Semantik ist nur ein Hype ?

Wie berechnet man Semantik ? Gestern – Heute – Morgen

Auf der Suche nach dem semantischen Raum

Glaskugel

© 2004 – 2011 Matthias Schneider

Wer spricht da eigentlich ?

Semager

Semantic Business – semantische Datenbanken und Dienstleistungen.
(Keyword-Datenbanken, Konzept-Extraktionen, Kategorisierung von Webseiten, Texten oder einzelnen Wörtern, …)

www.semager.de = Showcase

Referenzen:
- T-Online
- Yahoo Partner
- Kelkoo Partner
- Herold.at
- netzwelt.de
- Domainparking Sites, …

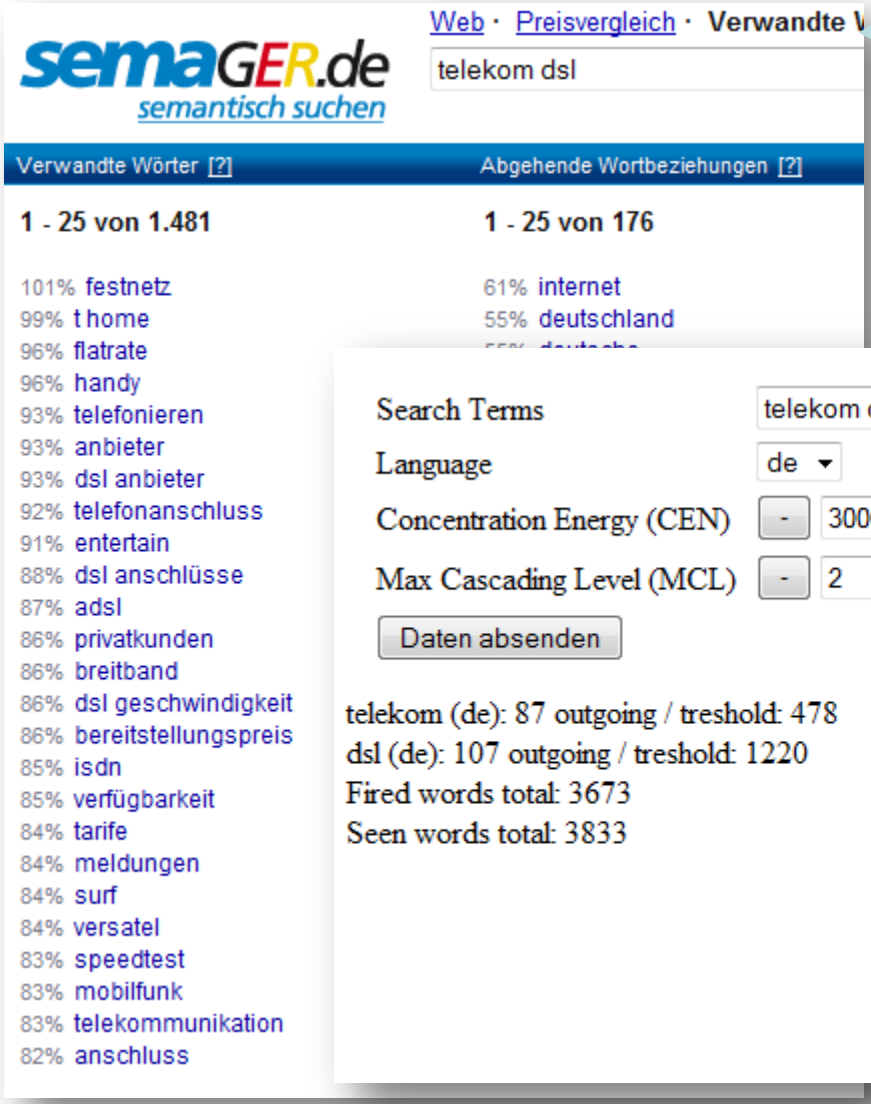
Matthias Schneider

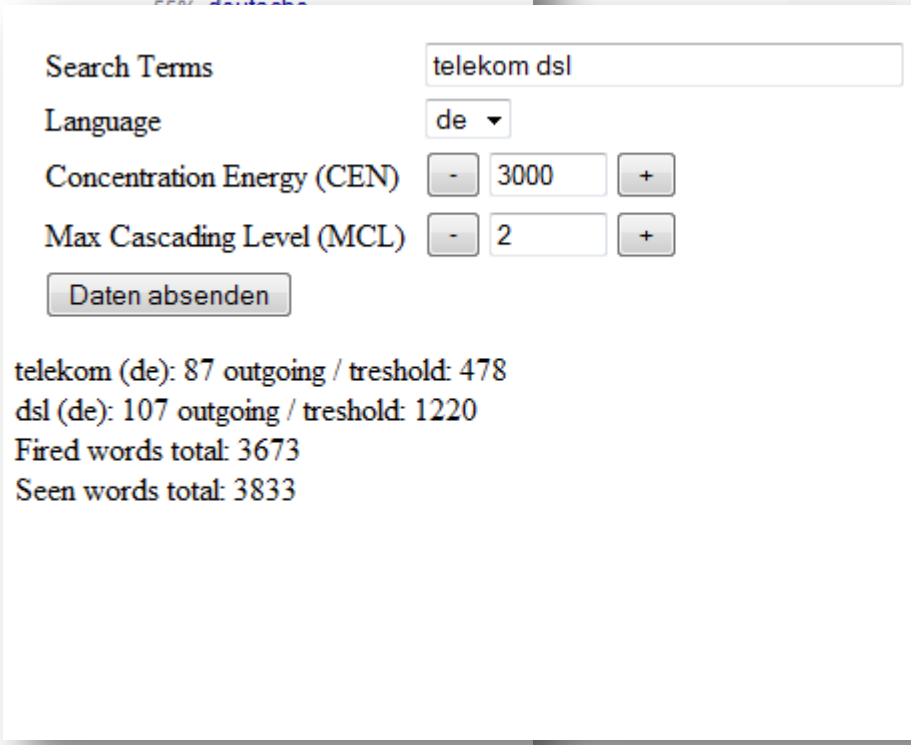
- Geschäftsführer Semager
- Referent auf der SES, SEMSEO, SEOCampixx, …
- Seminarleiter in verschiedenen Weiterbildungskursen und Handelskammern
- Gastautor in Fachzeitschriften

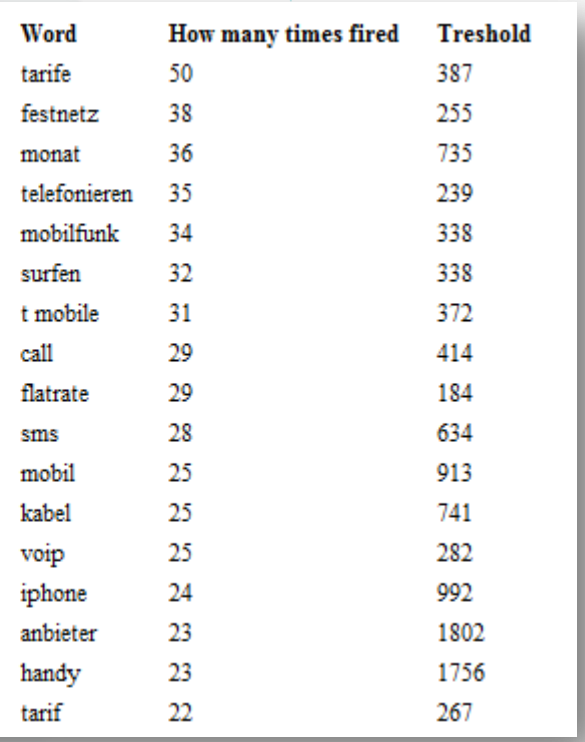
© 2004 – 2011 Matthias Schneider

Einfache Berechnung

Neuronale Berechnung







Word	How many times fired	Treshold
tarife	50	387
festnetz	38	255
monat	36	735
telefonieren	35	239
mobilfunk	34	338
surfen	32	338
t mobile	31	372
call	29	414
flatrate	29	184
sms	28	634
mobil	25	913
kabel	25	741
voip	25	282
iphone	24	992
anbieter	23	1802
handy	23	1756
tarif	22	267

