzugefügt werden mit einer Länge von maximal 500 Zeichen. Soekia beschränkt die Länge der Dokumente aus didaktischen Überlegungen, um eine überschaubare Experimentierumgebung zur Verfügung zu stellen. Im Unterricht kann so auch mit einer Dokumentenkollektion bestehend aus einem einzigen Dokument gestartet werden und der Aufbau des Indexes untersucht werden (siehe Abb. 5).

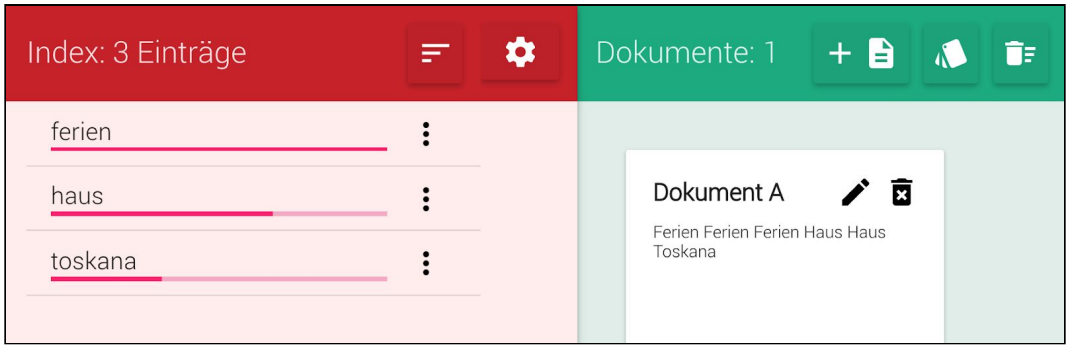



Abb.5: Entstehung des Indexes am Beispiel eines einfachen Dokumentes

Anschliessend bieten sich zahlreiche kleine Experimente rund um den Aufbau und die Länge des Indexes an. So können etwa die Auswirkung der Wortstammreduktion auf die Länge des Indexes untersucht werden und erste Überlegungen zu den Auswirkungen der Wortstammreduktion auf die Suchtreffer zu einer Suchanfrage angestellt werden.

