



Information Retrieval

Ulf Leser

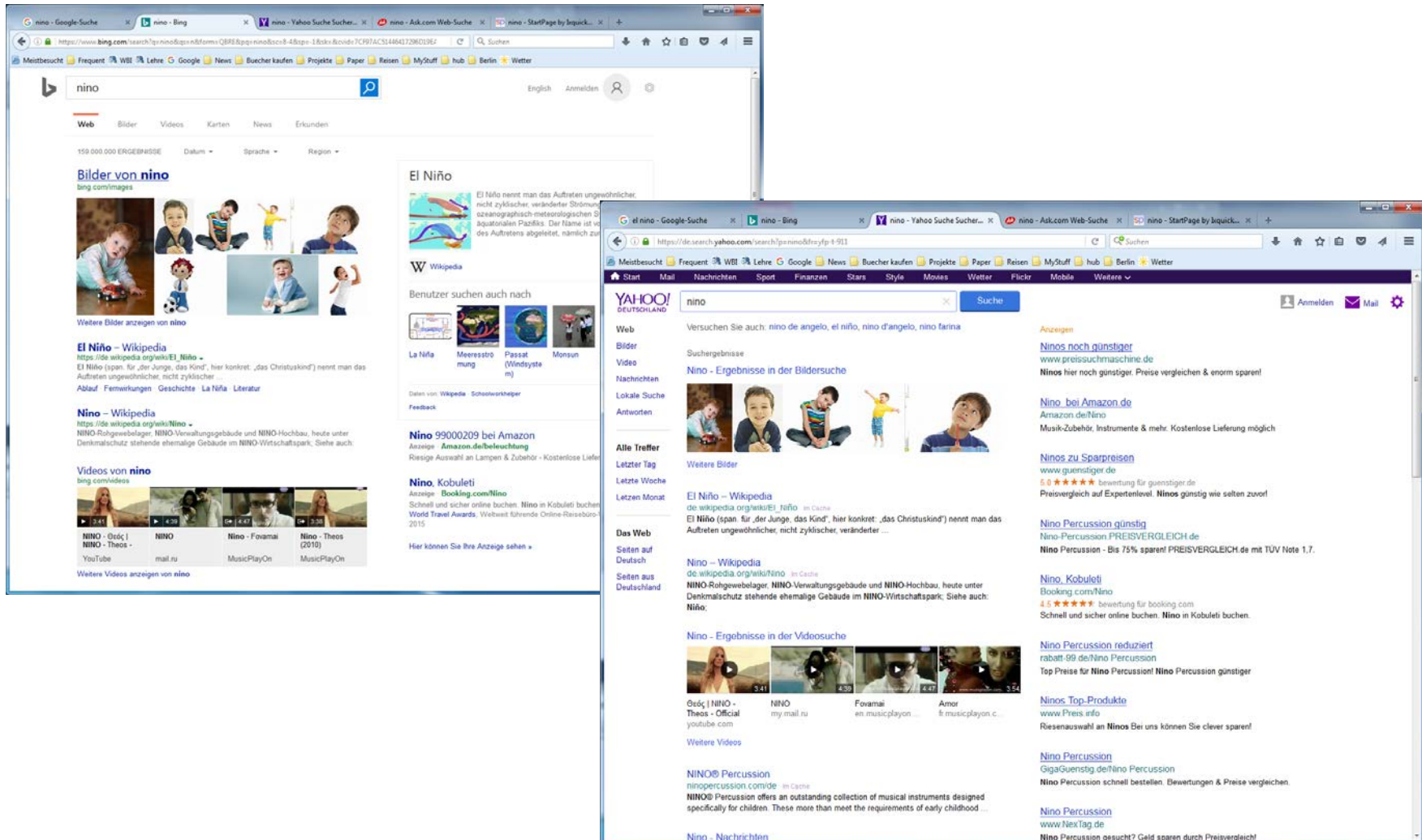
Web Search Engines

The screenshot shows a Google search for 'nino'. The search bar at the top contains 'nino'. Below the search bar, the results are categorized into 'Alle', 'Bilder', 'Videos', 'News', 'Maps', 'Mehr', and 'Suchoptionen'. The main results section shows 'Ungefähr 136.000.000 Ergebnisse (0,65 Sekunden)'. The first result is 'El Niño – Wikipedia' with a link to https://de.wikipedia.org/wiki/El_Niño. The snippet describes El Niño as a phenomenon in the Pacific Ocean. The second result is 'Nino – Wikipedia' with a link to <https://de.wikipedia.org/wiki/Nino>. The snippet describes Nino as a Spanish football player. Below the text results, there is a map showing the location of 'nino Permakultur' in Berlin. The map includes labels for 'Jüdisches Museum Berlin', 'Salumeria da Nino', 'NINO DER FRISEUR', and 'nino Permakultur'. Below the map, there are details for 'nino Permakultur' and 'Salumeria da Nino', including their addresses, phone numbers, and opening hours. At the bottom, there is a link to 'Die El Niño Infoseite' with the URL www.elnino.info/.

The screenshot shows a Bing search for 'nino'. The search bar at the top contains 'nino'. Below the search bar, the results are categorized into 'Web', 'Bilder', 'Videos', 'Karten', 'News', and 'Erkunden'. The