© 2004 – 2011 Matthias Schneider

1) Der natürliche Sprachprozess

Stichwort: NLP (natural language processing), Computerlinguistik

- Korrektur von Tipp- und Rechtschreibfehlern
- Prüfung auf grammatische Richtigkeit
- Automatische Übersetzung.
- Verschlagwortung von Literatur
- Anfertigung von Registern und Inhaltsverzeichnissen
- Herstellung von Zusammenfassungen und Abstracts.
- Unterstützung von Autoren beim Verfassen von Texten

© 2004 – 2011 Matthias Schneider

2) Das verstehen des Web-Kontextes

- Homonyme - gleiches Wort kann je nach Kontext andere Bedeutung haben.
- Auflösung syntaktischer Mehrdeutigkeiten - ein Satz lässt sich auf mehrere Arten deuten.

Beispiel:

„Peter sah Maria mit dem Fernglas“

Hat Peter Maria gesehen hat, die

a)ein Fernglas in der Hand hielt, oder hat
b)Peter Maria mit Hilfe eines Fernglases gesehen ?

© 2004 – 2011 Matthias Schneider

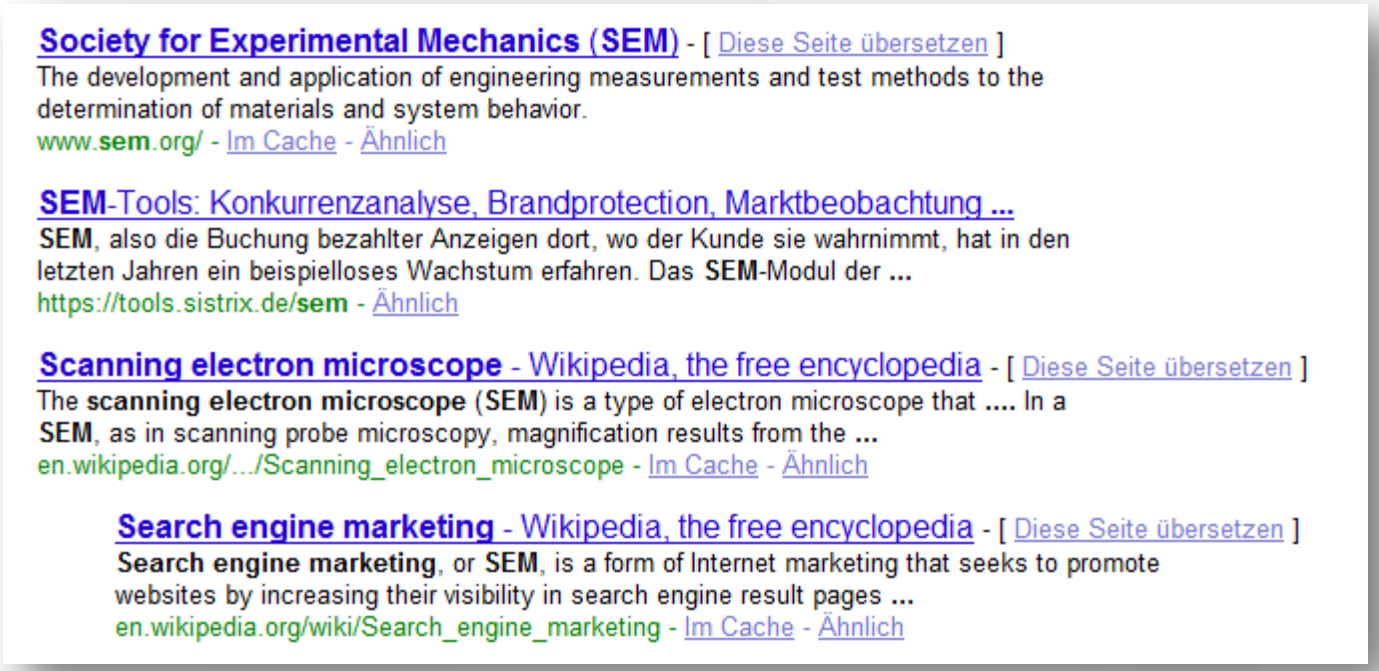
3) Das verstehen der Nutzerabsicht

Ein Suchender gibt das Wort „Kamera Canon EOS 450D“ ein.
Interessiert er sich nun für einen Preisvergleich oder einen Testbericht?

- Navigationsorientiert – der Nutzer sucht den Hersteller
- Informationsorientiert – der Nutzer sucht Testberichte
- Transaktionsorientiert – der Nutzer sucht Anbieter
- Evtl. auch weitere wie z.B. Ressourcen (Downloads) oder Media (Videos, Bilder)

→ Erkenntnis nutzen von Nutzerprofile, Standort, Suchhistorie, Klickhistorie

© 2004 – 2011 Matthias Schneider



Society for Experimental Mechanics (SEM) - [Diese Seite übersetzen]
The development and application of engineering measurements and test methods to the determination of materials and system behavior.
www.**sem**.org/ - Im Cache - Ähnlich

SEM-Tools: Konkurrenzanalyse, Brandprotection, Marktbeobachtung ...
SEM, also die Buchung bezahlter Anzeigen dort, wo der Kunde sie wahrnimmt, hat in den letzten Jahren ein beispielloses Wachstum erfahren. Das **SEM**-Modul der ...
https://tools.sistrix.de/**sem** - Ähnlich

Scanning electron microscope - Wikipedia, the free encyclopedia - [Diese Seite übersetzen]
The **scanning electron microscope** (**SEM**) is a type of electron microscope that .... In a **SEM**, as in scanning probe microscopy, magnification results from the ...
en.wikipedia.org/.../Scanning_electron_microscope - Im Cache - Ähnlich

Search engine marketing - Wikipedia, the free encyclopedia - [Diese Seite übersetzen]
Search engine marketing, or **SEM**, is a form of Internet marketing that seeks to promote websites by increasing their visibility in search engine result pages ...
en.wikipedia.org/wiki/Search_engine_marketing - Im Cache - Ähnlich

SEM = Society for Experimental Mechanics
= Scanning electron microscope
= Search engine marketing