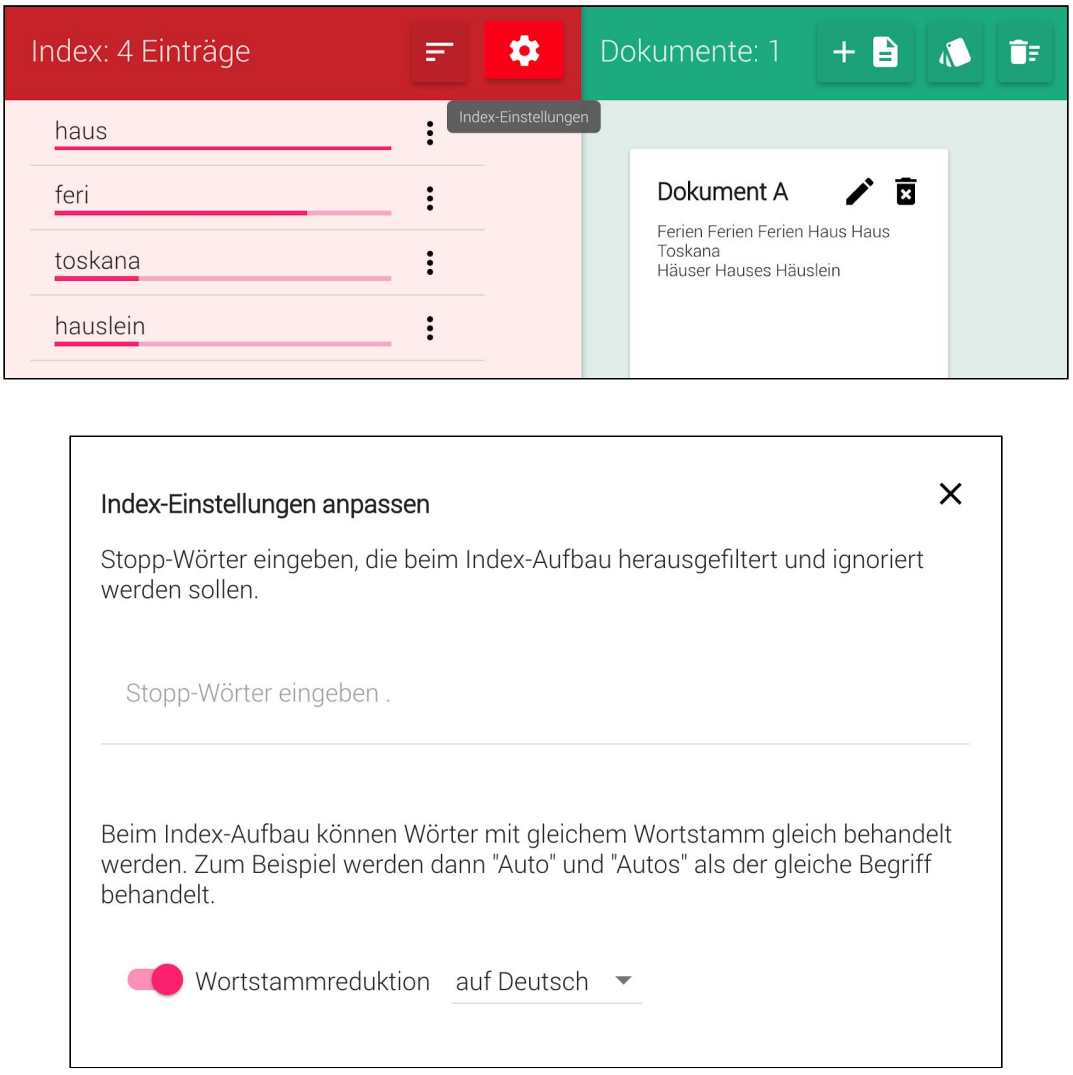



Abb. 6: Wortstammreduktion und seine Auswirkung auf die Länge des Indexes

## 4. Matching: Finden der zu einer Suchanfrage passenden Dokumente

Sinnvolle quantitative Vergleiche setzen immer einen dem Vergleich vorausgehenden Normalisierungsprozess voraus. Wir kennen das aus dem Alltag bestens: Um den Preis zweier Produkte zu vergleichen, müssen wir zuerst den Preis für gleich grosse Mengen des Produktes bestimmen. Genauso verhält es sich bei Suchdiensten im Internet. Eine wenig ausgereifte Suchmaschine vergleicht Wörter strikt Buchstabe für Buchstabe. "Ozonloch" ist somit ein anderer Begriff als "Ozonloches" oder "Ozonlöcher". Viele an und für sich relevante Dokumente werden so bei der Anfrage "Ozonloch" übersehen.

Ein besseres Verfahren reduziert Wörter auf den Wortstamm. Aus "Ozonloches" und "Ozonlöcher" wird bei diesem auch Stemming genannten Prozess der Wortstamm "Ozonloch". Wichtig ist, dass die Suchmaschine dieses Stemming-Verfahren nicht nur beim

Indexieren der Dokumente anwendet, sondern auch die Begriffe einer Suchanfrage demselben Prozess unterzieht. Nur so findet eine Anfrage mit "Ozonloches" ein Dokument mit “Ozonlöcher".

Soekia macht den Normalisierungsprozess transparent. Standardmässig werden Grossbuchstaben und Umlaute normalisiert. Obendrein kann man wählen, ob auch ein beschränktes, pseudo-linguistisches Wort-Stemming durchgeführt wird. Beim Stemming trennt Soekia häufige Endungen ab, darunter Substantivendungen wie -heit, -keit und -ung, Adjektivendungen wie -bar, -er, -sten und -lich sowie Verbalendungen -end, -et, -st. Zusätzlich werden die Vorsilben ge-, ver- und un- abgetrennt. Dabei kann sowohl over-stemming wie auch under-stemming auftreten. Beim over-stemming werden Wörter auf einen gemeinsamen Stamm reduziert, obwohl sie semantisch nicht miteinander verwandt sind. Beispiel: Das Substantiv "Versicherung" und das Reflexivpronomen "sich" werden beide auf "sich" abgebildet. Das Gegenteil heisst under-stemming und ist im Deutschen ohne Wörterbuch nicht zu verhindern. Stark gebeugte Verben wie "gehen, gingen, gegangen" lassen sich nicht durch Abtrennen von Endungen auf denselben Stamm reduzieren. Die Sprachwahl zwischen Deutsch, Französisch und Englisch erlaubt es sehr schön die Unterschiede zwischen verschiedenen Sprachen aus Sicht einer Suchmaschine aufzuzeigen. Für Suchmaschinen ist Deutsch ein deutlich schwierigere Sprache als Englisch, weil es in der deutschen Sprache unter anderem viele zusammengesetzte Wörter gibt.