main results section shows '159.000.000 ERGEBNISSE'. The first result is 'Bilder von nino' with a link to bing.com/images. The second result is 'El Niño – Wikipedia' with a link to https://de.wikipedia.org/wiki/El_Niño. The snippet describes El Niño as a phenomenon in the Pacific Ocean. The third result is 'Nino – Wikipedia' with a link to <https://de.wikipedia.org/wiki/Nino>. The snippet describes Nino as a Spanish football player. Below the text results, there is a section for 'Videos von nino' with a link to bing.com/videos. The video results show several videos related to 'nino'. At the bottom, there is a section for 'Nino 99000209 bei Amazon' with a link to <https://www.amazon.de/99000209>.

The screenshot shows a Google search for 'el nino'. The search bar at the top contains 'el nino'. Below the search bar, the results are categorized into 'Alle', 'News', 'Bilder', 'Videos', 'Bücher', 'Mehr', and 'Suchoptionen'. The main results section shows 'Ungefähr 17.800.000 Ergebnisse (0,65 Sekunden)'. The first result is '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.' The second result is 'El Niño – Wikipedia' with a link to https://de.wikipedia.org/wiki/El_Niño. The snippet describes El Niño as a phenomenon in the Pacific Ocean. At the bottom, there is a section for 'El Niño – Wikipedia' with a link to https://de.wikipedia.org/wiki/El_Niño.

