



Information Retrieval

Ulf Leser

Web Search Engines

The screenshot shows a Google search for 'nino' in German. The search results include:

- El Niño – Wikipedia**: https://de.wikipedia.org/wiki/El_Niño. El Niño (span. für „der Junge, das Kind“, hier konkret „das Christuskind“) nennt man das Auftreten ungewöhnlicher, nicht zyklischer, veränderter Strömungen im ...
Ablauf · Fernwirkungen · Geschichte · Vorhersagemöglichkeiten und ...
- Nino – Wikipedia**: <https://de.wikipedia.org/wiki/Nino>. Nino steht für Nino (Fußballspieler) (Juan Francisco Martínez Modesto; * 1980), spanischer Fußballspieler, Nino (Heilige), eine Heilige; Nino (Sänger) (Amir ...

Below the text results is a map of Berlin showing locations like 'Salumeria da Nino' and 'Nino Permakultur'. Below the map are details for these locations:

- nino Permakultur**: Keine Rezensionen · Garten-/Landschaftsbauunternehmen · Schierker Str. 9A · 030 71530407 · Geöffnet bis 18:00. Website and Route buttons are present.
- Salumeria da Nino**: 4,7 ★★★★★ (11) · Italienisches Restaurant · Italienisches Restaurant & Feinkostladen · Geisbergstraße 14 · 030 2134841 · Geöffnet bis 19:00. Website and Route buttons are present.

At the bottom, there is a link to 'Die El Niño Infoseite' with the URL www.elnino.info/. A small 'Info' icon is also visible.

The screenshot shows a Bing search for 'nino'. The search results include:

- Bilder von nino**: A grid of images showing children and a dog.
- El Niño – Wikipedia**: https://de.wikipedia.org/wiki/El_Niño. El Niño (span. für „der Junge, das Kind“, hier konkret „das Christuskind“) nennt man das Auftreten ungewöhnlicher, nicht zyklischer ...
Ablauf · Fernwirkungen · Geschichte · La Niña · Literatur
- Nino – Wikipedia**: <https://de.wikipedia.org/wiki/Nino>. NINO Zuhilfenahmeger, NINO Verwaltungsgelände und NINO-Hochbau, heute unter Denkmalschutz stehende ehemalige Gebäude im NINO-Wirtschaftspark. Siehe auch:
- Videos von nino**: A row of video thumbnails with titles like 'NINO - Theos', 'NINO - Fovamai', and 'Nino - Theos (2010)'.

On the right side, there is a 'Wikipedia' section with a 'Benutzer suchen auch nach' section listing related terms like 'La Niña', 'Meeresströmung', 'Passat (Windsystem)', 'Monsun', and 'Walker-Zirkulation'. Below that is a 'Nino 99000209 bei Amazon' section with a 'Amazon.de/Beleuchtung' link and a 'Nino, Kobuleti' section with a 'Booking.com/Nino' link.

The screenshot shows a Google search for 'el nino'. The search results include:

- El Niño** (span. für „der Junge, das Kind“, hier konkret „das Christuskind“) nennt man das Auftreten ungewöhnlicher, nicht zyklischer, veränderter Strömungen im ozeanographisch-meteorologischen System (El Niño-Southern Oscillation, ENSO) des äquatorialen Pazifiks.
- El Niño – Wikipedia**: https://de.wikipedia.org/wiki/El_Niño

Below the text results is a small map of the Pacific Ocean showing the ENSO system. At the bottom, there is a link to 'Die El Niño Infoseite' with the URL www.elnino.info/.

Web Search Engines

The screenshot shows a Bing search results page for the query "nino". The search bar at the top contains "nino" and the search button is a blue magnifying glass. Below the search bar, there are navigation tabs for "Web", "Bilder", "Videos", "Karten", "News", and "Erkunden". The main content area is divided into several sections:

- Bilder von nino**: A section titled "Bilder von nino" with a sub-link "bing.com/images". It displays a grid of six small images showing children playing and smiling.
- Wikipedia**: A section titled "El Niño" with a sub-link "Wikipedia". It contains a brief description: "El Niño nennt man das Auftreten ungewöhnlicher, nicht zyklischer, veränderter Strömungen im Pazifik. Der Name ist von dem spanischen Wort 'Niño' abgeleitet, nämlich zur..." and a small diagram of the Pacific Ocean showing the equatorial current.
- Benutzer suchen auch nach**: A section titled "Benutzer suchen auch nach" with sub-links for "La Niña", "Meeresströmung", "Passat (Windsysteme)", and "Monsun".
- Nino - Wikipedia**: A section titled "Nino - Wikipedia" with a sub-link "https://de.wikipedia.org/wiki/Nino". It contains a brief description: "El Niño (span. für ‚der Junge, das Kind‘, hier konkret ‚das Christuskind‘) nennt man das Auftreten ungewöhnlicher, nicht zyklischer..." and a sub-link "Ablauf Fermentierungen Geschichte La Niña Literatur".
- Nino - Wikipedia**: A section titled "Nino - Wikipedia" with a sub-link "https://de.wikipedia.org/wiki/Nino". It contains a brief description: "NINO-Rohgewebelager, NINO-Verwaltungsgebäude und NINO-Hochbau, heute unter Denkmalschutz stehende ehemalige Gebäude im NINO-Wirtschaftspark. Siehe auch:"
- Videos von nino**: A section titled "Videos von nino" with a sub-link "bing.com/videos". It displays a row of four video thumbnails with play buttons and durations: "NINO - Theos (2010)", "NINO - Theos", "Nino - Fovamai", and "Nino - Theos (2010)".

