Drupal Camp Essen 2013

Deutsche und mehrsprachige Volltextsuche mit Apache Solr

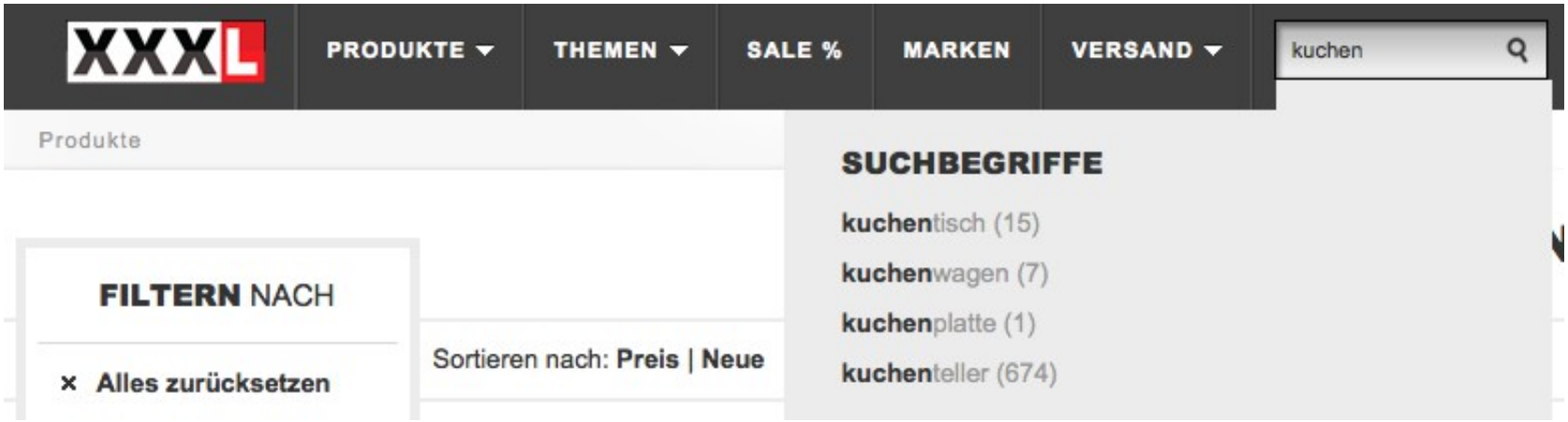
Markus Kalkbrenner (mkalkbrenner)

Frankfurt, 10/27/13

Volltextsuche ist super!

bio·logis

Noch besser mit Autovervollständigung



Autocomplete von „kuchen“: Kuchenplatte, Kuchenteller, …

PREISFRAGE: Was ist ein „Kuchenwagen“?

Kuchenwagen = Küchenwaage! WTF?

bio·logis

XXXL | PRODUKTE | THEMEN | SALE % | MARKEN | VERSAND | Ich suche… | (0) | (0) | (0)

Produkte

7 TREFFER FÜR „KUCHENWAGEN“

FILTERN NACH

× Alles zurücksetzen

SUCHE

kuchenwagen LOS

KATEGORIE

☐ Küchen, Essen, Haushalt (7)

FARBE

☐ Weiß (4)
☐ Grau/Silber (2)
☐ Rot (1)

PREIS

10 € bis 40 €

MARKE

☐ LEIFHEIT (1)

Sortieren nach: **Preis** | **Neue**

Produkte 1 bis 7

KÜCHENWAAGE

€ 19,99

Sale %

KÜCHENWAAGE

€ 34,99

KÜCHENWAAGE

€ 9,99

KÜCHENWAAGE

€ 24,95

KÜCHENWAAGE

KÜCHENWAAGE

KÜCHENWAAGE

22240 Kartoffel Rezepte vs. 7738 Kartoffeln Rezepte

97 Erdäpfel Rezepte

Ich bin gemein ;-)

“Apache Solr Search Integration“ is awesome!*

* ... Wenn man eine englischsprachige Webseite betreibt :-(

Out of the Box erzeugen die Drupalmodule „Apache Solr Search Integration“ und „Search API“ genau derartige Effekte auf nicht englischsprachigen Seiten.

Warum eigentlich?

Was ist das Problem?

bio·logis

- Stoppwörter (stop words)
- Wortstammbildung (stemming)
- Wortzusammensetzungen (compound words)
- Rechtschreibprüfung (spell checker)
- Synonymlisten (synonyms)
- …

=> Sämtliche Konfigurationen sind auf englischsprachige Inhalte ausgerichtet!