Ausserdem besteht in Soekia die Möglichkeit, häufige Wörter (sog. Stoppwörter) zu eliminieren. Stoppwörter wie "auf, der, die, das, ein, in" haben wenig Informationsgehalt. Deren Erfassung würde aber den Index stark anwachsen lassen. Bei der Suche nach "Das Traumhaus" möchte man nicht ganz viele Dokumente mit "Das" finden. Genau dies würde aber passieren, da die Suchmaschine keine semantische Unterscheidung zwischen "Das" und "Traumhaus" vornimmt. Das Ignorieren von Stoppwörter kann damit deutlich die Präzision erhöhen.

Dass eine gute Normalisierung für eine erfolgreiche Suche entscheidend ist, kann mit kleinen Experimenten in Soekia sehr gut gezeigt werden. In der Beispiel-Kollektion "Ozonschicht” findet man mit der Suchanfrage "Ozonloch" ohne Wortstammreduktion nur zwei Dokumente, mit Wortstammreduktion hingegen vier Dokumente. Eines der zusätzlichen Dokumente enthält das Wort “Ozonlochs”, das andere das Wort “Ozonloches”.

## 5. Rangierung: Präsentation der gefundenen Dokumente in der Reihenfolge ihrer Relevanz

Zum Suchbegriff "Ozonloch” werden im Internet über Hunderttausend Dokumente gefunden. Für einen Benutzer ein Ding der Unmöglichkeit, alle diese Dokumente zu sichten. Für die Qualität einer Suchmaschine ausschlaggebend ist deshalb auch die Reihenfolge, in welcher die gefundenen Dokumente dem Benutzer angezeigt werden. Die relevantesten Treffer sollen in der Rangliste weit oben erscheinen. Untersuchungen des Benutzerverhaltens haben gezeigt, dass weniger als ein Drittel der Benutzer drei oder mehr Dokumente in der Rangliste einer Suchmaschine besuchen. Treffer, die nicht auf der ersten Trefferseite erscheinen, werden von den wenigsten Benutzenden angeschaut. Für jede Suchmaschine ist es also ein Muss, die relevantesten Dokumente ganz oben in der Rangliste aufzuführen. Nun wird keine Suchmaschine je genau wissen können, was für den Anfragesteller relevant und was irrelevant ist. Heutige Suchsysteme verfügen aber über ausgeklügelte Verfahren zum Erstellen der Ranglisten. Das Zauberwort heisst “relevance ranking”. Damit meint man das Anordnen von Dokumenten gemäss absteigender Relevanz bezüglich einer Suchanfrage.

Relevance Ranking läuft in zwei Schritten ab: Nachdem die Benutzerin eine Suchanfrage gestellt hat, werden alle verfügbaren Dokumente mit der Anfrage verglichen. Bei diesem Vergleich entsteht für jedes Dokument ein Relevanzwert. Je höher der Relevanzwert ausfällt, desto wahrscheinlicher stuft das Suchsystem das Dokument bezüglich der Anfrage als relevant ein. Nun sortiert das Suchsystem die Dokumente aufgrund des Relevanzwertes in absteigender Reihenfolge. Die so entstehende geordnete Liste wird Rangliste genannt. Die Suchmaschine präsentiert die Rangliste der Benutzerin, die je nach ihrem Bedürfnis wenige oder viele Dokumente daraus auswählt und genauer betrachtet.

