Meinten Sie....?

Wie Suchmaschinen mit Hilfe des Levenshtein-Algorithmus und Soundex ähnliche Wörter finden.

24.11.2010, Christiane Olschewski

Kontakt: Olschewski@LaminARt.de , Olschewski@Olschewski.org

Download: olschewski.org

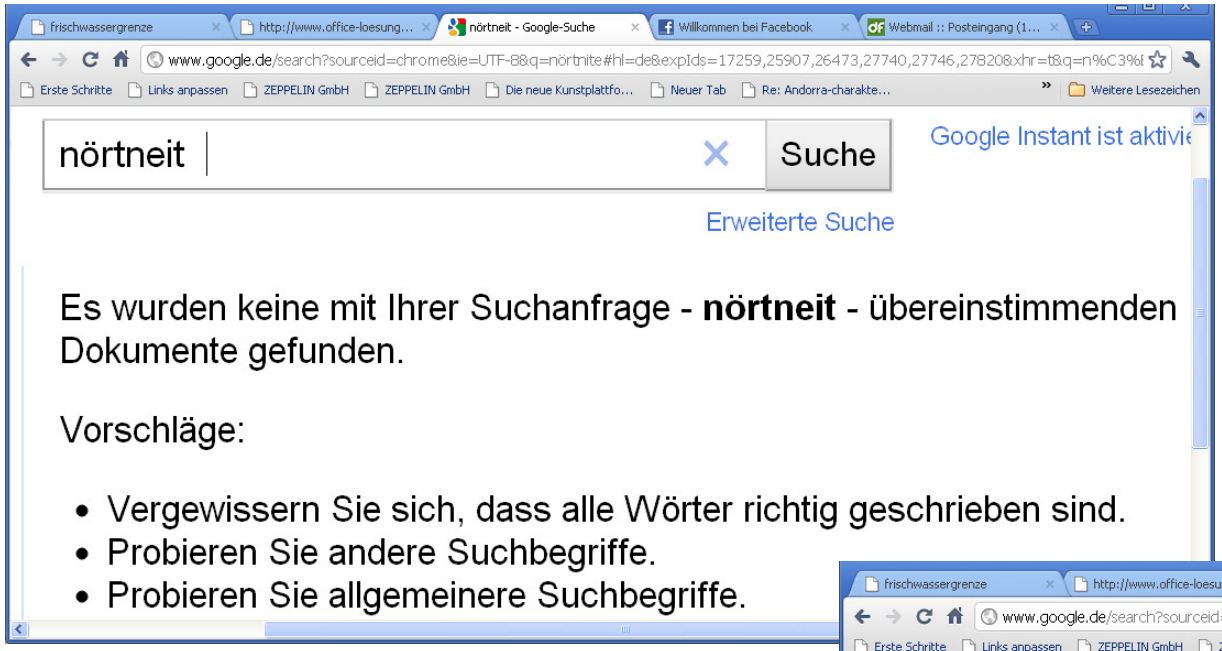
Menschen können meist ohne Probleme falsch geschriebene Wörter und Texte erkennen:

Afugrnud enier Sduite an enier Elingshcen Unvirestiät ist es eagl, in wlehcer Rienhnelfoge die Bcuhtsbaen in eniem Wrot sethen, das enizg wcihitge dbaei ist, dsas der estre und lzete Bcuhtsbae am rcihgiten Paltz snid. Der Rset knan ttolaer Bölsdinn sien, und du knasnt es torztedm onhe Porbelme lseen. Das ghet dseahlb, wiel wir nchit Bcuhtsbae für Bcuhtsbae enizlen lseen, snodren Wröetr als Gnaezs. Smtimt's?

Aufgrund einer Studie an einer Englischen Universität ist es egal, in welcher Reihenfolge die Buchstaben in einem Wort stehen, das einzig wichtige dabei ist, dass der erste und letzte Buchstabe am richtigen Platz sind. Der Rest kann totaler Blödsinn sein, und du kannst es trotzdem ohne Probleme lesen. Das geht deshalb, weil wir nicht Buchstabe für Buchstabe einzeln lesen, sondern Wörter als Ganzes. Stimmt's?

