Arbeiten mit SOEKIA

- Wie kommt eine Suchmaschine von Dokumenten zu Tabellen? -

1. Kennzeichne alle Wörter im Text farbig, die du mit einem Hashtag versehen würdest!

FACK JU GÖTHE Kleinganove Zeki Müller [...] landet bei der Suche nach seiner Diebesbeute als Aushilfslehrer an einer Schule. Den Lehrerberuf führt er laut eigener Aussage nur nebenberuflich aus und das merkt man schnell: Er bedient sich unkonventioneller Methoden, wie beispielsweise seiner an Schülern erprobten Paintball-Pädagogik, und hat auch sonst keinen blassen Schimmer von den Unterrichtsthemen. Als Neuer an der Schule bekommt er gleich die Problemklasse aufs Auge gedrückt. Mit seinen rabiaten Mitteln und ungewöhnlichen Lehrmethoden mischt er die Chaosklasse und auch die Lehrerschaft ordentlich auf. Und schließlich ist da noch die Referendarin Lisi Schnabelstedt [...], die ihm nicht nur dank ihrer pädagogischen Ratschläge etwas bedeutet... Zeki muss sich entscheiden, ob er die Chance auf ein anständiges Leben und die große Liebe ergreifen will. Quelle: http://www.filmstarts.de/kritiken/209260.html, zuletzt zugegriffen am 19.09.2017	FACK JU GÖTHE Zeki Müller [...] ist ein Prolet, wie er im Buche steht. Große Klappe, kein Respekt und aggressives Auftreten. Eine Schulklasse zu unterrichten ist das Letzte, was man sich bei ihm vorstellen kann. Doch genau damit hat der neue Aushilfslehrer nach anfänglichen Schwierigkeiten Erfolg und schafft es sogar, die Chaosklasse 10b nach seiner Pfeife tanzen zu lassen. Das beeindruckt vor allem die vorbildliche Referendarin Lisi Schnabelstedt [...], die seiner Art eigentlich rein gar nichts abgewinnen kann. Als sie herausfindet, dass Zeki gerade erst frisch aus dem Gefängnis kommt, erklärt das für sie so einiges. Als Lisi dann auch noch erfährt, dass er nur an seine Beute unter der neu gebauten Turnhalle gelangen will, ist sie Zekis rauhem Charme jedoch längst erlegen. Zeki muss sich entscheiden, ob er das Geld oder die Liebe haben will. Quelle: http://www.moviepilot.de/movies/f-you-gothe, zuletzt zugegriffen am 19.09.2017

Im Internet lassen sich noch viele weitere Inhaltsangaben finden und Google hilft einem beim Aussuchen, denn Google zeigt die besten Angaben gleich auf der ersten Seite. Einfacher kann man es nicht haben. Doch welche ist die bessere Inhaltsangabe von den beiden Texten?

2. Bestimme die bessere Inhaltsangabe in Bezug auf
   i. mehr Informationen,
   ii. mehr Hashtags,
   iii. die Wörter „Klasse Lehrer Zeki Müller“ bei einer Suchanfrage!

Wie erkennt ein Informatiksystem die bessere Inhaltsangabe, wenn dieser nur Buchstaben, Wörter und Sätze erkennt, aber den Inhalt nicht versteht?

3. Zähle alle Wörter mit dem jeweiligen Wortstamm in den Inhaltsangaben, nutze dafür die Tabelle!

Wortstamm	Häufigkeit: FACK JU GÖTHE	Häufigkeit: FACK JU GÖTHE 2
Klasse		
Lehr		
Zeki		
Müller		
Inhaltsangabe		

Soekia ist eine didaktische Suchmaschine, die einen Einblick in die Funktionsweise erlaubt. So lässt sich beispielsweise die interne Datenstruktur, der Index, anzeigen.

Bei einer weiteren Suchanfrage kann nach anderen Wörtern gesucht werden. Bevor jetzt per Hand eine neue Tabelle erstellt wird, sollten ein Informatiksystem zum Zählen eingesetzt werden. Dieses kann gleich alle Wörter zählen und sie kategorisieren, diese dann entstehende Tabelle wird Index genannt.

4. Erstelle einen Index mit der Webseite „soekia.ch“!
   i. Du kannst Texte über Plus-Button hinzufügen, aber immer nur 500 Wörter mit einmal.
   ii. Durch drücken des Buttons oben rechts, wird der Index erzeugt.
   iii. Auf der rechten Seite kannst du alle Elemente des Index und deren Häufigkeit ablesen.

Eine Suchmaschine versteht unter einem **Dokument** nicht nur Textdateien, sondern auch Webseiten, E-Mails, Bücher, Zeitungsartikel, SMS, Tweets, Präsentationsfolien, PDFs, Patente, Chat-Sessions, Forumbeiträge, ... Die Gemeinsamkeit aller ist der hohe Anteil an Text, geschrieben in natürlicher Sprache.

Arbeiten mit SOEKIA

- Welchen Einfluss kann man auf die Indexerstellung nehmen? -

Ist Herr Müller ein klasse Lehrer? Die Antwort auf diese Frage ist SOEKIA uns noch schuldig.