Man kann sich zwei vollkommen unterschiedliche Bedürfnisse bei einer Recherche vorstellen: Die Physikerin auf der Suche nach dem Zahlenwert von Pi auf 40 Stellen genau gibt sich mit einem einzigen relevanten Dokument zufrieden. Sie ist an einer hohen Präzision interessiert. Ein Patentanwalt hingegen muss abklären, ob für eine neue Erfindung bereits ein Patent existiert. Deshalb möchte er natürlich möglichst alle relevanten Dokumente auffinden, die ähnliche Erfindungen beschreiben. Er ist an einer möglichst hohen Ausbeute interessiert. Er wird also einen grösseren Teil der Rangliste in Betracht ziehen als die Physikerin. Durch das Sortieren der Dokumente in der Rangliste gemäss ihrer Relevanzwerte wird diesen zwei völlig entgegengesetzten Bedürfnissen gleichzeitig Rechnung getragen.

Wie geht nun ein Suchsystem konkret vor, um die Relevanz eines Dokuments bezüglich einer Anfrage zu berechnen? Im Wesentlichen kommen hier zwei Arten von Kriterien zum Zuge:

### Dokumentenbasierte Rangierungsregeln

Diese Regeln basieren auf einer wichtigen Annahme: Die Vorkommen von Suchbegriffen in einem Dokument geben Hinweise auf die Relevanz dieses Dokuments. Diese Annahme bildet die theoretische Grundlage für wissenschaftliche Modelle zur Berechnung der Relevanz. Die teilweise komplexen mathematischen Hintergründe sollen uns hier nicht interessieren. Wir illustrieren drei wichtige Rangierungsregeln, die bei Soekia zum Einsatz kommen und auch von den meisten Suchmaschinen genutzt werden. Die Gewichtung der einzelnen Rangierungsregeln kann bei Soekia mittels eines Schiebereglers vorgenommen werden. Für erste Experimente mit der Dokumentenkollektion “Ozonschicht” empfiehlt es sich, alle drei Schieberegler ganz links zu setzen und die folgenden Beispiele ohne Wortstammreduktion zu erproben.