Etwas Theorie: Stoppwörter

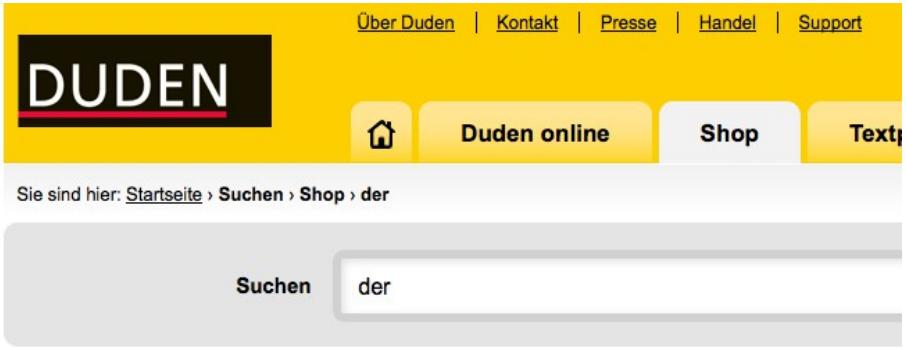
bio·logis

- Grundsätzlich eine gute Idee:
  - Wörter nicht in den Suchindex aufnehmen, die zu unspezifisch sind.
- Typische Stoppwörter sind z.B. bestimmte Artikel:
  - Englisch: „the“
  - Deutsch: „der“, „die“, „das“
- Stoppwörter unterscheiden sich von Sprache zu Sprache!

Auch die Domäne ist wichtig!

- Die Stoppwortliste hängt nicht nur von der Sprache, sondern auch vom Inhalt und dem Zielpublikum ab, der so genannten Domäne.
- Auf duden.de ist „der“ korrekterweise kein Stoppwort.

Getrennte Suchindexe für jede Domäne wären besser!



Suchergebnisse

Die Suche nach **"der"** lieferte 135019 Treffer.

Duden online (133765) | **Shop (443)** | **Sprachwissen (811)**

Duden - Lexikon der Familiennamen

Mit bekannten Namensträgerinnen und -trägern.

Sofort lieferbar

12.95 €

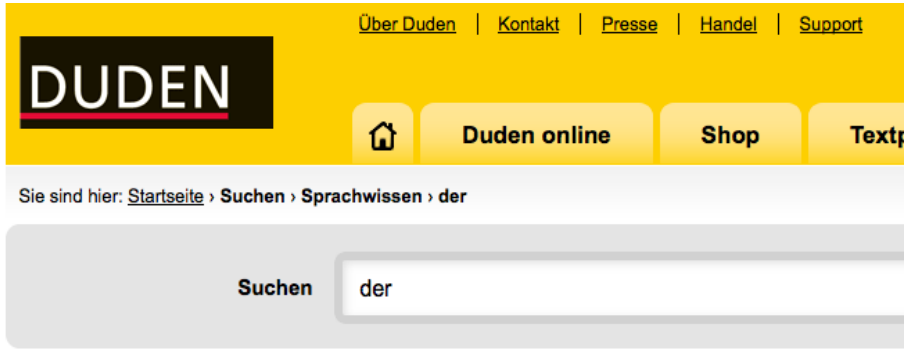
Duden - Lexikon der Vornamen

5., völlig neu bearbeitete Auflage

Sofort lieferbar

8.99 €

Duden - Lexikon der Vornamen



Suchergebnisse

Die Suche nach **"der"** lieferte 135019 Treffer.

Duden online (133765) | **Shop (443)** | **Sprachwissen (811)**

Sprachratgeber: Der Anlagenvermerk

Für alle, die an einer soliden Anlagenberatung interessiert sind, haben wir die wichtigsten Informationen zusammengestellt. Erfahren Sie, wie **der** Anlagenvermerk in **der** geschäftlichen Korrespondenz gestaltet werden sollte. Kleinere Gegenstände oder schriftliche Unterlagen wie Zeugnisse, Arbeitsproben oder Formulare, die Briefen beigefügt sind, werden Anlagen genannt. Nach den Empfehlungen **der** DIN 5008 steht **der** Anlagenvermerk, also **der** Hinweis, dass ein Schreiben eine Beilage enthält, mit mindestens drei Leerzeilen Abstand unter dem Gruß oder **der** Firmenbezeichnung. Falls **der** Unterzeichner maschinenschriftlich angegeben wird, schließt sich **der** Anlagenvermerk nach einer Leerzeile an. Ist wenig Platz vorhanden,

Aktuelle Meldung: Journalistenpreis der Hauptstadt der deutschen Sprache

Mama, stirb?

Falsche Freunde (False Friends)

bio·logis

- Identische Wörter haben in verschiedenen Sprachen unterschiedliche Bedeutungen, z.B. „die“:
  - Deutsch: „die“
  - Englisch: „sterben“, aber auch „Würfel“ oder „Chip“ usw.
  - Niederländisch: „jene“
  - …

    => Das nennt man „False Friends“.
- Identische „Buchstabenkombinationen“ sind in einer Sprache Stoppwörter, in einer anderen nicht!
- Eine einzelne große sprachübergreifende Stoppwortliste ist nicht möglich!