The screenshot shows a Yahoo! search results page for the query "nino". The search bar at the top contains "nino" and the search button is a blue "Suche" button. Below the search bar, there are navigation tabs for "Start", "Mail", "Nachrichten", "Sport", "Finanzen", "Stars", "Style", "Movies", "Wetter", "Flickr", "Mobile", and "Weitere". The main content area is divided into several sections:

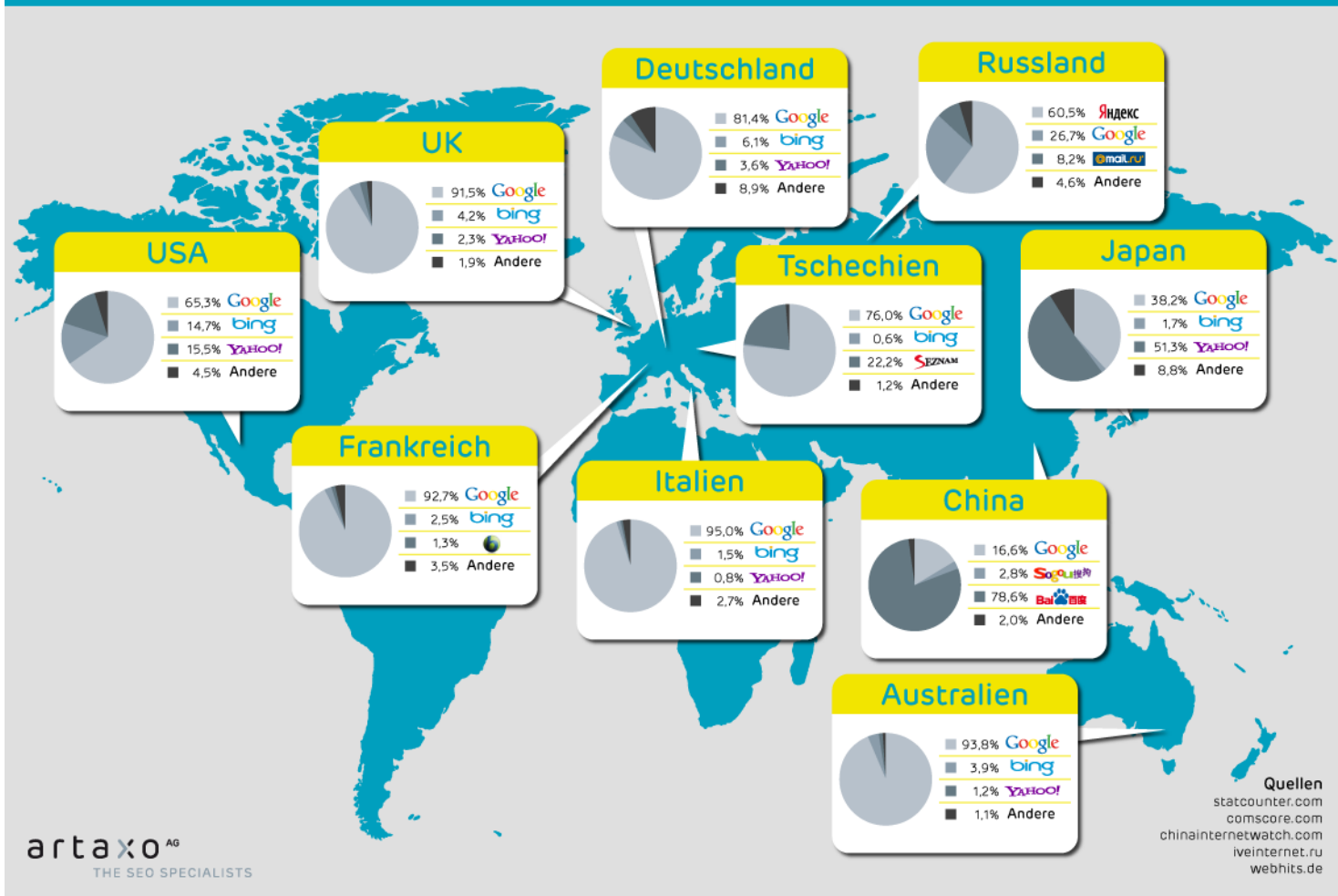
- Web**: A section titled "Web" with a sub-link "Versuchen Sie auch: nino de angelo, el niño, nino d'angelo, nino farina".
- Bilder**: A section titled "Bilder" with a sub-link "Suchergebnisse".
- Video**: A section titled "Video" with a sub-link "Nino - Ergebnisse in der Bildersuche".
- Nachrichten**: A section titled "Nachrichten" with a sub-link "Lokale Suche".
- Alle Treffer**: A section titled "Alle Treffer" with a sub-link "Letzter Tag", "Letzte Woche", and "Letzen Monat".
- Das Web**: A section titled "Das Web" with a sub-link "Seiten auf Deutsch" and "Seiten aus Deutschland".
- El Niño - Wikipedia**: A section titled "El Niño - Wikipedia" with a sub-link "de.wikipedia.org/wiki/El_Niño". It contains a brief description: "El Niño (span. für ‚der Junge, das Kind‘, hier konkret ‚das Christuskind‘) nennt man das Auftreten ungewöhnlicher, nicht zyklischer, veränderter..."
- Nino - Wikipedia**: A section titled "Nino - Wikipedia" with a sub-link "de.wikipedia.org/wiki/Nino". It contains a brief description: "NINO-Rohgewebelager, NINO-Verwaltungsgebäude und NINO-Hochbau, heute unter Denkmalschutz stehende ehemalige Gebäude im NINO-Wirtschaftspark. Siehe auch: Niño."
- Nino - Ergebnisse in der Videosuche**: A section titled "Nino - Ergebnisse in der Videosuche" with a sub-link "Weitere Videos". It displays a row of four video thumbnails with play buttons and durations: "Theos | NINO - Theos", "NINO - my mail.ru", "Fovamai - en.musicplayon...", and "Amor - fr.musicplayon...".
- NINO® Percussion**: A section titled "NINO® Percussion" with a sub-link "ninopercussion.com/de". It contains a brief description: "NINO® Percussion offers an outstanding collection of musical instruments designed specifically for children. These more than meet the requirements of early childhood..."

Estimated Scale [Beware: Diverging evidences]

- Queries (only google, 2016)
 - World-wide: ~150.000.000.000 queries / month
 - Per day: ~5.000.000.000
 - Per second: ~50.000
 - Germany: ~5.000.000.000 queries / month
- Web ([how to count](#) / estimate?)
 - 14.3 Trillion webpages (www.factshunt.com, 31.12.13)
 - >4.29 billion webpages (www.worldwidewebsite.com, 15.10.14)
 - >1 billion sites (www.internetlivestats.com, 15.10.14)
 - ~5 billion sites (WorldWideWebSize.com, June 2016)

Market Shares (2014)

Der internationale Suchmaschinenmarkt



Web Basics



Server T:

\index.html
\comm\pic.jpg
\comm\product.html
...



Server S:

\index.html
\main\pic.jpg
\main\text.html
...



Client/Browser:

S: Gib „\index.html“...