© 2004 – 2011 Matthias Schneider

Lehrstellen = Lehrstellenmarkt

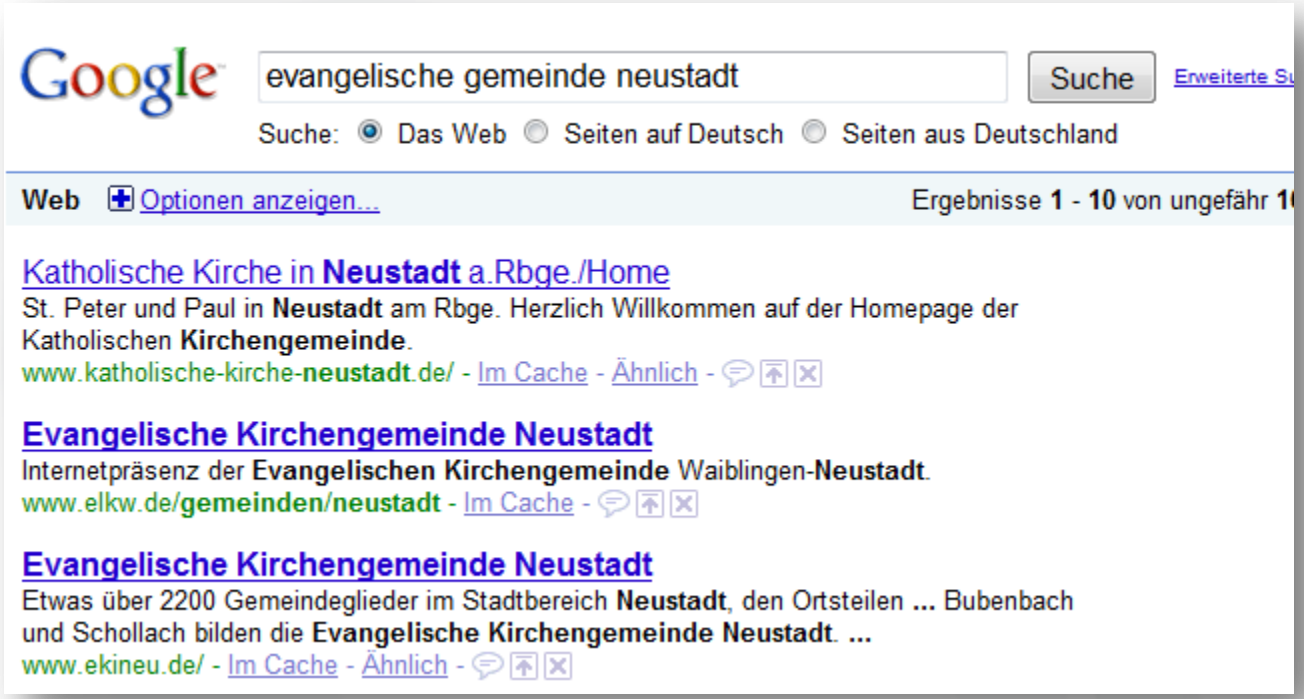
© 2004 – 2011 Matthias Schneider

Stellensuche = Stellengesuche

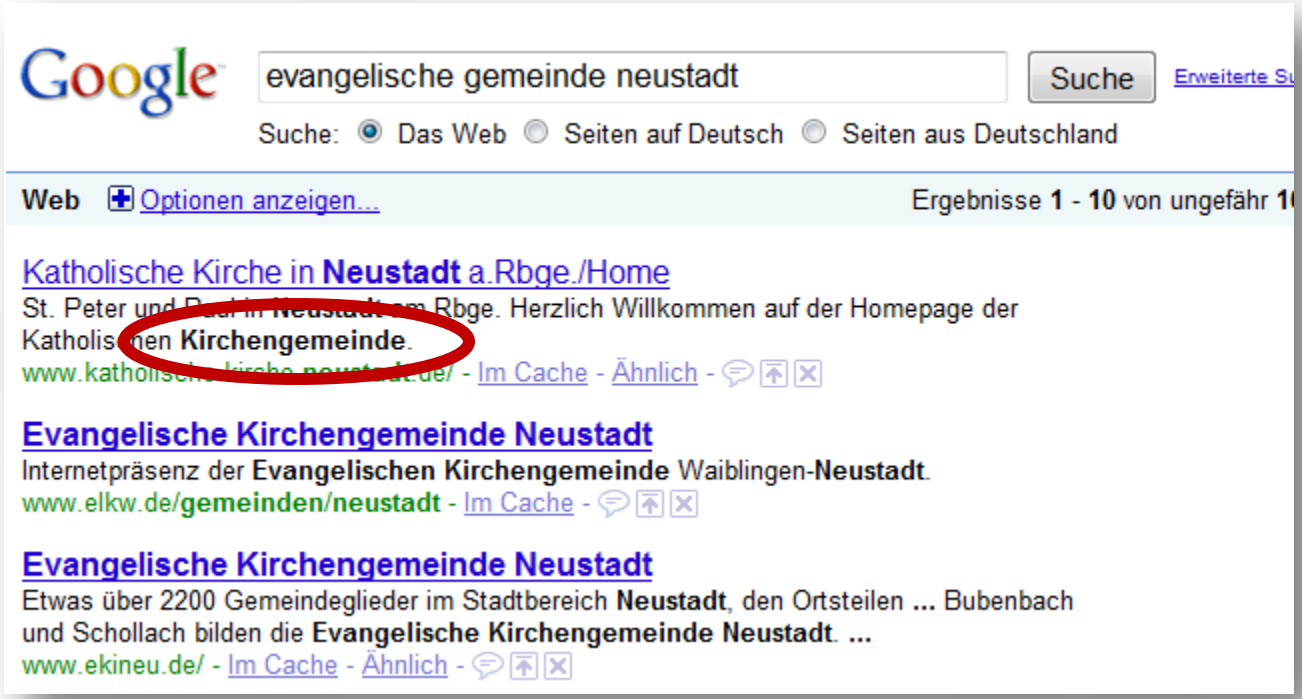
© 2004 – 2011 Matthias Schneider

Stellenanzeigen = Stellenangebote

© 2004 – 2011 Matthias Schneider



© 2004 – 2011 Matthias Schneider



- Sie sehen das nicht „ökumenisch“ genug, Google schon
- Rom lässt sich den Alleinvertretungsanspruch etwas kosten
- Sie haben beim Verwendungszweck für den Kirchenbeitrag „Suchmaschinenoptimierung“ angekreuzt

© 2004 – 2011 Matthias Schneider

Google ergänzt/ändert Suchbegriffe

Stellenanzeigen Berlin	→ Stellenangebote Berlin
Webagentur Berlin	→ Werbeagentur Berlin
Webseiten Design	→ Webdesign
song words	→ „words“ wurde ergänzt durch „lyrics“
what state has the highest murder rate	→ „homicide“ wurde ergänzt für „murder“
himalayan kitten breeder	→ "cat breeder" ist das gleiche wie „kitten breeder“

Kontextual:

dura ace track bb axle njs	→ "bb" here means "bottom bracket".
software update on bb color id	→ „bb“ steht für „blackberry“
bb cream dark	→ hier steht „bb“ für „blemish balm“
southeastern usa bb fitness & figure	→ „bb“ steht hier für „bodybuilding“
arm reduction oder arms reduction	→ keine Wortstammreduzierung

Quellen:
http://googleblog.blogspot.com/2010/01/helping-computers-understand-language.html

© 2004 – 2011 Matthias Schneider

Google ergänzt/ändert Suchbegriffe

Google Patent 7,409,383

Methode, um Synonyme oder anderen Ersatz-Klauseln zu bestimmen.

Für jede Suchanfrage wird eine Vielzahl von Pseudo-Suchanfragen bestimmt, jede Pseudo-Suchabfrage wird abgeleitet von Suchanfragen bei denen ein Phrase ausgetauscht wurde.

Ein potenzielles Synonym ist ein Begriff, der

a)innerhalb einer benutzerdefinierten Abfrage an die Stelle in einer Suchanfrage verwendet wurde

b)und im Kontext einer Pseudo-Suchanfrage auftaucht.