1. Prüfe als Erstes wie häufig SOEKIA die einzelnen Worte gefunden hat.

Ist: Herr: Müller: ein: klasse: Lehrer

Einige von den Wörtern sind so nicht im Index zu finden, da Vereinfachungen in der Schreibweise vorgenommen worden sind. Diese Vereinfachungen heißen Buchstaben-Normalisierungen.
(Hinweis: Die Wortstamm-Reduktion muss aktiviert sein.)

2. Nenne vier Buchstaben-Normalisierung die SOEKIA durchführt!

___ → ___ / ___ → ___ / ___ → ___ / ___ → ___

Vier Dokumente wurden auf die Frage gefunden. Das zweite Dokument enthält die Anfragewörter „ist“ und „ein“ und wird daher als Treffer ausgegeben. Aber ist dieses Dokument relevant?

3. Nenne zehn Wörter aus dem Text, denen du keinen Hashtag geben würdest und trage sie als Stoppwörter ein!

Stoppwörter nennt man in der Informationsrückgewinnung Wörter, die bei einer Volltextindexierung nicht beachtet werden, da sie sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumenentinhalts besitzen.

Zipfsches Gesetz besagt, dass die Wahrscheinlichkeit eines Wortes umgekehrt proportional zum Platz in der geordneten Häufigkeitsliste ist.

$$p(n)=1/n.$$

Die Top 50 der englischen Wörter machen 40 % der Texte aus.

- Wettbewerb: Wortstamm-Reduktion -

Bei diesem Wettbewerb sollst du so viele verschiedene Wörter mit dem gleichen Wortstamm finden, sodass die Indexgröße bei SOEKIA „1“ beträgt.

. Erstelle ein neues Dokument und trage so viele Wörter wie möglich mit gleichem Wortstamm ein! Dein Ergebnis siehst du, indem du den Index erzeugst.

Gleichartige Dokumente enthalten viele Wörter mit gleichem Wortstamm, sodass beim Hinzufügen von gleichartigen Dokumenten zur Indexierung sich der Umfang des Index kaum ändert. Bei **fremdartigen Dokumenten** ist der Gegenteil der Fall, der Index wird dadurch deutlich größer.

Arbeiten mit SOEKIA

- Wie kommt eine Suchmaschine von Tabellen zu Ergebnislisten? -

Unabhängig wie präzise die Anfrage gestellt wird, es gibt immer mehr als ein Dokument, dass als Treffer gewertet werden kann. Auf welcher Grundlage erfolgt die Ordnung der Ergebnisliste von am meisten relevant zu am wenigsten relevant? Diese Grundlagen werden Rangierungsprinzipien genannt.

1. Ergänze die vier Satzanfänge zu den wichtigsten Rangierungsprinzipien.

   i. Ein Dokument ist relevanter, desto _____

   _____

   ii. Ein Dokument ist relevanter, desto _____

   _____

   iii. Ein Dokument ist relevanter, desto _____

   _____

   iv. Ein Dokument ist relevanter, desto _____

   _____

Nach diesen Prinzipien können Informatiksysteme bzw. Suchmaschinen ein Ranking erstellen.

2. Kreuze das relevantere Dokument für die folgenden drei Anfragen an! Begründe deine Entscheidung durch Angabe des verwendeten Rangierungsprinzips!

Anfrage: *Ozonloch Antarktis*	
Dokument 1	**Dokument 2**
Treibhauseffekt, Ozonloch, Überfischung der Weltmeere, Pinguinsterben in der Antarktis – die Menschheit zerstört ihren Planeten fortlaufend.	Im Winter entsteht über den Polkappen jeweils ein Ozonloch. Das Ozonloch über der Antarktis ist seit 1985 bekannt.
Anfrage: *Ausdehnung Ozonloch Antarktis*	
Dokument 1	**Dokument 2**
Das Ozonloch über der Antarktis erreichte im September 2001 eine Ausdehnung von 29 Millionen Quadratkilometern.	Die Ausdehnung des Packeises über der Antarktis beträgt im Winter über 20 Millionen Quadratkilometer.
Anfrage: *Geschichte Ozonloch Antarktis*	
Dokument 1	**Dokument 2**
Geschichte der Antarktis Am 16. Januar 1820 erreichten zum ersten Mal Menschen die Antarktis. Leiter der russischen Expedition war der baltische Deutsche Fabian Bellinghausen.	Ozonloch über der Antarktis Seit 1978 wird die Ozonschicht über der Südhalbkugel regelmäßig gemessen. Anzeichen für ein Ozonloch wurden aber erst 1985 ernst genommen

Die Qualität der Ergebnisliste hängt von der **Ausbeute** und der **Präzision** ab. D. h. die Ausbeute gibt an, wie viele Dokumente gefunden wurden und die Präzision gibt an, wie relevant sind diese in Bezug auf die inhaltliche Suchanfrage. Je höherer diese sind, desto besser ist das Ergebnis.