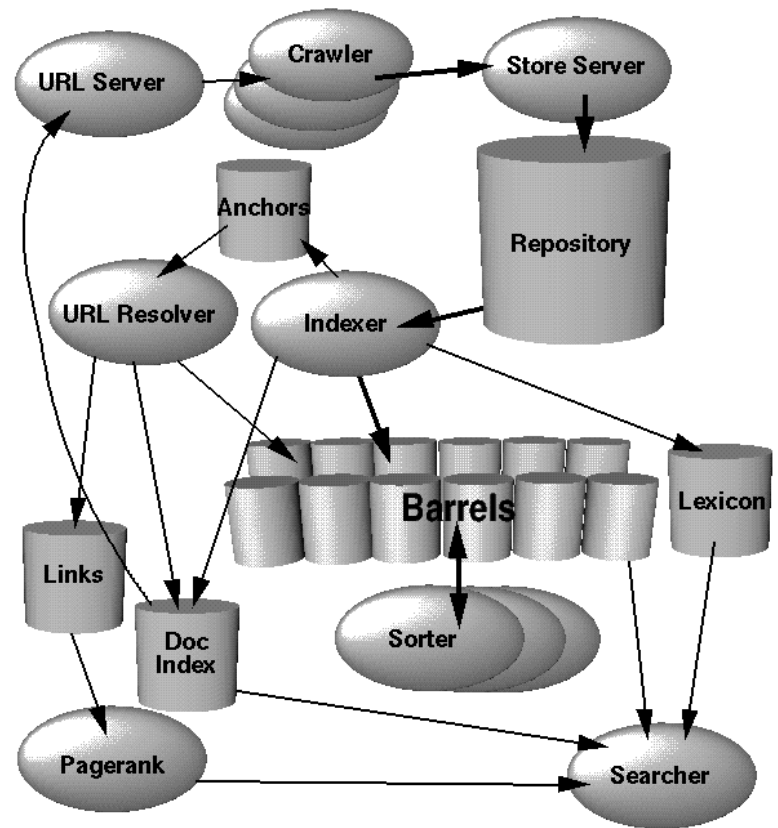
```
<html>  
  blabla <a href=„http:T\index.html“>blublu</a>  
</html>
```

Searching the Web?

- Browser needs server name and page name (URL)
 - Mostly taken from a link
- Browser loads page from server for display
- Web consists of >1.000.000.000 sites
- How can we **search 1 billion sites** in milliseconds?
 - Corresponding to 100? 1000? billion web pages

Crawling

- At query time, only one server is searched – located at the search engine
- Every search engine has a (partial) **copy of the web**
- Created and maintained by a **crawler**



Not (easily) Indexed: The Deep Web

The screenshot shows a Firefox browser window displaying the Galeria Kaufhof website. The address bar shows the URL www.galeria-kaufhof.de/.... The search bar contains the text "gummistiefel kinder" and a green "Suchen" button. The search results page displays a message: "Ihre Suche nach 'gummistiefel kinder' ergab 4 Treffer" and "den Filtern in der linken Spalte". The left sidebar shows filter options for "Shop-Kategorie" (Schuhe (2), Spielwaren (2)) and "Marken" (Manguun (2), RAVENSBURGER SPIELE (1), Sigikid (1)). The main content area shows two product listings: "Sigikid Pinky Queeny Gummistiefel Gr. 23" and "manguun Stiefel".

What is a good Search result?

The image displays three side-by-side browser windows, each showing search results for the query "el nino".

- Left Window (Google):** Shows the Google search interface with the query "el nino". It reports approximately 71.7 million results. The top results include a Wikipedia entry for "El Niño", an "Infoseite" (information page) from "elnino.info", and a NOAA research page. A cookie consent banner is visible at the top.
- Middle Window (Yahoo!):** Shows the Yahoo! search interface with the query "el nino". It features a "Versuchen Sie auch" (Try also) section with suggestions like "El Niño im Vergleich - Spitzen El Niño zu Top-Preisen" and "El Niño noch günstiger". It also lists "Anzeigen zum Thema" (Ads for the topic) and "Weitere Sponsoren" (Further sponsors).
- Right Window (Bing):** Shows the Bing search interface with the query "el nino". It reports 74,700,000 results. The top results include "El Niño im Vergleich - Spitzen El Niño zu Top-Preisen", "El Niño noch günstiger", "El Niño günstig | preisvergleich.de", "Die El Niño Infoseite", "El Niño - Wikipedia", "Das Inhaltsverzeichnis - Die El Niño Infoseite", "Bilder von el nino", and "Das ENSO-Phänomen: El Niño und La Niña". A "ÄHNLICHE SUCHEN" (Similar searches) sidebar is on the right.

What do **you** Expect?

- Climate researcher: The weather phenomena
- Traveler to Peru: Implications of the weather phenomena
- Citizens of Weimar: The Restaurant
- Cineastes: The movie
- Outdoor fan: The brand
- ...

Challenges to Keyword Search

- How can we measure **relevance of a page** given a query?
- Interpreting a query is difficult
 - **Users have different** intentions and understandings
 - Many words have many senses: **Homonyms**
 - Usually you look for only one sense
 - Usually a web side uses only one sense: **One sense per discourse**
 - Many things have many names: **Synonyms**
- One remedy: **Longer queries**
 - Use semantically close word to narrow down: „El nino pazifik klima“
 - But: These again have homonyms
 - Large corpus (web): Precision increases, recall doesn't matter
 - Small corpus (library): Precision may decrease, recall increase

Boolean Keyword Search (with AND semantics)

- Naive: A page is relevant iff it **contains all query token**
- Disadvantages
 - <El nino> – many false positives (because homonyms – el, nino)
 - “EL nino jugaba futbol” – But who searches for “the boy”?
 - <El nino pazifik klima> – many false negatives
 - „El Nino ist ein Phänomen, dass im pazifischen Ozean auftritt und das Wetter weltweit beeinflusst“
- Web problem
 - There are anyway 100.000+ hits
 - FP are not really important, but **ranking** is
- Boolean information retrieval: **From the 80ths**
 - Does not work for lay people (Web)
 - Does not work for very large corpora (Web)