Quellen:
http://arnoldit.com/wordpress/2009/12/24/google-nails-patent-for-query-synonyms-in-query-context/

http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1=7,636,714.PN.&OS=pn/7,636,714&RS=PN/7,636,714

© 2004 – 2011 Matthias Schneider

Google squared (ver)sucht semantisch

Google squared labs

mozart birthday | Square it | Add to this Square

mozart birthday

Item Name	Image	Description	Death	Date Of Birth
Wolfgang Amadeus Mozart		**Wolfgang Amadeus Mozart** (German: [ˈvɔlfgaŋ amaˈdeus ˈmoːtsart], full baptismal name Johannes Chrysostomus Wolfgangus Theophilus Mozart (27 January 1756 – 5 ...	Vienna, Austria	27 January 1756
Wolfgang Amadeus Mozart Biography		**Wolfgang Amadeus Mozart**. (1756 - 1791) Probably the greatest genius in Western musical history, **Wolfgang Amadeus Mozart** was born in Salzburg, Austria, Jan.	Dec 5, 1791	1 possible value
Franz Xaver Wolfgang Mozart		**Franz Xaver Wolfgang Mozart** (26 July 1791 – 29 July 1844), also known as F. X. Mozart, W. A. Mozart Son, or Wolfgang Amadeus Mozart, Jr., was the youngest ...	29-Jul-1844	2 possible values
Anna Maria		**Maria Anna** Mozart, or "Nannerl" as family called her, is the older sister of the ... **Maria Anna** Nannerl Mozart, Wikimedia Commons **Maria Anna** Nannerl Mozart ...	2 possible values	4 possible values
Add items Add	or Add next 10 items			

© 2004 – 2011 Matthias Schneider

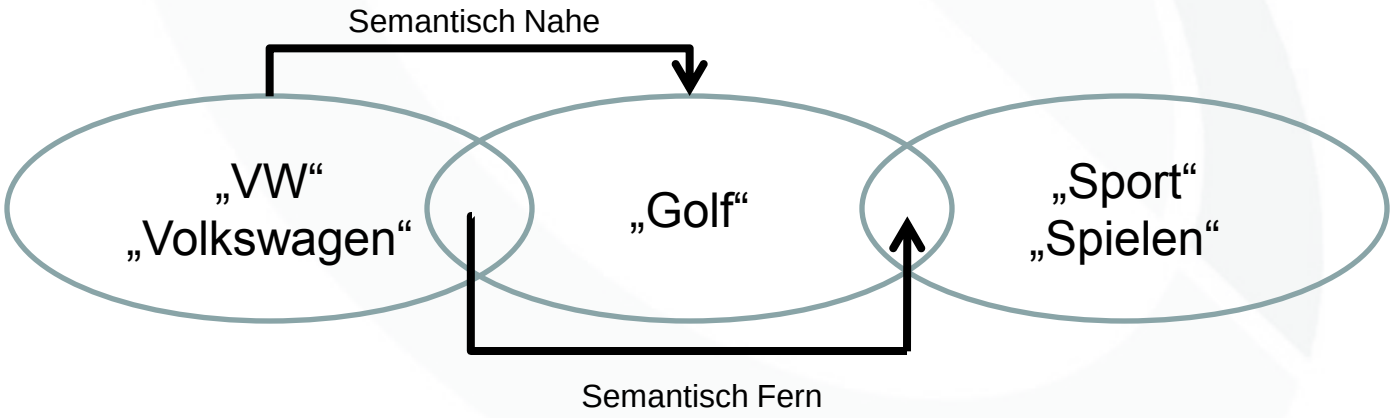
Wie berechnet man Semantik ? Gestern – Heute – Morgen

© 2004 – 2011 Matthias Schneider

„semantischen Nahe“ ← → „semantisch Fern“

In Webseiten bei denen es um „Golf“ geht, wird auch oft „VW“ und „Volkswagen“ genannt. Webseiten in denen zwar „Golf“ genannt wird, aber in anderem Zusammenhang (stattdessen mit „Sport“ und „Spielen“), sind *semantisch Fern* zu diesen. *Semantisch Nahe* sagt man, wenn Webseiten zwar „VW“ und „Volkswagen“ haben, aber eben nicht „Golf“.

+ Semantisch Nähe
- Längere Such- und Analysezeiten
- Eben doch nur Latent (und deswegen auch nicht *Synonym*)



LSI ← Scott Deerwester, 1990

© 2004 – 2011 Matthias Schneider

Content DNA

The closer the content of your webpage matches the ContentDNA, the higher the search engine concerned will rank your webpage on the content score for the specific search term.

Gravitationszentrum eines Themenclusters = ContentDNA

ABER: *Common Neurolinguistic Map*

es macht einen Unterschied was man von Schnee hält, ob man nun in der Arktis oder in der Karibik wohnt (→ kulturell und geographischer Sprach- und Meinungsraum)

© 2004 – 2011 Matthias Schneider

Phonetik

Hat eigentlich nichts mit Semantik zu tun, sollte aber der Vollständigkeit halber mal genannt sein: berechnet die Aussprache eines Wortes:

TABLE 1: SOUNDEX and Double Metaphone Encoding

Word	SOUNDEX	Primary Metaphone	Secondary Metaphone
Smith	S530	SM0	XMT
Smythe	S530	SM0	XMT
phone	P500	FN	no code
pony	P500	PN	no code
George	G620	JRJ	KRK
garage	G620	KRJ	KRK
benign	B525	PNN	PNKN
benignant	B525	PNNNT	PNKNNT
poignant	P255	PNNT	PKNNT

Interessant in Verwendung mit Levenshtein-Distanz, um z.B. Falschschreibweisen zu finden.

Soundex ← Robert Russel, 1918
Metaphone ← Lawrence Philips, 1990

© 2004 – 2011 Matthias Schneider

Thesaurus

Anbindung/Import einer bestehenden Thesaurus Datenbank und vergleich der Suchanfragen mit dieser

- Lexikalisch
  - Wordnet , Germanet , Wortschatz Uni Leipzig

- GPL
  - Openthesaurus
  - Wikipedias Wictionary

- Fachthesauri:
  - Standardthesaurus Wirtschaft (STW)
  - Medizin

- Multilinguale Thesauri
  - UNESCO

- Kostenpflichtige
  - Dornseiff

wictionary.org:

bugsieren

bugsieren (Deutsch) [Bearbeiten]

Verb [Bearbeiten]

Silbentrennung:
bug·sie·ren, Präteritum: bug·sier·te, Partizip II: bug·siert

Aussprache:
IPA: [buk'si:ʀən], Präteritum: [buk'si:ɐ̯tə], Partizip II: [buk'si:ɐ̯t]
Hörbeispiele: —

Bedeutungen:
[1] *transitiv*, Seefahrt: ein Schiff durch Lotsen ins Schlepptau nehmen und an einen bestimmten Ort schleppen
[2] *transitiv*, *umgangssprachlich*: mit Mühe an einen anderen Ort bringen
[a] Dinge
[b] Menschen, auch Tiere

Herkunft:
Lehnwort des 17. Jahrhunderts zunächst in den Schreibweisen *buxiren, buchsieren* aus dem Niederländischen *boegseren*, weiterentwickelt aus den alten Formen *boesjaarden, boesjaren*, kombiniert mit dem niederländischen Wort *boeg* (Bug); die niederländischen Urformen gehen auf das portugiesische *puxar* (ziehen, schleppen) zurück, das sich seinerseits aus dem Lateinischen *pulsare* (stoßen) ableitet; die jetzige Schreibweise stammt aus dem 19. Jahrhundert [1]