Quelle: http://www.heise.de/tp/r4/artikel/15/15701/1.html

Auch Suchmaschinen tolerieren so manche Rechtschreib-, und Tippfehler:



Tippfehler (Buchstaben doppelt, ausgelassen, vertauscht) Nerd Nite

Es gibt zwei grundlegende Verfahren, um ein fehlertolerantes Suchen oder Vergleichen zu ermöglichen:

- buchstabenorientierte Verfahren prüfen, inwieweit sich die **Schreibweise** von Wörtern unterscheidet und gewichtet die Anzahl / Art der Unterschiede (**bei erd nite fehlt nur 1 Buchstabe**)

  → Levenshtein-Distanz

- sprachorientierte Verfahren berücksichtigen die **phonetische Ähnlichkeit** von Wörtern (**nört neit klingt wie nerd nite**)

  → phonetische Algorithmen wie Soundex, Kölner Phonetik

Der Levenshtein-Algorithmus

Der Levenshtein-Algorithmus errechnet die Mindestanzahl von Einfüge-, Lösch-, und Ersetz-Operationen um eine bestimmte Zeichenkette in eine andere umzuwandeln.

(Wie viele Buchstaben/Zeichen muss man austauschen, löschen oder einfügen, um aus einem Wort ein anderes zu machen?)

(Anzahl = **Levenshtein-Distanz**)

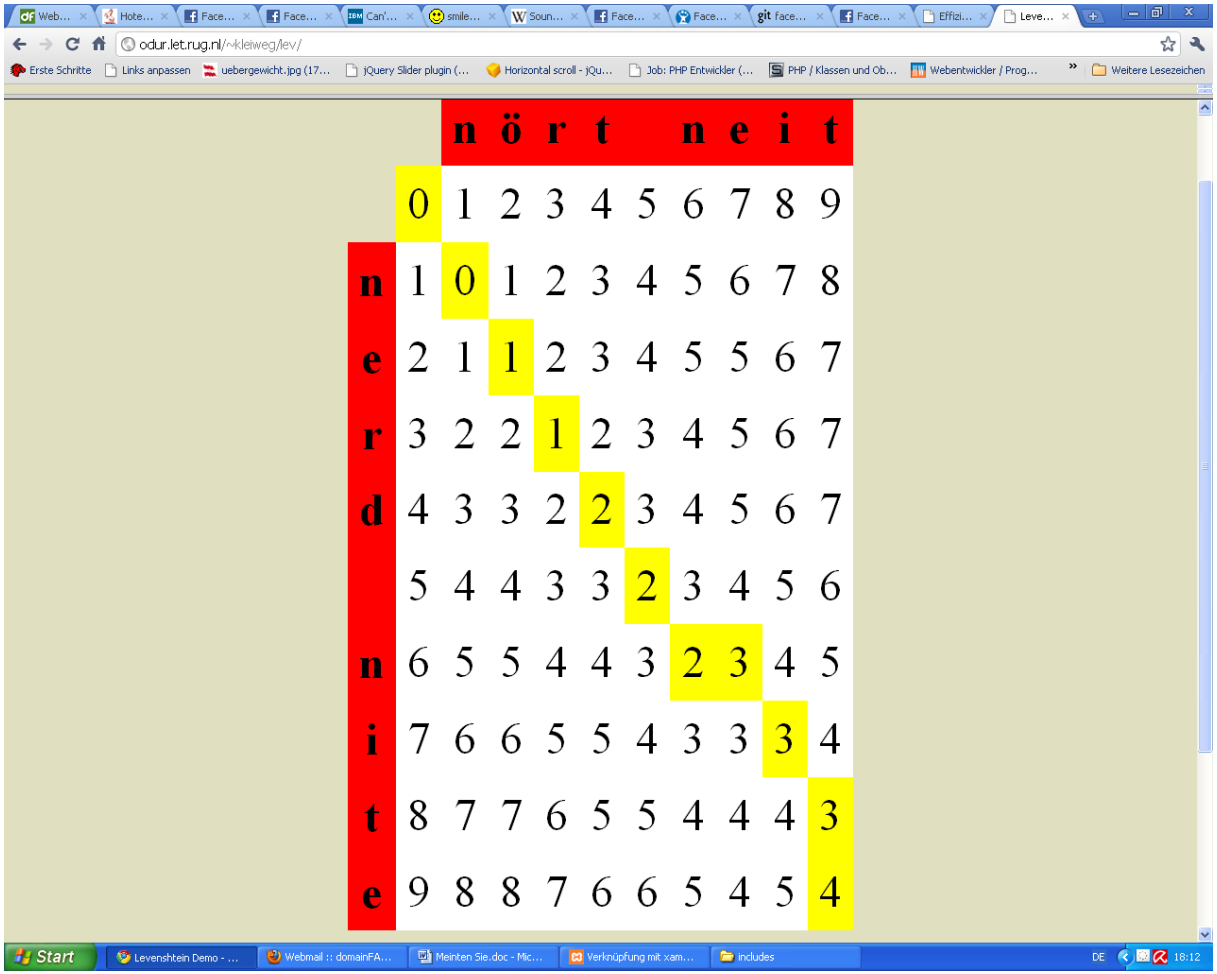
Ein **Algorithmus** ist eine eindeutige Anleitung zur Lösung eines Problems in mehreren Schritten, hier: die geringstmögliche Anzahl von Änderungen zu finden.