Vector Space Model

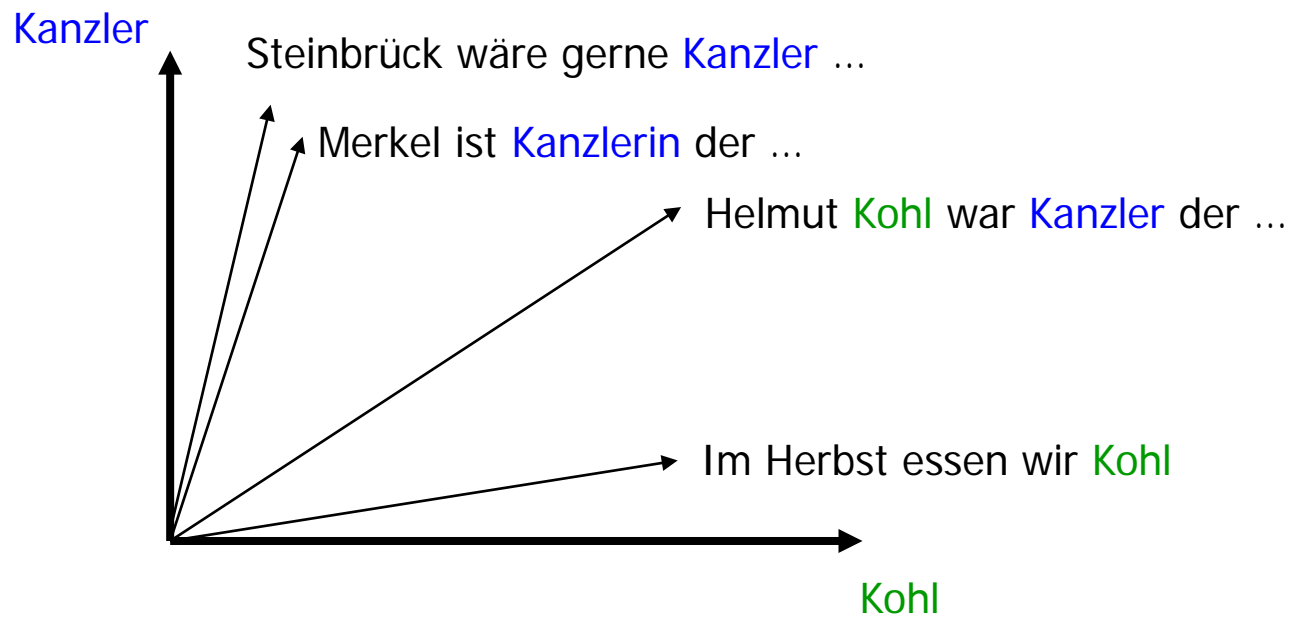
- Transform each page into a **high dimensional vector**
- Every unique token is a dimension
- Value can be binary, or count occurrences, or ...
- Vector as has many dimensions as there are **unique tokens on the web**

Example (after linguistic preprocessing)

	Text	verkauf	haus	italien	gart	miet	blüh	woll
1	Wir verkaufen Häuser in Italien	1	1	1				
2	Häuser mit Gärten zu vermieten		1		1	1		
3	Häuser: In Italien, um Italien, um Italien herum		1	1				
4	Die italienischen Gärtner sind im Garten			1	1			
5	Der Garten in unserem italienischen Haus blüht		1	1	1		1	

Comparing Vectors

- Assumption: Texts with semantically similar content will share many tokens
- Their **vectors are similar** (in some sense)



Pages and Queries

	Text	verkauf	haus	italien	gart	miet	blüh	woll
1	Wir verkaufen Häuser in Italien	1	1	1				
2	Häuser mit Gärten zu vermieten		1		1	1		
3	Häuser: In Italien, um Italien, um Italien herum		1	1				
4	Die italienischen Gärtner sind im Garten			1	1			
5	Der Garten in unserem italienischen Haus blüht		1	1	1		1	
Q	Wir wollen ein Haus mit Garten in Italien mieten		1	1	1	1		1

Using the Angle between Vectors

$$sim(d, q) = \frac{\sum (v_q[i] * v_d[i])}{\sqrt{\sum v_d[i]^2}}$$

1	1	1	1				
2		1		1	1		
3		1	1				
4			1	1			
5		1	1	1		1	
Q		1	1	1	1		1

	Q: Wir wollen ein Haus mit Garten in Italien mieten
1	d ₂ : Häuser mit Gärten zu vermieten
2	d ₅ : Der Garten in unserem italienischen Haus blüht
3	d ₄ : Die italienischen Gärtner sind im Garten
	d ₃ : Häuser : In Italien , um Italien , um Italien herum
5	d ₁ : Wir verkaufen Häuser in Italien

A Solution?

- <El nino pazifik klima>
 - Problem: „El Nino ist ein Phenomen, dass im pazifischen Ozean auftritt und das Wetter weltweit beeinflusst“
 - **Missing words** are not decisive any more – just a wider angle
 - The more shared words, the smaller the angle, the **better the rank**
 - Problem: Ranking for short queries with very many results
 - Pages having the same token in common with the query all get the same rank
 - We need more ranking power: **PageRank**

Modul Information Retrieval

- Lecture 2 SWS
- Exercises 2 SWS
- Slides are English
- Examination: Written (Klausur)

- Contact
Ulf Leser
Raum: IV.401
Tel: (030) 2093 – 3902
eMail: leser (..) informatik . hu-berlin . de

Literatur

- Manning, C. D., Raghavan, P. and Schütze, H. (2008). "Introduction to Information Retrieval", Cambridge UP
- Other
 - Grossmann, Frieder: „Information Retrieval“, Springer, 2004
 - Henrich (2007): „Information Retrieval 1 “, Online-Lehrbuch
 - Witten, Moffat, Bell (1999): „Managing Gigabytes: Compressing and Indexing Documents and Images“, Morgan Kaufmann
- Also interesting
 - Lemnitzer, L. and Zinsmeister, H. (2010). "Korpuslinguistik - Eine Einführung", narr Studienbücher.
 - Lüdeling, A. (2009). "Grundkurs Sprachwissenschaft". Stuttgart, Klett Lerntraining.
 - Manning, C.D., Schütze, H. (1999). „Foundations of Statistical Natural Language Processing“, MIT Press.

managing gigabytes - Goo... x Information Retrieval — Wi... x +

https://www.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1415/vl_inforet/

Meistbesucht Nachsehen Frequent WBI Lehre Google News Buecher kaufen Projekte Paper suchen Reisen MyStuff hub Berlin Wetter

Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik
Wissensmanagement in der Bioinformatik