Synonyme:
[1] geleiten, lotsen, ins Schlepptau nehmen, schleppen
[1,2a] drücken, schieben, ziehen
[2a] schleifen, stoßen
[2b] dirigieren, drängen, unter die Arme greifen, schubsen, weiterhelfen

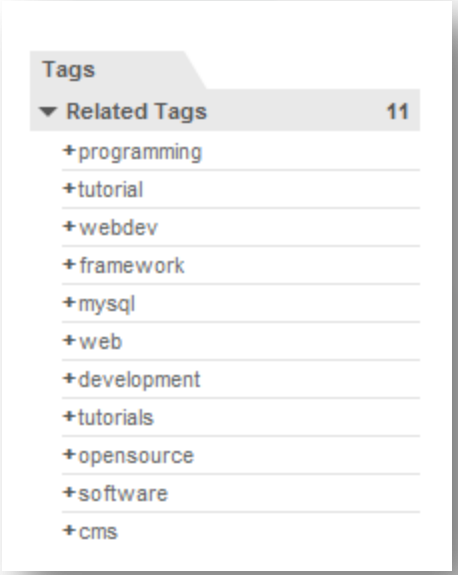
Gegenwörter:
[2a] karren, wegrollen, werfen

© 2004 – 2011 Matthias Schneider

Folksonomy / Social Tagging

Eine durch Benutzer erzeugte Stichwortsammlung zu einer Internetseite, Bild oder Artikel.

+ Sehr genau, da menschlich generiert
- Kein Algorithmus und somit nicht generisch anwendbar
- Kein kontrolliertes Vokabular
- Manipulierbar



Quellen:
Delicous.com, mister-wong.de (jeweils Screenshots der TagCloud)

© 2004 – 2011 Matthias Schneider

Clustern

Indem man z. B. die Wörter in den Titeln der ersten 100 Treffern einfach mathematisch gruppiert.

+ rel. schnell
- Ungenau (da nicht semantisch, sondern eben nur geclustert)

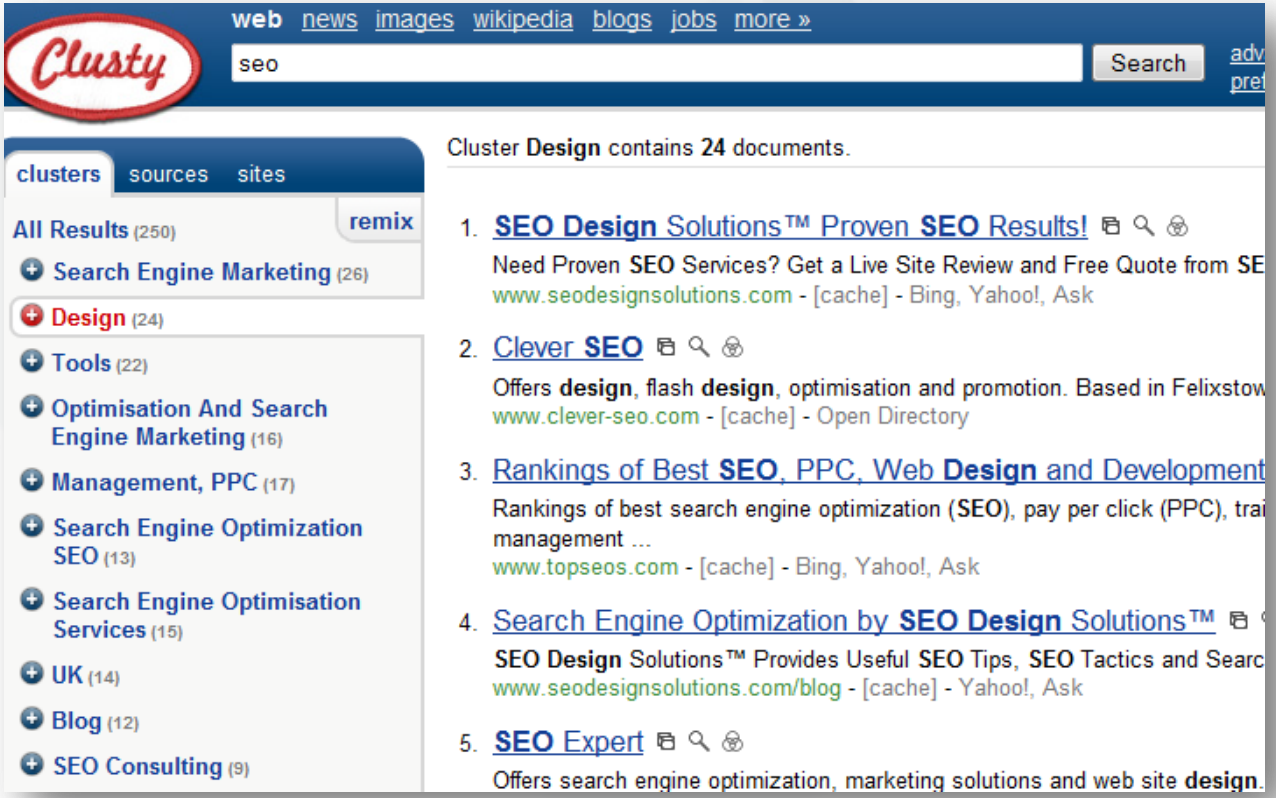


Bild:
www.clusty.com (Vivisimo)

© 2004 – 2011 Matthias Schneider

HAL (Hyperspace Analogue to Language)

Jedes Wort wird durch die Gesamtheit seiner Nachbarschaften im Kontext repräsentiert, oder einfacher ausgedrückt: **Wörter mit ähnlicher Bedeutung erscheinen in ähnlichen Sätzen.**

+ Semantische Nähe
= Ergebnisse ähnlich LSI

Wir analysieren ein paar Texte und stellen ein **Kookkurrenzen** fest:
(sprich „a“ findet sich oft im Zusammenhang mit „b“ genannt)

	mountain	valley	river	mouse	cat	dog
mountain	0	12	8	0	0	0
valley	12	0	9	0	0	0
river	8	9	0	0	0	0
mouse	0	0	0	0	9	0
cat	0	0	0	9	0	5
dog	0	0	0	0	5	0

Naive HAL Matrix

HAL ← Kevin Lund and Curt Burgress ,1996 http://www.psychonomic.org/search/view.cgi?id=1105

© 2004 – 2011 Matthias Schneider

LSI – Latent Semantic Indexing

1)Man bilde die Term-Dokumenten-Matrix aller URLs

Sample Term by Document matrix

	access	*document*	*retrieval*	*information*	*theory*	*database*	*indexing*	*computer*
Doc 1	x	x	x			x	x	
Doc 2				x*	x			x*
Doc 3			x	x*				x*

2)lege darüber den semantischen Raum
(via Singularitätswertzerlegung der wichtigsten Konzepte/Wörter)
(Idealerweise mit vorhanden Synonymdatenbanken noch weiter verkleinern)

➔ Die Matrix wird kleiner , lässt sich schneller rechnen,

© 2004 – 2011 Matthias Schneider

PLSI (Probabilistic Latent Semantic Indexing) ←Thomas Hofmann 1999

Im Vergleich zu LSI welches auf einer Matrixzerlegung basiert, hat die probabilistische Variante statistische Grundlagen (**bedingte Wahrscheinlichkeit**), um eine höhere Präzision zu erreichen. Die Dimensionsreduktion erfolgt nicht via SVD (Singularitäts-Wert-Zerlegung), sondern auf **Bayes** beruhende Wahrscheinlichkeitsberechnungen.