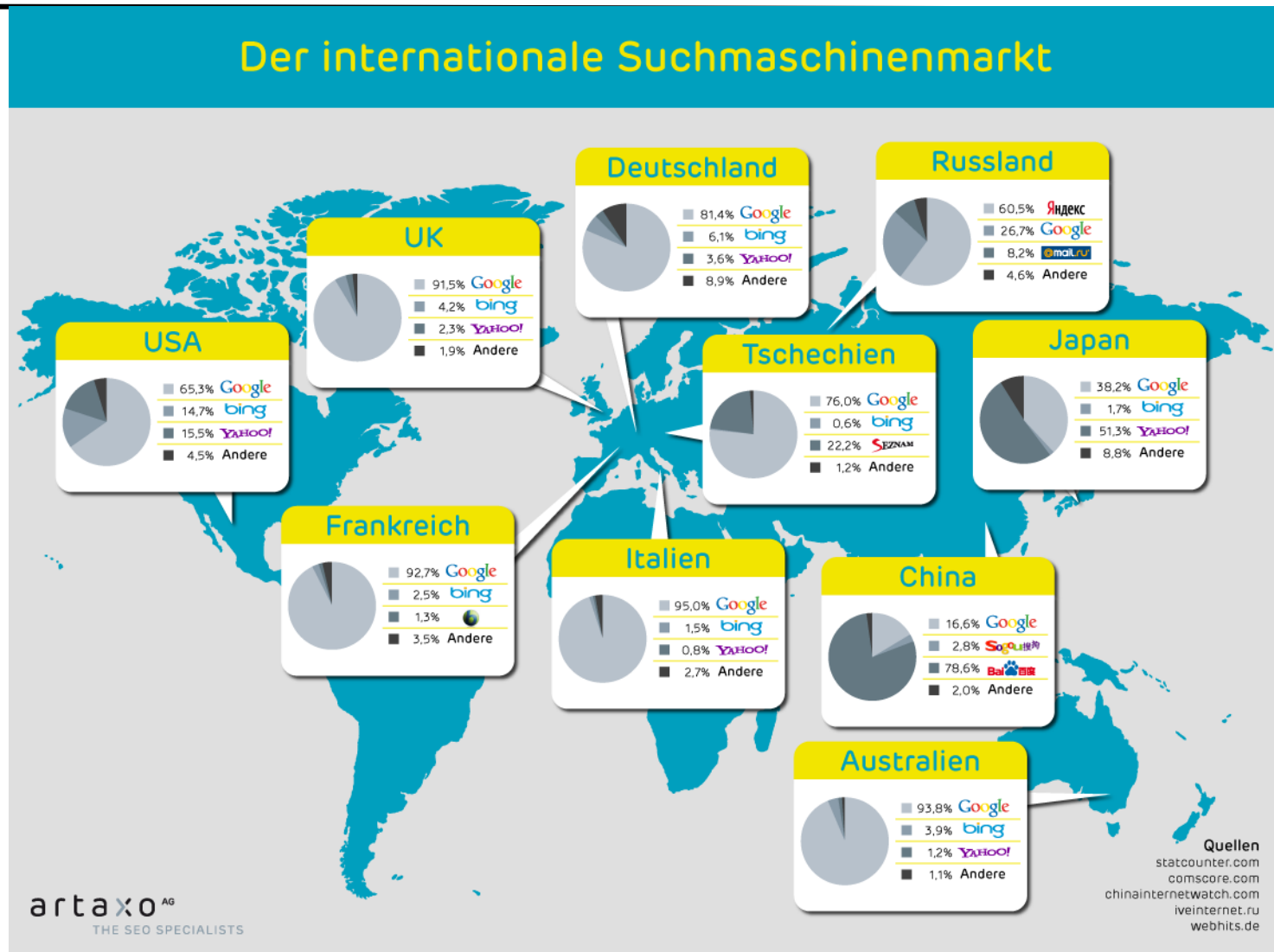
Web Search Engines



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)



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“...

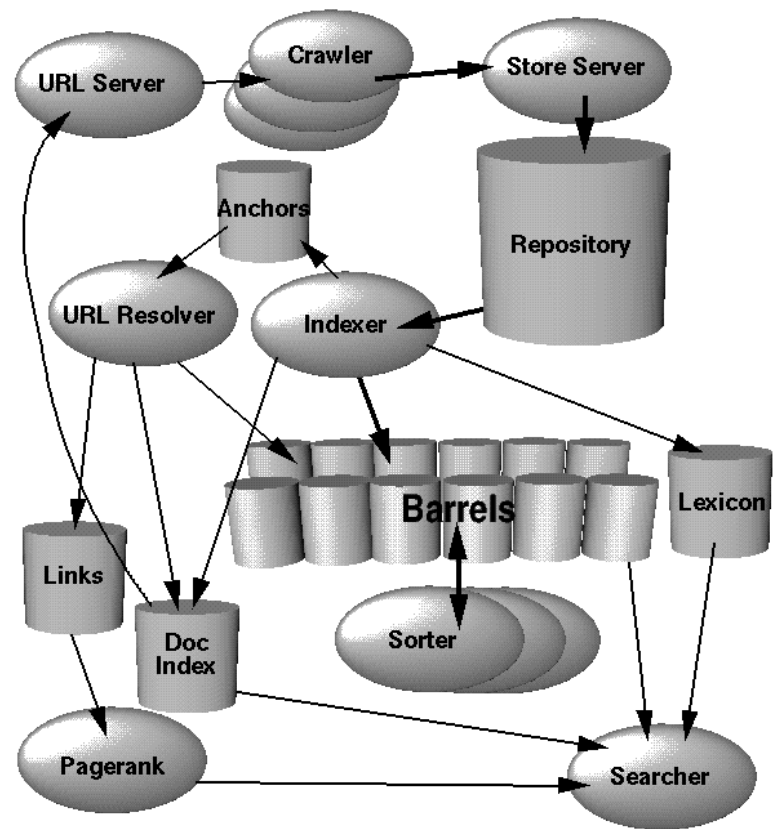
```
<html>
  blabla <a href=„http:T\index.html“>blublu</a>
</html>
```