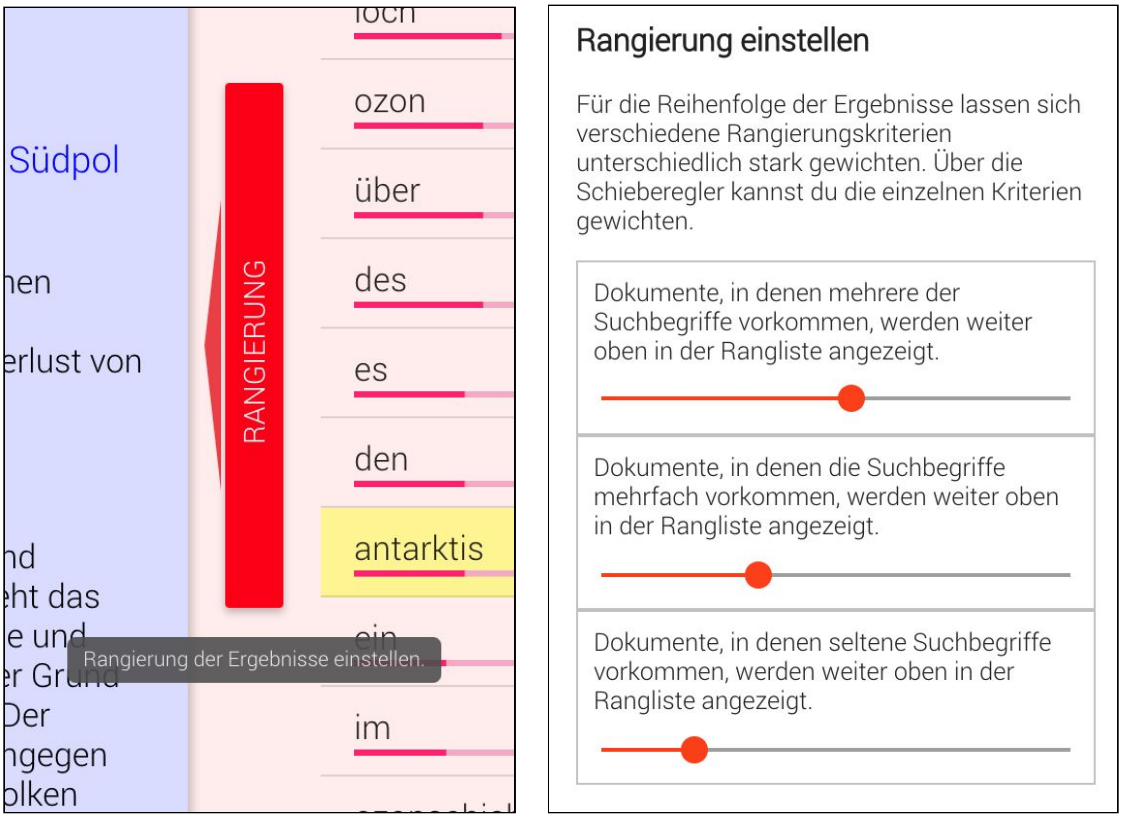
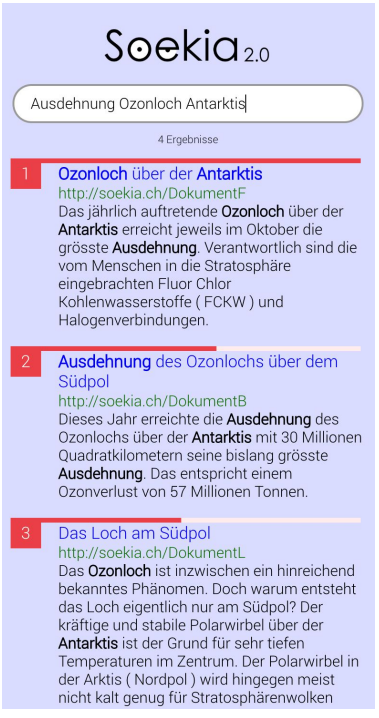
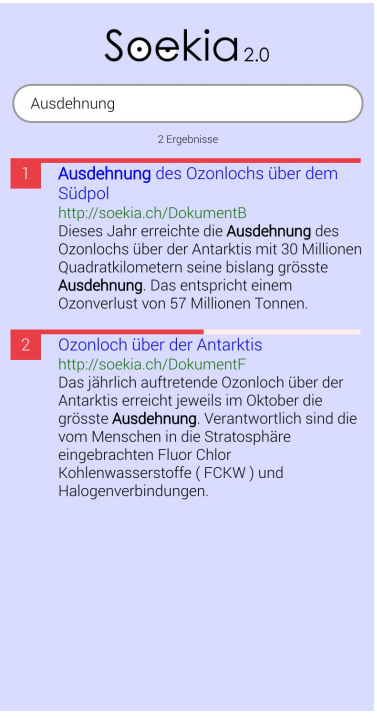
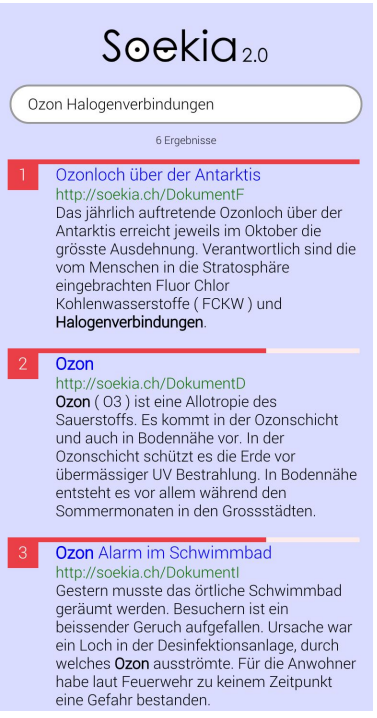