Kontakt
Mitarbeiter
Veranstaltungen
Lehre
Archiv
WS 14/15
Modul Data Warehousing und Data Mining
Modul Information Retrieval
Übung zu Information Retrieval
Seminar Infrastrukturen für BIG DATA Anwendungen
Forschungsseminar
SS14
WS 13/14
SS13
WS 12/13
SS12
WS 11/12
SS 11
WS 10/11
SS 10
WS 09/10
SS 09
WS 08/09
SS 08
WS 07/08
SS 07
WS 06/07
SS 06
WS 05/06
SS 05
WS 04/05
SS 04
WS 03/04
SS 03
WS 02/03

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Deutsch English Kontakt/Impressum

Humboldt-Universität zu Berlin | Mathematisch-Naturwissenschaftliche Fakultät | Institut für Informatik | Wissensmanagement in der Bioinformatik | Lehre | Archiv | WS 14/15 | Modul Information Retrieval

Website durchsuchen

Information Retrieval

Professor Dr. Ulf Leser

Das Modul "Information Retrieval" behandelt Methoden zur Suche in (sehr grossen) Textsammlungen, insbesondere im Web. Vorgestellt werden Algorithmen und Verfahren zur Textvorverarbeitung, Anfragesprachen, Relevanzmodelle, Indexierung, und spezielle Probleme bei Web-Suchmaschinen. Am Ende der Vorlesung werden auch kleinere Auszüge in die Computergestützte Sprachverarbeitung unternommen (Language Models, Word Sense Disambiguation). Immer werden sowohl algorithmische Grundlagen als auch konkrete Anwendungen behandelt.

Die Vorlesung wird durch eine [Übung](#) begleitet. Diese vertieft die gelernten Methoden durch praktische Umsetzung. In Gruppen werden verschiedene Probleme des Information Retrieval, teilweise unter Benutzung existierende Frameworks, gelöst.

Voraussetzungen

Voraussetzung für den Besuch sind gute Kenntnisse in "Algorithmen und Datenstrukturen" und der Programmierung mit Java.

Prüfungen

Prüfungen sind mündlich.

Anrechnung

Das Modul (Vorlesung + Übung) kann angerechnet werden für

- Bachelor Informatik, 5 SP
- Bachelor INFOMIT, 5 SP

Literatur zur Vorlesung

- Schütze, Manning, Raghavan: "Introduction to Information Retrieval", MIT Press, 2009 ([Komplette Onlineversion](#))
- Weitere Literatur und Links

Themen und Termine im Einzelnen

Folien sind hier jeweils nach der Vorlesung als PDF verfügbar. Änderungen möglich. All slides are English, but the course will be held in German.

- Overview
- Introduction to Information Retrieval
- Evaluation of IR Systems; document normalization
- IR Models I: Boolean, Vector Space, Relevance Feedback
- IR Models II: Probabilistic Retrieval, Latent Semantic Indexing
- Exact online substring search: Z-Box, Boyer-Moore
- Indexing terms: Inverted files

Topics we Shall Discuss

- Evaluating IR systems
- Relevance models: Semantics of queries (IR model)
- User feedback (relevance feedback)
- Searching strings (exact, token-based, substring, ...)
- Building efficient search indexes
- Search on the web / PageRank
- Language models
- Word collocations

Related Topics we shall not Discuss

- Information Extraction, Named Entity Recognition
- Entity Search
- Personalized, social-media based, local, mobile, ... search
- Search Engine Optimization
- Detecting similar texts (plagiarism)
- Computational Linguistics
- Text classification
- Text clustering

- See lecture „Computational Natural Language Processing“
 - Maschinelle Sprachverarbeitung

Exercises

- We will form **teams**
- Five exercises, **all must be passed**
 - IMDB crawler
 - Boolean Information Retrieval the hard way
 - Information Retrieval with Lucene
 - Synonym expansion with Lucene and Wordnet
 - Significant co-occurrences
- There will be a competition
- First exercise: 23.4.2018

Questions

- Diplominformatiker?
- Bachelor?
- Semester?

- Special expectations, experiences, questions?