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

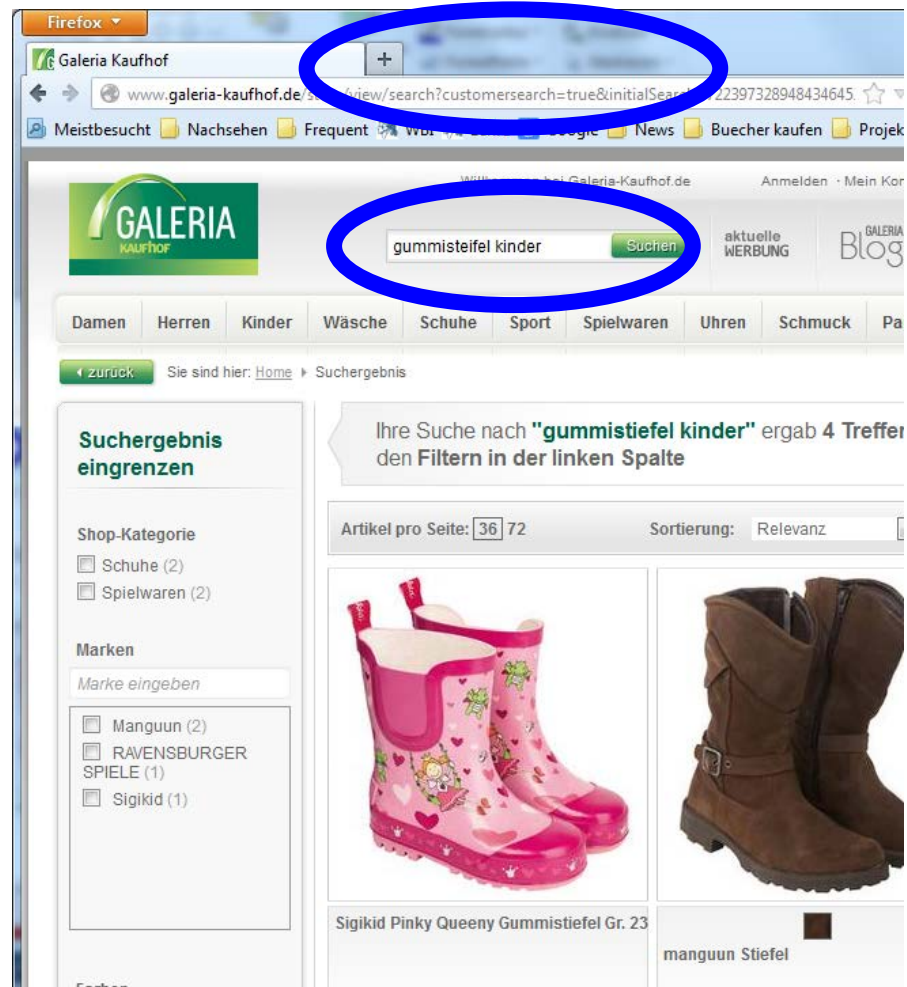
- At query time, only one server is searched – located at the search engine
- Every search engine has a (partial) **copy of the web**
- Created and maintained by a **crawler**



Careful!

- Use 1000ds of servers for parallel crawling
- No server overload (DDoS)
- Adapt frequency of visits to change rate
- Watch your bandwidth
- DNS resolution is a bottleneck (caching helps)
- Never stop
- ...

Not (easily) Indexed: The Deep Web



What is a Search result?

The image displays three side-by-side screenshots of search engine results for the query "el nino".

- Left Screenshot (Google):** Shows the Google search results page. The search bar contains "el nino". Below the search bar, it indicates "Ungefähr 71.700.000 Ergebnisse (0,42 s)". The first result is "El Niño – Wikipedia" with a link to de.wikipedia.org/wiki/El_Niño. Other results include "Die El Niño Infoseite" and "NOAA El Niño: Research, Forecasts".
- Middle Screenshot (Yahoo!):** Shows the Yahoo! search results page. The search bar contains "el nino". Below the search bar, it indicates "Suche: Das Web". The first result is "Versuchen Sie auch: el-ni" with a link to www.guenstiger.de. Other results include "El Niño noch günstiger" and "El Niño günstig".
- Right Screenshot (Bing):** Shows the Bing search results page. The search bar contains "el nino". Below the search bar, it indicates "74.700.000 ERGEBNISSE". The first result is "El Niño im Vergleich - Spitzen El Niño zu Top-Preisen" with a link to www.guenstiger.de. Other results include "El Niño noch günstiger" and "El Niño günstig".