Abb. 7: Die drei Rangierungsregeln in Soekia

| | | |
|---|---|---|
| **Rangierungsprinzip 1:**<br>Dokumente, in denen mehrere der Suchbegriffe vorkommen, werden weiter oben in der Rangliste angezeigt. | **Rangierungsprinzip 2:**<br>Dokumente, in den die Suchbegriffe mehrfach vorkommen, werden weiter oben in der Rangliste angezeigt. | **Rangierungsprinzip 3:**<br>Dokumente, in denen seltene Suchbegriffe vorkommen, werden weiter oben in der Rangliste angezeigt. |
| **Beispiel Suchanfrage:**<br>“Ausdehnung Ozonloch Antarktis”<br><br>Die Dokumente, in denen alle drei Suchbegriffe vorkommen, werden weiter oben angezeigt. | **Beispiel Suchanfrage:**<br>“Ausdehnung”<br><br>Beim ersten Treffer kommt der Suchbegriff dreimal vor, im zweiten Treffer nur einmal. | **Beispiel Suchanfrage:**<br>“Ozon Halogenverbindungen”<br><br>In der Dokumentenkollektion kommt der Begriff “Ozon” häufig vor, der Begriff “Halogenverbindung” nur einmal. Es handelt sich deshalb um einen aussagekräftigeren Suchbegriff (sog. Diskriminator). |
|  |  |  |
| **Resultierender Tipp:**<br>Will man möglichst relevante Treffer erhalten, lohnt es sich, bei einer Suchanfrage mehrere Suchbegriffe zu verwenden. | **Resultierender Tipp:**<br>Für die Wahl guter Suchbegriffe ist es eine gute Strategie, sich die gesuchten Dokumente vorzustellen und darin Wörter zu bestimmen, die in den gesuchten Dokumenten mit grosser Wahrscheinlich-keiten häufig vorkommen. | **Resultierender Tipp:**<br>Die bewusste Verwendung von allgemein seltenen, aber thematisch passenden Begriffen kann bei der Suche helfen, mehr relevante Dokumente zu finden. |

### Nicht dokumentenbasierte Rangierungsregeln

Die oben beschriebenen drei Rangierungsprinzipien werden von den meisten Suchmaschinen berücksichtigt. Daneben gibt es noch unzählige andere Kriterien, die von Suchmaschine zu Suchmaschine variieren. Mit Soekia nicht illustrieren lassen sich die dokumentenunabhängigen Rangierungsprinzipen, etwa der von Google verwendete PageRank-Algorithmus, der Webseiten auch aufgrund ihrer Popularität gewichtet. PageRank nutzt -stark vereinfacht beschrieben - wie eine Webseite mit anderen Webseiten verknüpft ist. Im wesentlichen interpretiert Google einen Link von Seite A zu Seite B als eine Stimme von Seite A für Seite B. Google beurteilt die Wichtigkeit einer Seite also nach der Anzahl der abgegebenen Stimmen. Google berücksichtigt jedoch nicht nur die Anzahl der Stimmen bzw. Links, sondern analysiert auch die Seite, von der die Stimme ausgeht. Stimmen von Seiten, die selbst als "wichtig" eingestuft werden, haben eine grössere Bedeutung bei der Bewertung der Wichtigkeit anderer Seiten. Heute verwendet Google nicht mehr den ursprünglichen PageRank-Algorithmus, der den Grundstein für die Erfolgsgeschichte von Google legte. Sehr schön zeigt sich das Prinzip des PageRank bei den Seiten der Wikipedia. Bei vielen Suchanfragen erscheinen die entsprechenden Wikipedia-Artikel weit oben in der Rangliste. Zum einen verlinken viele Seiten auf Wikipedia, zum anderen handelt es sich bei diesen Seiten oft um Informationen von grossen Medienhäusern, die selbst prominent im Web vertreten sind.

Eine Suchmaschine wie Google bezieht noch viele weitere Kriterien zur Rangierung der Treffern ein. Ein Beispiel ist der aktuelle Standort (soweit bekannt) der Person, die eine Suchanfrage stellt. Sucht man etwa nach “Pizza”, werden Webseiten von umliegenden italienischen Restaurants weit oben in der Rangliste angezeigt. Ebenso spielen die bisherigen Suchanfragen eines bei Google angemeldeten Benutzers eine Rolle. So entsteht über die Zeit ein individuelles Profil der Benutzer, weshalb die gleiche Suchanfrage bei unterschiedlichen Nutzenden unterschiedliche Ergebnisse liefern kann.
Eine Aufforderung wie “Gib … ein und schau dir mal den dritten Treffer an” macht deshalb im Unterricht oder gegenüber anderen Personen meist wenig Sinn.