Feedback 2016/2017

	Konzeption	Viel neues	Lernziele	Materialien	Lehrwille	Beispiele	Klare Struktur	Verbindungen	Klar verständlich	Übung hilft	Kritische Auseinandersetzung	Gute Atmosphäre	Verständliche Antworten	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	Abweichung vom Optimum	Abweichung pro Frage	Gefühl	Warum kommen?	Studiengang	Fachsemester
2	5	6	5	6	5	5	5	5	5	4	5	6	6	3	4	5	1	1	13	0,72	0	3,4	BA	5
3	6	6	6	6	6	6	6	6	6	6	6	6	6	3	3	3	1	1	0	0,00	1	3,4	KB	5
4	6	6	6	5	6	5	5	6	5	5	5	5	5	3	3	4	2	2	11	0,61	0	3	BA	3
5	5	5	5	5	6	4	5	5	6	5	3	5	4	4	4	4	2	2	20	1,11	0	4	BA	5
6	4	5	5	4	5	5	4	5	6	6	5	6	6	3	3	3	2	2	14	0,78	1	4	BA	5
7	6	5	5	5	6	6	5	5	6	4	5	6	5	3	3	4	2	3	15	0,83	0	3,4	BA	7
8	6	6	6	3	4	5	4	5	6	6	5	6	5	3	3	5	2	2	15	0,83	2	2,3,4	BA	3
9	5	5	5	5	6	5	5	5	6	6	5	5	5	3	3	3	1	2	10	0,56	0	3,4	KB	5
10	6	5	5	4	6	6	6	6	6	5	5	6	6	3	4	6	1	2	9	0,53	1	3,4	BA	5
11	5	6	4	5,0	6	5	4	5	6	3	6	5	6	4	3	4	1	2	15	0,88	0	2,3	BA	5
12	6	5	5	6	6	6	4	6	6	4	3	5	6	1	2	3	1	2	14	0,78	3	1,2,3,4	BA	7
13	6	6	6	5	6	6	6	6	6	5	5	6	6	3	3	3	1	1	3	0,17	2	3,4	BA	5
14	5	5	4	6	6	5	5	5	6	6	4	6	5	3	3	3	2	2	12	0,71	1	4	BA	13
15	6	6	6	6	6	6	6	6	6	6	6	6	6	3	3	4	1	1	1	0,06	1	2,3,4	BA	5
16	5	5	5	5	5	4	5	5	4	4	4	5	5	3	3	5	2	2	21	1,17	0	3,4	BA	5
17	5	5	5	5	6	6	6	6	6	5	6	6	6	3	3	4	1	1	5	0,29	1	3,4	BA	5
18	6	6	5	5	6	6	5	5	6	5	6	6	6	3	3	3	1	1	5	0,28	1	3,4	KB	7
19	6	6	6	5	6	5	5	6	6	6	5	5	5	3	3	4	1	1	8	0,44	1	3,4	BA	5
20	6	6	6	6	6	6	5	5	6	5	5	6	6	3	3	3	1	1	4	0,22	0	4	BA	5
21	5	6	4	5	6	5	5	5	6	6	5	6	5	3	4	3	2	2	12	0,67	0	2,3,4	BA	5
22	6	6	6	4	6	5	5	6	6	5	4	5	5	3	4	4	2	2	13	0,72	0	1,4	BA	5
23	5	6	4	5	5	5	5	4	6	4	5	6	5	3	3	3	2	2	15	0,83	3	3,4	BA	9
24	5	6	6	5	5	6	5	5	6	4	5	5	5	3	4	4	2	2	14	0,78	0	4	BA	5
25	5	5	5	5	5	4	6	4	5	6	5	5	5	3	3	3	2	2	14	0,82	1	2,3,4	BA	7
26	6	6	6	6	6	5	5	5	6	6	6	4	5	3	3	4	1	1	6	0,40	2	3,4		7
27	6	6	6	6	6	6	6	6	6	6	6	6	6	3	3	3	1	1	0	0,00	0	3,4	M	9
28	5	6	4	4	5	6	5	5	6	4	3	5	5	4	2	3	2	1	18	1,00	2	4	BA	3
29	6	6	6	6	6	6	6	6	6	6	6	6	6	3	3	3	1	1	0	0,00	1	4	BA	5
30	Durchschnitt	5,50	5,64	5,29	5,00	5,68	5,30	5,14	5,37	5,82	5,00	4,79	5,54	5,33	3,04	3,14	3,64	1,46	1,59					
31	Wunschzahl	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	6,00	3,00	3,00	3,00	1,00	1,00					5,8
32	Abweichung	0,50	0,36	0,71	1,00	0,32	0,70	0,86	0,63	0,18	1,00	1,21	0,46	0,67	-0,04	-0,14	-0,64	-0,46	-0,59					
33		0,50	0,36	0,71	1,00	0,32	0,70	0,86	0,63	0,18	1,00	1,21	0,46	0,67	0,04	0,14	0,64	0,46	0,59	10,48				

Free Text Feedback – Room for Improvement

- Nicht so gut
 - Auf Folien mit vielen Variablen deren Bedeutung irgendwo angeben
 - Folien Englisch, Sprache Deutsch
 - Ich kann Google nicht mehr benutzen (?)
 - Fehler in Folien (aber korrigiert)
 - Verbindung VL - UE nur schwach ausgeprägt (aber Übung ist gut)
- Zu wenig
 - Aktuelle IR Forschung
- Zu viel
 - Zu viel Politik, Zeit fehlte später
 - 4 Aufwand für Übungen
 - Zu viel biomedizinische Beispiele
- Weitere Anregungen
 - Themenreihenfolge unklar
 - 2 Auf 4SWS ausbauen

Beyond Information Retrieval

Entities

The image displays three overlapping browser windows illustrating search results for different entities. The leftmost window shows search results for 'der pate', listing various entries with titles like 'Der Pate (F...)', 'Der Pate --', and 'Der Pate | T...'. The middle window shows search results for 'Francis Ford Coppola', listing entries such as 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', 'Francis Ford C...', and 'Francis Ford C...'. The rightmost window shows search results for 'Detroit', featuring a detailed information card on the right side. This card includes a map of Detroit, the title 'Detroit', the subtitle 'Großstadt, Michigan', a descriptive paragraph, and several key statistics: Fläche: 370 km², Höhe: 183 m, Ortszeit: Freitag, 06:42, Wetter: 9 °C, Wind aus NW mit 23 km/h, 88 % Luftfeuchtigkeit, Bevölkerung: 688.701 (2013), and Arbeitslosenrate: 10,2% (Apr. 2015). Below the statistics, there is a section for 'Kommende Veranstaltungen' with a table listing events like 'Green Day' and 'Luke Bryan'. At the bottom of the card, it says 'Über 25 weitere ansehen' and 'Interessante Orte' with 'Über 15 weitere ansehen'.

Entity Search

- Very often people search **information about an entity**
 - Location, person, movie, product, football player, pop singer, ...
- Entity search
 - **Detect entities** in text and build a structured **knowledge base**
 - Despite homonyms, synonyms, collocations, abbreviations, spelling variants and spelling mistakes ...)
 - Extract related facts (Wikipedia, Freebase, ...)
 - Person age, address, spouse, income, place of birth, ...
 - Detect entities in queries
 - Answer **with extracted data** (not “just” a page page)
- Which entities?
 - Today: Wikipedia

Applications in Business

- Given an incoming complaint mail: **Which product** (line) is affected?
 - Recognize and normalize product; forward mail or link to FAQ
- Given twitter etc.: **What problems** are most frequently reported by our customers?
 - Recognize and normalize “problems”; assign to product (lines)
- Improved **customer self service**
 - Entity Search for product and problem
 - Precise routing and prioritization of requests

WBI Research in Text Mining

- **Entity recognition** and search in biomedical texts
 - Genes, diseases, mutations, species, drugs, ...
- Relationships: Gene regulation, protein-protein-interaction, disease-drug-mutation ...
- Text classification: Molecular ... cancer ... colon ...
- Table similarity search
- We mostly work on **scientific literature**
 - But also web crawls, patent search, ...

High-Accuracy BTM: GeneView (Thomas et al., 2012)

The screenshot displays the GeneView web application interface in a Firefox browser. The main content area shows search results for the query "ENTREZ%3A1956&gv_result_order_desc=desc&gv_result_order=DATE&gv_page_size=20&gv_page_size=;". The results are sorted by date of publication, showing a list of 10 items. The first item is "1. Protein Kinase...".

The interface includes a sidebar with navigation options and a search bar. The main content area is divided into sections for "Options", "Annotations", and "Chemicals". The "Annotations" section shows a table of gene annotations for TP63 and TP73 genes across various species.