HTMM (Hidden Topic Markov Model) ← Amit Gruber 2007

- Annahme: alle Wörter in einem Satz haben das gleiche Thema.
- Annahme: nachfolgende Sätze haben das gleiche Thema.
- „Latent“, sprich das eigentliche Thema ist aber versteckt.
- Anwendung des „Hidden Markov Modells“ zur **Mustererkennung verborgener Zustände**.

= Hidden Topic Markov Modell

(verborgende Zustände von DNA-Sequenzen, 23andme ??)

© 2004 – 2011 Matthias Schneider

Auf der Suche nach dem semantischen Raum

© 2004 – 2011 Matthias Schneider

Suche nach dem „semantischen Raum“

Wort: evangelisch
Anzahl: 109
Häufigkeitsklasse: 15 (d.h. *der* ist ca. 2^15 mal häufiger als das gesuchte Wort)
Morphologie: evangel|isch
Grammatikangaben: Wortart: Adjektiv
Pragmatikangaben: etym: griech.
Relationen zu anderen Wörtern:

- Synonyme: protestantisch
- wird referenziert von: protestantisch

Links zu anderen Wörtern:

- Grundform: evangelisch
- Haupteintrag (Polysem): evangelisch, evangelisch
- ist ein(e) protestantisch
- Form(en): evangelischen, evangelische, evangelischer, evangelisch, evangelischem, evangelisches
- Untereinträge: evangelisch, evangelisch

Dornseiff-Bedeutungsgruppen:

- 22.1 Religiosität, Glaube: evangelisch, gläubig, gottergeben, gottesfürchtig, gottgefällig, heilig, kanonisch, katholisch, orthodox, protestantisch, puritanisch, rein, religiös, tugendhaft, unschuldig
- 22.12 Gebet, Frömmigkeit: biblisch, eingegeben, erbaulich, evangelisch, geheiligt, göttlich, inspiriert, prophetisch

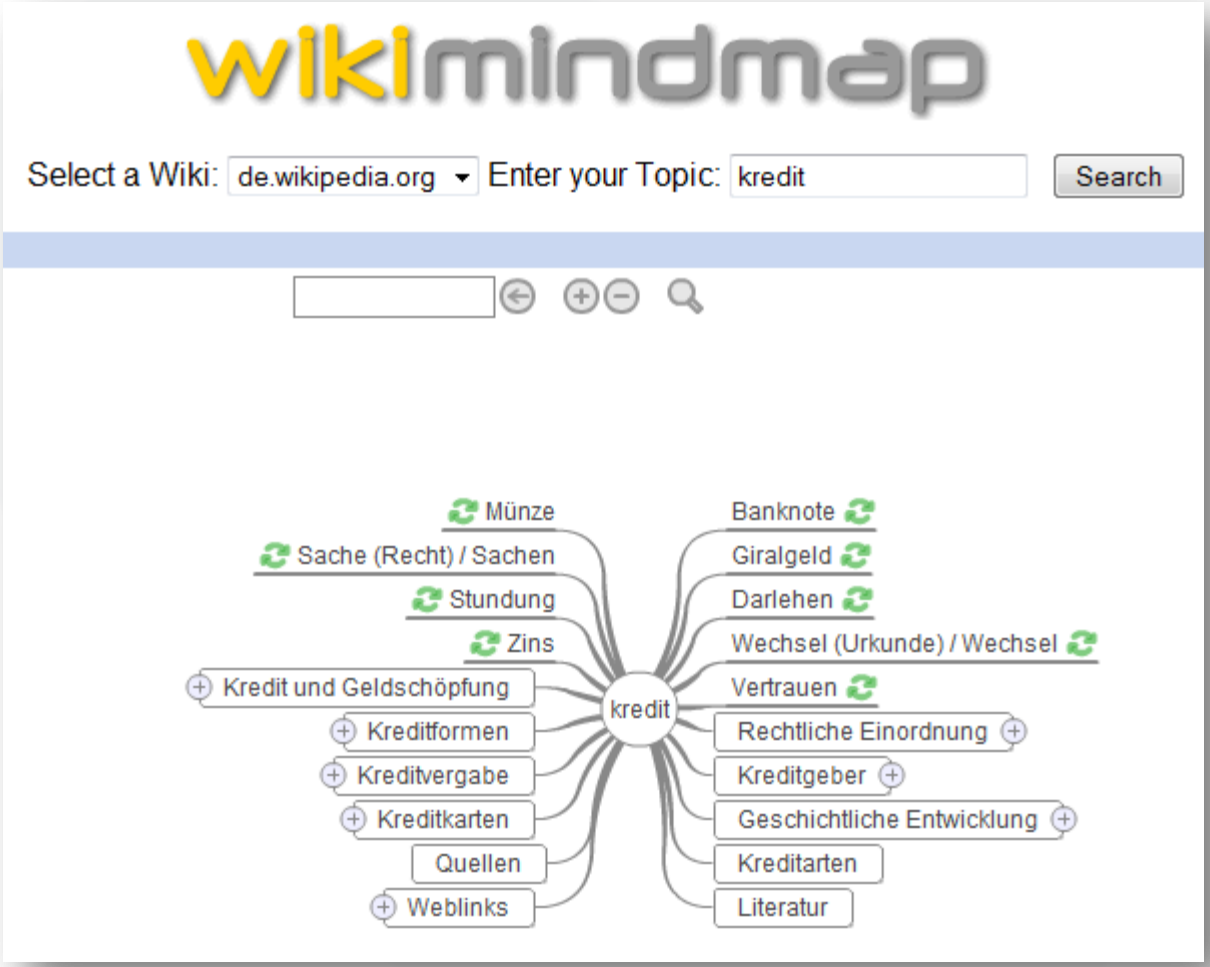
Beispiel(e):
Nur eins macht mir Sorge: Angela Merkel ist **evangelisch**. (Quelle: *fr-aktuell.de vom 08.01.2005*)
Die 54-jährige Juristin ist in Frankfurt geboren, wohnt seit 21 Jahren in Neu-Isenburg, ist **evangelisch**, verheiratet und hat zwei erwac
Söhne. (Quelle: *fr-aktuell.de vom 13.01.2005*)
Moshammer war **evangelisch**, trat jedoch aus der Kirche aus. (Quelle: *n-tv.de vom 19.01.2005*)
weitere Beispiele

Signifikante Kookkurrenzen für evangelisch:
katholisch (506), getauft (55), Kirche (55), Krippen (53), Krippentraditionen (45), Taufunterlagen (45), jüdisch (41), getaufter
Getaufte (38), Heimatmuseum (35), geboren (30), geschieden (30), konfessionell (29), Oelsner (27), muslimisch (27), herumge
geprägten (26), Katholiken (25), Einwohner (25), draufsteht (24), Kindheit (23), Evangelium (23), Seid (23), Religionszugehör