Nach dem Prinzip der dynamischen Programmierung wird die optimale Lösung von Teilproblemen in einer Tabelle gespeichert und die Gesamtlösung Schritt für Schritt zusammengesetzt.

Eine Anordnung von Zahlenwerten in Tabellenform, mit denen man rechnen kann, nennt man in der Mathematik **Matrix.**

Wie viele Buchstaben muss man austauschen, löschen oder hinzufügen, um „nört neit“ in „nerd nite“ umzuwandeln?

Diese Matrix wurde nach dem Levensthein-Algorithmus von links oben nach rechts unten berechnet. Das Ergebnis ist eine Levenshtein-Distanz von 4.



		n	ö	r	t		n	e	i	t
	0	1	2	3	4	5	6	7	8	9
n	1	0	1	2	3	4	5	6	7	8
e	2	1	1	2	3	4	5	5	6	7
r	3	2	2	1	2	3	4	5	6	7
d	4	3	3	2	2	3	4	5	6	7
	5	4	4	3	3	2	3	4	5	6
n	6	5	5	4	4	3	2	3	4	5
i	7	6	6	5	5	4	3	3	3	4
t	8	7	7	6	5	5	4	4	4	3
e	9	8	8	7	6	6	5	4	5	4

http://odur.let.rug.nl/~kleiweg/lev/

Es gibt zahlreiche **Abwandlungen** des Levensthein-Algorithmus. Im oberen Beispiel werden alle Änderungen, ob Ersetzen, Löschen, Einfügen gleich bewertet (mit dem Wert 1).

Bei der **Gewichteten Levenshtein-Distanz** können die „Kosten“ der einzelnen Operationen unterschiedlich gewichtet werden, sogar abhängig von den beteiligten Zeichen.

Bei der Damerau-Levenshtein-Distanz gelten zwei vertauschte Zeichen als eine einzige Operation, bei Levensthein wären es 2. So kann man das Maß für Ähnlichkeit noch genauer bestimmen.

Der Levensthein-Algorithmus wurde von dem russischen Mathematiker **Wladimir Iossifowitsch Lewenstein** 1965 erfunden. http://www.uni-bielefeld.de/(en)/ZIF/FG/2002Combinatorics/02-reception.html

(Auf dem Foto: der Herr, der gerade ein Würstl bekommt.)