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

- Naive: A page is relevant iff it **contains all query token**
- Disadvantages
 - <El nino> – many false positives (because homonyms – el, nino)
 - <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

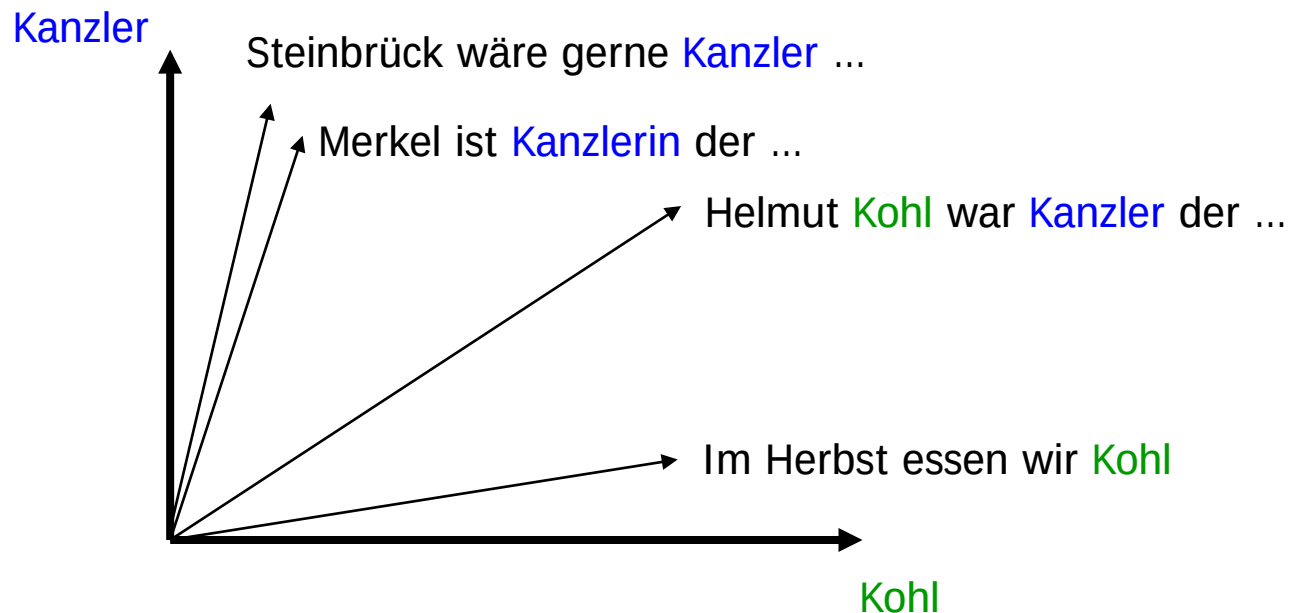
- Transform each page into a **high dimensional vector**
- Every unique token is a dimension
- Value can be binary, or count occurrences, or ...
- Vector as has many dimensions as there are **unique tokens on the web**

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

- Page with semantically similar content usually share many token
- 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 italienschen Haus blüht
3	d ₄ : Die italienschen 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>
 - „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**
 - Small queries, large 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.

The screenshot shows a web browser window with the address bar displaying https://www.informatik.hu-berlin.de/forschung/gebiete/wbi/teaching/archive/ws1415/vl_inforet/. The browser's address bar also shows the text "managing gigabytes - Goo..." and "Information Retrieval — Wi...".

The page header includes the Humboldt-Universität zu Berlin logo and the text "HUMBOLDT-UNIVERSITÄT ZU BERLIN". Below the header, there is a navigation bar with links: "Deutsch", "English", and "Kontakt/Impressum".

The main content area is titled "Information Retrieval" and is led by "Professor Dr. Ulf Leser". The text describes the course as a module that deals with methods for searching in (very large) text collections, particularly in the web. It mentions that the course will cover algorithms and procedures for text preprocessing, query languages, relevance models, indexing, and special problems in web-search engines. It also states that the course will be accompanied by exercises and that the course will be held in German.