## 6. Qualität der Suche: Ausbeute versus Präzision

Hat man eine Suche erfolgreich abgeschlossen, sollte man sich auch über die Qualität der Antworten Rechenschaft ablegen. Bei offenen Fragestellungen, also im Zusammenhang mit umfangreichen Recherchen, spielt die Ausbeute eine grosse Rolle. Wie viele der in der Dokumentenkollektion zur gestellten Anfrage relevanten Dokumente wurden wirklich gefunden? Bei den grossen Suchmaschinen kann keine Aussage zur Ausbeute einer Suche gemacht werden. Man weiss nie, ob nicht noch weitere relevante Dokumente vorhanden wären. Bei Soekia ist aber die Dokumentenkollektion vorgegeben und kontrolliert. Man kann also zu einer Fragestellung in einem Experiment beispielsweise fünf relevante Dokumente in der Dokumentenkollektion "verstecken" und dann beobachten, wie viele Dokumente die Schülerinnen und Schüler finden. Damit kann relativ einfach das Bewusstsein für den Trade-Off zwischen Ausbeute und Präzision geschärft werden.

# 7. Weiterführende Ideen für den Unterricht

**n-Gram Suche:**
Jede Normalisierung hat aber auch ihre Grenzen. So findet sich zum Beispiel in der Beispiel-Kollektion “Ozonschicht” mit der Suchanfrage “Kopfweh” kein Treffer, obwohl es ein durchaus relevantes Dokument gibt, das aber den Begriff “Kopfschmerzen” enthält. Suchmaschinen bieten auch dazu Lösungen an. Eine Methode ist die sogenannte n-Gram-Suche. Vereinfacht gesagt werden zum Beispiel bei der 4-Gram-Suche in den Index alle Teilwörter der Länge “4 Buchstaben” aufgenommen. Der Index wird dadurch natürlich viel länger, aber mit der Suchanfrage “Kopfschmerzen” würde man auch Dokumente finden, die den Begriff “Kopfweh” enthalten. Dazu würde man aber auch alle Dokumente finden, die das Teilwort “schm” enthalten würden, also auch Dokumente zu Schmerikon, Schmid, Schmetterling, Schmuck und viele mehr. Suchmaschinen wie Google nutzen deshalb keine n-Gram-Suche. Die n-Gram-Suche wird in der Regel meistens nur auf kleinen Dokumentenkollektionen genutzt. Google stellt mit dem Ngram Viewer (https://books.google.com/ngrams) eine Suchmaschine auf Buchinhalten zur Verfügung, die interessante Experimente mit der n-Gram-Suche ermöglicht.

**Wortzerlegung und Synonyme:**
Für den Sprachunterricht von Interesse sind Suchmaschinen, die Suchanfragen und Dokumente nicht nur aufgrund der Übereinstimmung von Wortstämmen ermitteln, sondern auch Synonyme berücksichtigen. Bei einer Suchanfrage mit dem Suchbegriff “Rabatt” würden automatisch auch Dokumente mit den Begriffen “Preisnachlass”, “Vergünstigung” etc. gesucht. Sehr schön experimentieren lässt sich mit einer solchen Suche auf den Leitentscheiden des Schweizer Bundesgerichts ab 1954 (https://www.bger.ch/ext/eurospider/live/de/php/clir/http/index.php?lang=de). Hier werden in den als Treffer angezeigten Gerichtsentscheiden alle Begriffe gelb hervorgehoben, auf deren Grund ein Gerichtsentscheid als Treffer angezeigt wird. In Form von entdeckendem Lernen können die Schülerinnen und Schüler selbst versuchen herauszufinden, welche Methoden diese Suchmaschine anwendet. Besonders spannend sind dabei Suchbegriffe wie “Sekundar-schulanlage”, die zu auf den ersten Blick völlig irrelevanten Treffern führen, bedingt durch die Wortzerlegung von hinten: lage, ulan, arsch, sekunde ;-) Oder die Schülerinnen und Schüler können selbst versuchen, komplexe Rechtsfragen zu beantworten, etwa ob man ohne Besitz eines Fahrausweises einen Personenwagen umparkieren darf, wenn man den Motor nicht anstellt.