The "Annotations" table:

Gene (species)	Entrez	Count
TP63 (Human)	7157	41
TP63 (Mouse)	22059	23
TP63 (Human)	8626	14
TP63 (Mouse)	22061	10
TP73 (Human)	7161	4
Hic1 (Mouse)	15248	2
HIC1 (Human)	3090	1
SF1 (Human)	7536	1
EGFR (Human)	1956	1

The "Chemicals" section shows a table of chemical annotations for TP63 and TP73 genes across various species.

The "Chemicals" table:

SNP	Count
R248W	1
R283H	1
Y220C	1
R175H	1
D281G	1
R249S	1

The "Abstract" section contains the following text:

p53 Family: Role of Protein Isoforms in Human Cancer
Wei Jinxiang, Zaika Elena, Zaika Alexander
Journal of Nucleic Acids – 2012

Abstract
 TP53, TP63, and TP73 genes comprise the p53 family. Each gene produces protein isoforms through multiple mechanisms including extensive alternative mRNA splicing. Accumulating evidence shows that these isoforms play a critical role in the regulation of many biological processes in normal cells. Their abnormal expression contributes to tumorigenesis and has a profound effect on tumor response to curative therapy. This paper is an overview of isoform diversity in the p53 family and its role in cancer.

1. Introduction
 Alternative splicing allows a single gene to express multiple protein variants. It is estimated that 92–95% of human multiexon genes undergo alternative splicing [1, 2]. Abnormal alterations of splicing may interfere with normal cellular homeostasis and lead to cancer development [2–5].

The p53 protein family is comprised of three transcription factors: p53, p63, and p73. Phylogenetic analysis revealed that this family originated from a p63/73-like ancestral gene early in metazoan evolution [2, 4]. Maintenance of genetic stability of germ cells seems to be its ancestral function [2]. The p53 family regulates many vital biological processes, including cell differentiation, proliferation, and cell death/apoptosis [2, 15]. Dysregulation of the p53 family plays a critical role in tumorigenesis and significantly affects tumor response to therapy. This review summarizes current data on the regulation of p53, p63, and p73 isoforms and their roles in cancer.

2. Structure and Function
 p53, p63, and p73 genes are located on chromosomes 17p13.1, 3q27-29, and 1p36.2 amino acid sequence homology in the transactivation, DNA-binding and oligomerization domains. Evolutionally, this domain is the most conserved, suggesting that regulation of transcription is found in the oligomerization and transactivation domains (~30%).

The founding member of the p53 family, the p53 protein, had been discovered more than 20 years ago. However, when it had been found that the p63 and p73 genes encode proteins that form multiple variants.

Transcriptions of p53, p63, and p73 genes are regulated by similar mechanisms. It is well established that the p53 gene is regulated by a third putative p53 promoter (Figure 1). One study *in silico* provided evidence for the existence of a third putative p53 promoter which will be found in the future. An extensive alternative splicing of the p53 gene results in a large number of p53 isoforms. TA variant isoforms are generally categorized into two main groups, termed TA and ΔN [15, 16]. TA variant isoforms were initially thought that ΔN isoforms are only generated by the P2 promoter. However, recent analysis of alternative mRNA splicing revealed that some transcriptionally deficient isoforms are products of the P1 promoter. For example, the P1 promoter of the TP73 gene regulates TP73 isoforms and isoforms, which lack the TA domain: ΔEx2p73, ΔEx2/3p73, and ΔNp73. The latter isoforms are missing either exon 2 (ΔEx2p73) or both exon 2 and 3 (ΔEx2/3p73) or contain an additional exon 3 (ΔNp73) [17, 18]. Other ΔNp73 transcripts are products of the P2 promoter. Similar to p73, the P1 promoter of the p53 gene produces transcriptionally active isoforms [1]. The alternative splicing is responsible for transcriptionally deficient isoforms of Δ40p53, which missing the first 40 amino acids at the N-terminus [19, 20]. Additional

Detecting Gene Names

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

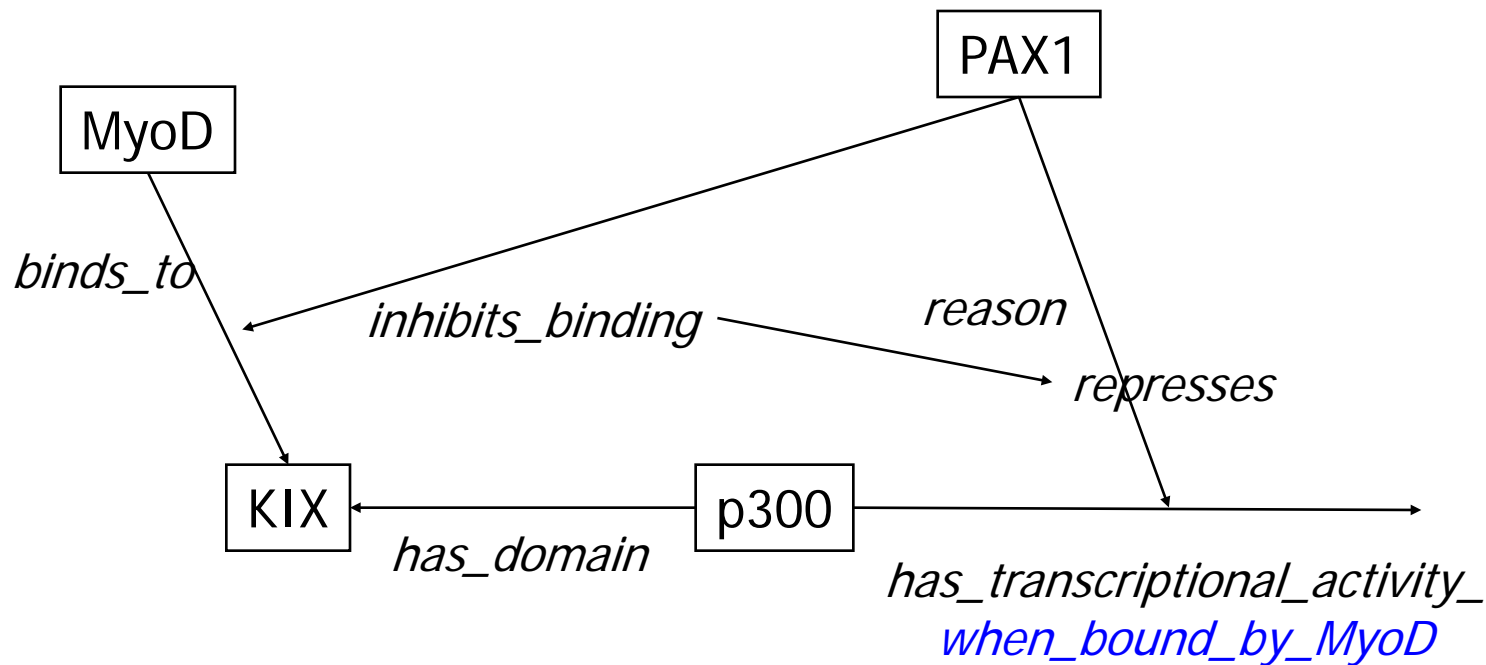
Detecting Gene Names

The human T cell leukemia lymphotropic virus type 1 Tax protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.