Quelle:
Wortschatz Uni-Leipzig

© 2004 – 2011 Matthias Schneider

Suche nach dem „semantischen Raum“



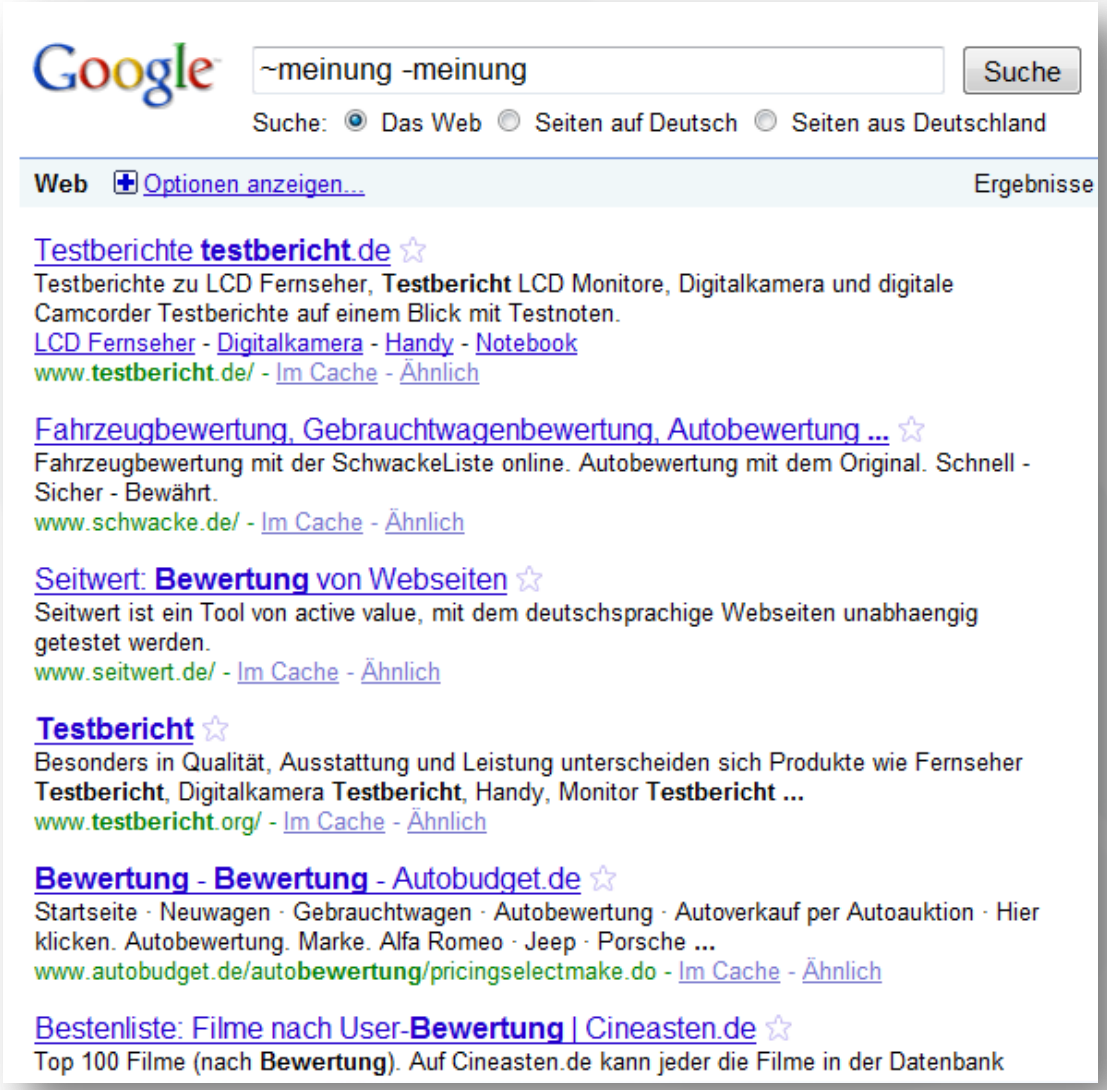
© 2004 – 2011 Matthias Schneider

Suche nach dem „semantischen Raum“

Predicted Items
evangelische
kirche
evangelisch
protestantismus
katholische
gesellschaft
homepages
reformation
diakoniestation
ehelosigkeit
zölibat

© 2004 – 2011 Matthias Schneider

Suche nach dem „semantischen Raum“



© 2004 – 2011 Matthias Schneider

Semager Keyword API

All In- and Output UTF-8 encoded

Query

Language (optional, default: de)
de

Output (optional, default: xml)
xml

Count (optional, max. number of results, default: 20)

Threshold (optional, min. correlation to query in percentage

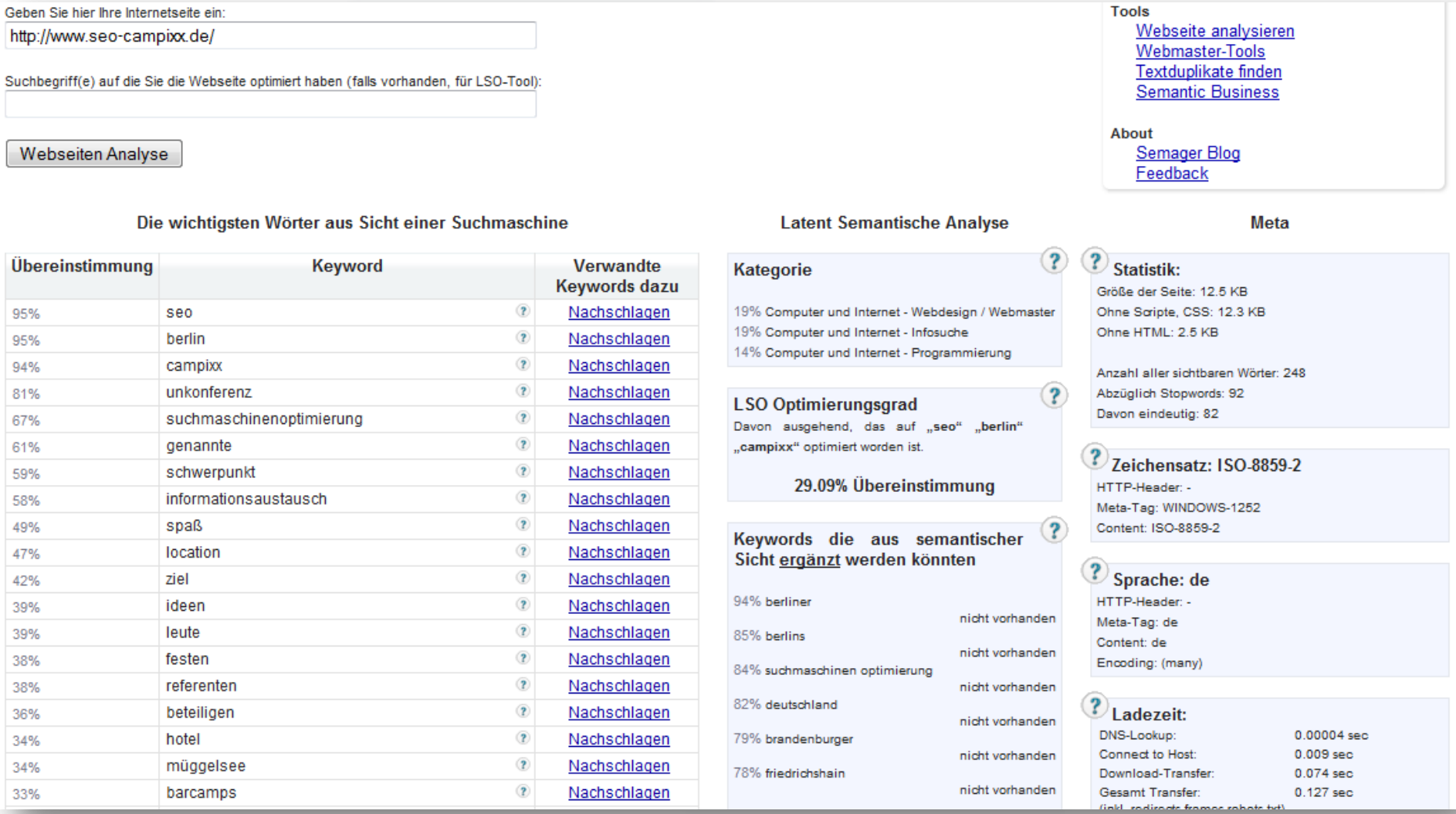
Sort (optional, sort after the correlation to query or absolut
(Freq. means its sorts after how much common a word is)
correlation