Wortstammbildung (Stemming)

bio·logis

- Die Reduktion eines Suchbegriffs auf seinen Wortstamm ermöglicht das Finden von passenden Inhalten unabhängig von der Flexion des Suchbegriffs, z.B. Singular oder Plural:
  - `Tomate  => tomat`
  - `Tomaten  => tomat`
- Aber: Die Algorithmen zur Wortstammbildung unterscheiden sich von Sprache zu Sprache!
- Für einige Sprachen gibt es gar keinen Algorithmus!
- Der englische „Default Stemmer“ führt bei allen anderen Sprachen zu Fehlern!

Algorithmus zur Wortstammbildung

bio·logis

- Algorithmus für deutsche Sprache:
  - `Tomate   => tomat`
  - `Tomaten  => tomat`
  - `tomato   => tomato`
  - `tomatoes => tomato`
- Algorithmus für englische Sprache:
  - `tomato   => tomato`
  - `tomatoes => tomato`
  - `Tomate   => tomat`
  - `Tomaten  => tomaten`

Synonyme

bio·logis

- Wortstammbildung hilft nicht immer.
  - Deutsch:

    `Kartoffel, Kartoffeln`
  - Englisch:

    `goose, geese`

    `mouse, mice`

  => Derartige Pluralformen müssen in sprachspezifische Synonymlisten eingetragen werden!

Protected Words

bio·logis

- In Abhängigkeit von der Domäne der Webseite kann es sinnvoll sein, einzelne Wörter von der Wortstammbildung oder der Anwendung von Stoppwörtern aus zu schließen.
- Produkt- oder Markennamen beinhalten oft einen Plural oder ein Stoppwort:
  - Drupal Gardens
  - Die Ärzte
  - The Who
  - Pittsburgh Steelers

Wortzusammensetzungen (Compound Words)

bio·logis

- Typisch für die deutsche Sprache: Wörter aus zusammengesetzten Substantiven, z.B. „Dampfschifffahrt“.
- Je nach Anwendungsfall wünscht man sich einen Suchtreffer bei folgenden Suchbegriffen:
  - Dampf
  - Dampfschiff
  - Schiff
  - Schifffahrt
  - Fahrt
- Solr bietet dafür den CompoundWordFilter. Die Standardkonfiguration der Drupalmodule unterstützt diesen aber nicht!

Rechtschreibprüfung (Spell Checker)

bio·logis

- Begriff „Rechtschreibprüfung“ ist eigentlich falsch, besser wäre „Wortkatalog“
- Ein Katalog ist die Grundlage für verschiede Features:
  - „Did you mean ...“
  - Autovervollständigung
- Diese Kataloge sind fraglos sprachabhängig!
- Die Rechtschreibprüfung berücksichtigt Stoppwörter!
  => Vermutlich ist eine dedizierte Stoppwortliste sinnvoll
- Kataloge können statische Listen sein. Default bei Drupal ist eine dynamische Generierung aus dem Content!
  => Das kann bei User-generated-Content peinlich werden ;-)

ISOLatin1Accent Filter

bio·logis

- In der Drupalstandardkonfiguration wird z.B. aus einem „Ä“ ein „A“:
  - Stoppwörter mit Umlauten funktionieren nicht!
  - Die Wortstammbildung verursacht „Fehler“:
    - Kuchen => kuch
    - Küche => kuch
    - Küchen => kuch
- Abschalten des Filters ist keine Lösung, sonst bekommt man andere Probleme, z.B. mit „Crème fraîche“
- Besser: In jeder Sprache die „eigenen“ diakritischen Zeichen gesondert behandeln und die „fremden“ verwerfen.

War es das jetzt endlich?

- Nein, es ist noch (viel) komplexer:
  - Word Delimiter
  - Length Filter
  - Lower Case
  - Elevator
  - Ranking und Boosts
  - Multisite
  - N-Gram
  - …
- Und mehrsprachige Webseiten?

Drei Optionen

bio·logis

- Die Solr-Dokumentation oder ein Buch lesen und den Index perfekt auf den eigenen Anwendungsfall hin konfigurieren.
- Jemanden beauftragen, der sich damit auskennt ;-)
- Mithilfe der folgenden Dupalmodule brauchbare Ergebnisse Out-of-the-box erzielen und die wichtigsten Einstellungen innerhalb von Drupal verfügbar haben:
  - Apache Solr Config Generator
    - Inkl. Apache Solr Advanced Settings
    - Inkl. Apache Solr Text Files
  - Apache Solr Multilingual
    - Apache Solr Multilingual Config Generator

Live-Demo der Module

bio·logis

- Übrigens: Ohne Apache Solr Multilingual kann die Apache Solr Search Integration Inhalte, die per Entity Translation übersetzt werden, gar nicht verarbeiten!

Danke

bio·logis

- drupal.org:
  - mkalkbrenner: https://drupal.org/user/124705
  - https://drupal.org/project/apachesolr_multilingual
  - https://drupal.org/project/apachesolr_confgen
- twitter: https://twitter.com/mkalkbrenner
- github: https://github.com/mkalkbrenner
- Apache Solr Multilingual in Action:
  - https://pgsbox.de/
  - https://pgsbox.com/
  - https://pgsbox.nl/