Seine Homepage: www.keldysh.ru

Der Soundex-Algorithmus

Das Soundex-Verfahren wurde in den USA als Hilfsmittel bei der Volkszählung von Margaret K. Odell und Robert C. Russell entwickelt und 1918 patentiert. Es wird hauptsächlich zum Auffinden von Namen benutzt.

Soundex ist ein phonetischer Algorithmus, der Wörtern nach ihrem (englischen) Sprachklang eine Zeichenfolge zuordnet, den phonetischen Code. Ziel dieses Verfahrens ist es, gleich klingenden Wörtern denselben Code zuzuordnen. Dadurch können sie leicht auf Gleichheit geprüft werden.

Allerdings besteht der Soundex-Code für ein Wort nur aus seinem **Anfangsbuchstaben und 3 Ziffern** für die Konsonanten. Der Rest wird entfernt bzw. abgeschnitten.

„nört neit“, „nerd night“, „nerd nite“ , aber auch z.B. „neuere Themen“ haben denselben Code, nämlich N635

„nört neit“, „nerd night“, „nerd nite“ , aber auch z.B. „neuere Themen“ werden zu N635

Konvertierungs-Algorithmus: (Grundregeln)

1. nimm den Anfangsbuchstaben
2. entferne die doppelten Konsonanten (ß wird s)
3. entferne alle Vokale sowie h,w,y und (im Deutschen) ä,ö,ü
4. ersetze die Konsonanten duch die entsprechende Ziffer (gleiche Ziffern dürfe nur aufeinanderfolgen, wenn die Konsonanten durch einen Vokal oder Y getrennt waren, sonst auslassen); nach 3 Ziffern ist Schluss
5. fülle gegebenenfalls mit 0 auf

Ziffer	Repräsentierte Buchstaben
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M,N
6	R

Alternativen: - Metaphone (genauer und variable Länge)

- Kölner Phonetik: ist besser auf die deutsche Sprache abgestimmt

(Jeder Buchstabe eines Wortes wird auf eine Ziffer zwischen "0" und "8" abgebildet, wobei für die Auswahl der jeweiligen Ziffer maximal ein benachbarter Buchstabe als Kontext benutzt wird .Die Länge des phonetischen Codes ist nicht beschränkt.)

Vergleich Soundex versus Levenshtein

Bei Verfahren wie Soundex werden Zeichenketten zunächst verschlüsselt, in Gruppen mit gleicher Kodierung eingeteilt und bei einer Suchanfrage nur auf Übereinstimmung der Kodierung geprüft.

Das geht schnell, aber leider fallen gut passende Begriffe durch das Raster, wenn beispielsweise der Anfangsbuchstabe nicht stimmt, andererseits tauchen auch komplett unsinnige Ergebnisse auf, weil gerade Soundex zu stark vereinfacht.

Innerhalb einer Gruppe mit tatsächlich oder vermeintlich gleich klingenden Begriffen fehlt auch ein zusätzliches Kriterium, um bessere und schlechtere Ergebnisse zu filtern und zu sortieren.

Dagegen lassen Verfahren wie der Levensthein-Algorithmus eine kontinuierliche Bewertung der Ähnlichkeit zweier Wörter zu, indem er einzelne Fehler zählt und bewertet, was allerdings erheblichen Rechenaufwand bedeutet.

Eine Kombination aus mehreren Verfahren bringt grundsätzlich die besten Ergebnisse.

Quellen:

http://www.heise.de/tp/r4/artikel/15/15701/1.html

http://www.smile-datentechnik.de/projekte/dipl/main_dipl_kapitel_02.html

http://sound-ex.de

http://www.levenshtein.de

http://de.wikipedia.org/wiki/Koelner_Phonetik

http://de.wikipedia.org/wiki/Levenshtein-Distanz

http://www.uni-bielefeld.de/(en)/ZIF/FG/2002Combinatorics/02-reception.html

http://www.functions-online.com/soundex.html

http://odur.let.rug.nl/~kleiweg/lev/