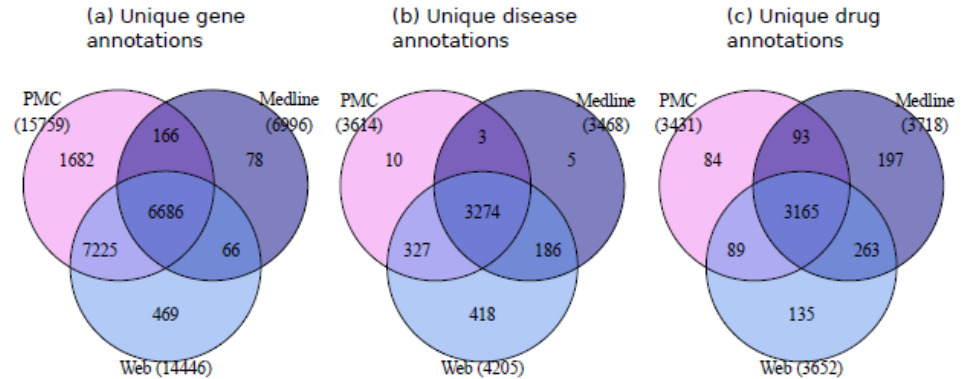
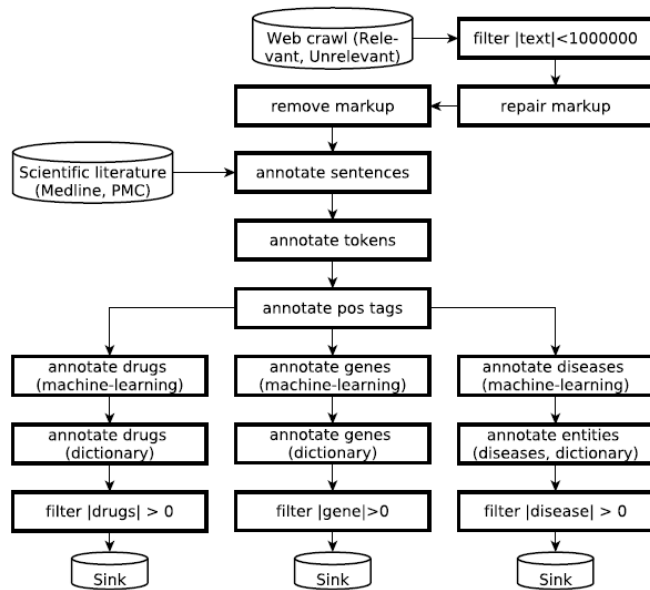
- Typical problems
 - Multi-token entities with ill-defined boundaries
 - Abbreviations
 - Synonyms, homonyms, polysemy
 - Irregular spelling, naming variations
 - ...

Beyond Entities: Understanding Text is Difficult (even for us)

„The PAX1 protein represses MyoD-dependent transcription by inhibiting MyoD-binding to the KIX domain of p300.“



Biomedical Web



rank	Drug ID, Drug name	#PMC	#Web	Log ₂ Fold	Indication
1	DB01619 Nolahist D27.505.519.625.375.425.400	0	143	-Inf	allergies and the common cold
2	D05.750.716.579.159 Cuemid D27.505.519.186.071.202	0	86	-Inf	hypercholesterolemia
3	DB00237 Butalan D27.505.954.427.210.350	0	59	-Inf	insomnia, anxiety disorders
4	D04.615.181.384.535 D27.505.954.427.700.122.055 Pamelor (Nortriptyline)	3	214	-6.52	depression, chronic pain, ...
5	D03.066.515.530 D27.505.696.377.605.600.708 Nicobion	1	65	-6.38	Nicotinamid deficiency (vitamin)
6	D03.383.129.308.207 Krebiozen (Creatine)	1	61	-6.29	nutritional supplementation
7	DB00729 Prantal	1	53	-6.09	peptic ulcer, gastric hyperacidity, hyperhidrosis
8	DB01178 Chlormezanon D27.505.954.427.210.950.015	40	1838	-5.88	anxiety, muscle spasm
9	D03.383.097.217.900 D27.505.519.124.035 Triaziquone	2	82	-5.72	chemotherapy
10	DB00241 Butalbital	3	119	-5.67	headaches, migraines, and pain
11	DB00653 Bitter salt D27.505.696.663.850.014	11	395	-5.53	magnesium deficiency, acute nephritis in children, ...

rank	Drug ID, Drug name	#PMC	#Web	Log ₂ Fold	Indication
1	D03.132.760.864 D27.505.954.122.136 Tomatine	64	2	4.64	fluorescein photoactive dye
2	D04.615.885.347.300 Phloxin-B	50	2	4.28	red-colored dye
3	D04.345.891.900 Trichodermin	49	2	4.25	acts as an antifungal and protein synthesis inhibitor
4	D04.345.891.870 Fusariotoxin	90	5	3.81	trichothecene mycotoxin
5	DB02859 Soraphen A	57	4	3.47	inhibitor of acetyl CoA carboxylase activity
6	D03.383.129.708.867 Thiadiazoles	109	9	3.24	oxidation inhibitors, cyanine dyes, metal chelating agents
7	DB02929 K201	441	39	3.14	anti-arrhythmic potential
8	D03.383.533 Oxazines	70	8	2.77	heterocyclic compounds (used as fluorescent dyes)
9	D05.750.716.822.300 Bone Cement	83	11	2.55	anchor artificial joints
10	D05.500.249.600 Phycobilisomes	83	11	2.55	light harvesting antennae from e.g. algae (used as fluorescent dyes)
11	DR00503 ΔRT-F38	86	12	2.48	HIV