Daten absenden

```
- <results>
 - <resultset title="Semager: Keyword API" query="seo" spell="" lang="de"
   - <item>
       <score>83</score>
       <tag>suchmaschinen optimierung</tag>
     </item>
   - <item>
       <score>76</score>
       <tag>sem</tag>
     </item>
   - <item>
       <score>76</score>
       <tag>webseiten</tag>
     </item>
   - <item>
       <score>70</score>
       <tag>suchmaschinen</tag>
     </item>
   - <item>
       <score>70</score>
       <tag>optimierung</tag>
     </item>
   - <item>
       <score>70</score>
       <tag>suchmaschinenoptimierung</tag>
     </item>
```

© 2004 – 2011 Matthias Schneider

Semager URL Analyse



Geben Sie hier Ihre Internetseite ein:
http://www.seo-campixx.de/

Suchbegriff(e) auf die Sie die Webseite optimiert haben (falls vorhanden, für LSO-Tool):

Webseiten Analyse

Tools
- Webseite analysieren
- Webmaster-Tools
- Textduplikate finden
- Semantic Business

About
- Semager Blog
- Feedback

Die wichtigsten Wörter aus Sicht einer Suchmaschine

Übereinstimmung	Keyword	Verwandte Keywords dazu
95%	seo	Nachschlagen
95%	berlin	Nachschlagen
94%	campixx	Nachschlagen
81%	unkonferenz	Nachschlagen
67%	suchmaschinenoptimierung	Nachschlagen
61%	genannte	Nachschlagen
59%	schwerpunkt	Nachschlagen
58%	informationsaustausch	Nachschlagen
49%	spaß	Nachschlagen
47%	location	Nachschlagen
42%	ziel	Nachschlagen
39%	ideen	Nachschlagen
39%	leute	Nachschlagen
38%	festen	Nachschlagen
38%	referenten	Nachschlagen
36%	beteiligen	Nachschlagen
34%	hotel	Nachschlagen
34%	müggelsee	Nachschlagen
33%	barcamps	Nachschlagen

Latent Semantische Analyse

Kategorie

19% Computer und Internet - Webdesign / Webmaster
19% Computer und Internet - Infosuche
14% Computer und Internet - Programmierung

LSO Optimierungsgrad

Davon ausgehend, das auf „seo“ „berlin“ „campixx“ optimiert worden ist.

29.09% Übereinstimmung

Keywords die aus semantischer Sicht ergänzt werden könnten

94% berliner	nicht vorhanden
85% berlins	nicht vorhanden
84% suchmaschinen optimierung	nicht vorhanden
82% deutschland	nicht vorhanden
79% brandenburger	nicht vorhanden
78% friedrichshain	nicht vorhanden

Meta

Statistik:
Größe der Seite: 12.5 KB
Ohne Scripte, CSS: 12.3 KB
Ohne HTML: 2.5 KB

Anzahl aller sichtbaren Wörter: 248
Abzüglich Stopwords: 92
Davon eindeutig: 82

Zeichensatz: ISO-8859-2
HTTP-Header: -
Meta-Tag: WINDOWS-1252
Content: ISO-8859-2

Sprache: de
HTTP-Header: -
Meta-Tag: de
Content: de
Encoding: (many)

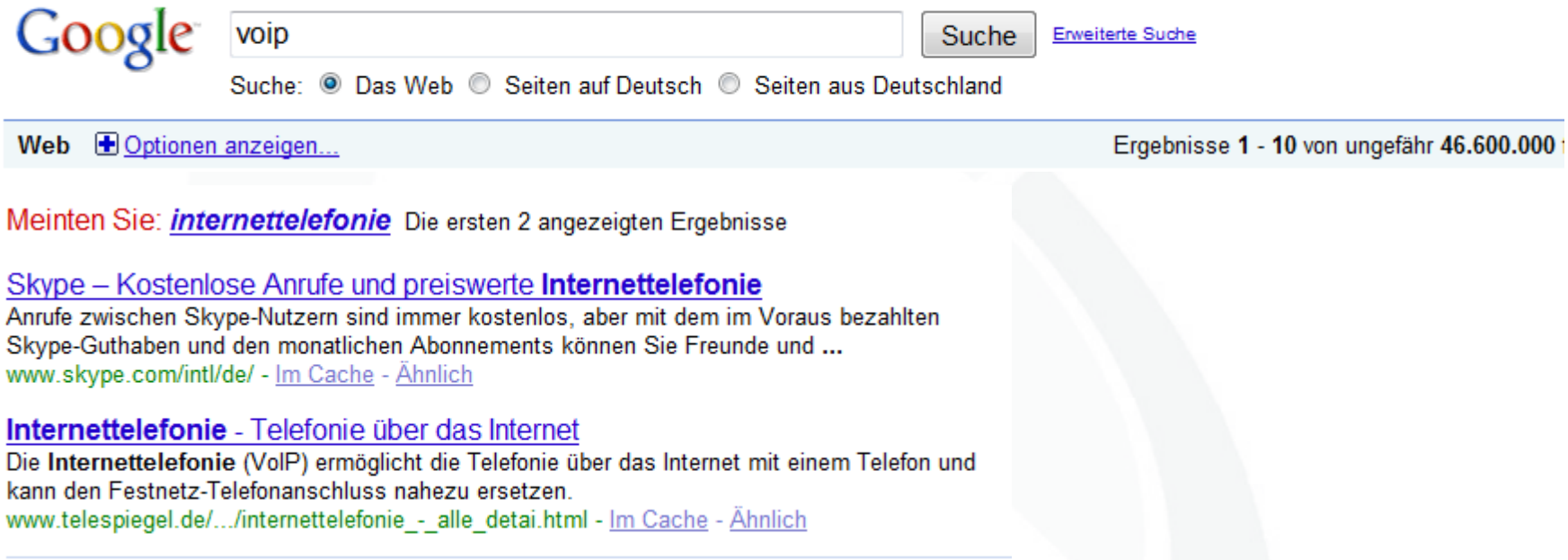
Ladezeit:

DNS-Lookup:	0.00004 sec
Connect to Host:	0.009 sec
Download-Transfer:	0.074 sec
Gesamt Transfer:	0.127 sec

© 2004 – 2011 Matthias Schneider

Glaskugel

© 2004 – 2011 Matthias Schneider



Fake (noch) ;-)

© 2004 – 2011 Matthias Schneider

Wie finde ich den ContentDNA zu einem Wort / Thema heraus?

www.semager.de/keywords/
das ist ein Affiliate Link ;-)

© 2004 – 2011 Matthias Schneider