The sidebar on the left contains a list of navigation links: "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".

The main content area also includes a section titled "Voraussetzungen" (Prerequisites) which states that the course requires good knowledge in "Algorithms and Data Structures" and programming in Java. It also includes a section titled "Prüfungen" (Examinations) which states that the exams are oral. A section titled "Anrechnung" (Credit Transfer) states that the course can be credited for the Bachelor in Informatics (5 SP) and the Bachelor in Information Systems (5 SP). A section titled "Literatur zur Vorlesung" (Literature for the Lecture) lists the course material, including "Introduction to Information Retrieval" by Schütze, Manning, and Raghavan (MIT Press, 2009) and a link to the "Onlineversion". A section titled "Themen und Termine im Einzelnen" (Topics and Dates in Detail) lists the topics: 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, and 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
- Language models
- Word collocations
- Word sense disambiguation

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: 31.10.2014

Questions

- Diplominformatiker?
 - Bachelor?
 - Semester?
-
- Special expectations, experiences, questions?

Feedback vom Letzten Mal

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
		Konzeption	Viel neues	Lernziele	Materialien	Rhetorik	Beispiele	Klare Struktur	Verbindungen	Klar verständlich	Übung hilft	Verständliche Antworten	Kritische auseinander setzen	Gute Atmosphäre	Tempo	Schwierigkeit	Arbeitsaufwand	Dozent	Vorlesung	Abweichung vom Optimum	Abweichung pro Frage	Geschlecht	Gefühlt	Warum kommen?	Studiengang	Fachsemester	
1																											
2			6	6	6	6	6	6	6	6	6	6	6	6	3	3	3	1	1		0	0,00	M	1	1,2	BA	7
3			6	5	6	6	5	6	5	5	6	4	4	6	3	2	3	1	1		9	0,50		1	2	BA	7
4			6	5	6	6	6	6	6	6	5	5	5	6	3	4	3	1	1		4	0,22	M	2	2,3	M	9
5			6	6	6	6	6	5	5	5	6	5	6	6	3	3	4	1	2		6	0,33	M	0	2,3,4	BA	9
6			5	5	5		6	6	6	5	6	5	5	5	3	3	3	2	2		10	0,59	M	1	2,4	BA	5
7			5	6	5	5	6	6	5	6	6	6	5	6	2	2	3	1	1		8	0,44		0	2,3	BA	5
8			5	6	5	5	6	6	5	6	6	5	5	6	3	3	4	1	1		7	0,39	M	0	2,3	BA	5
9			5	6	5	5	5	6	6	6	5	6	6	5	3	3	3	1	1		7	0,39		0	1,2	BA	11
10			6	6	5	4	6	5	4	5	6	5	4	6	3	3	3	1	2		11	0,61		4	2,3	BF	9
11			5	5	6	4	6	6	5	6	6	5	6	6	2	2	2	2	1		10	0,56		0	1,2,3	BA	7
12	Durchschnitt	5,50	5,60	5,50	5,22	5,80	5,80	5,30	5,60	5,90	5,20	5,20	5,80	5,70	2,80	2,80	3,10	1,20	1,30								
13	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				0,9			7,4	
14	Abweichung	0,50	0,40	0,50	0,78	0,20	0,20	0,70	0,40	0,10	0,80	0,80	0,20	0,30	0,20	0,20	-0,10	-0,20	-0,30								
15		0,50	0,40	0,50	0,78	0,20	0,20	0,70	0,40	0,10	0,80	0,80	0,20	0,30	0,20	0,20	0,10	0,20	0,30	6,88							

Freitext

- **Besonders gut**

- Art der Vermittlung
- Selbsttest
- Dozent
- Übungsaufgaben
- 4 Hochschulpolitik
- Viel Programmieren
- 3 Atmosphäre
- 2 Interessante Aufgaben
- Anwendungsorientierung
- Wiederholungen

- **Verbesserung**

- 2 Tafelbild
- Zu wenig probabilistische Methoden
- Ruhig anspruchsvoller machen
- Zu wenig aktuelle Forschung
- Folien vor VL online stellen
- Algs im Pseudocode angeben
- Erste Aufgabe behalten
- Übung: „Für Nicht-Kerninformatiker viel Aufwand“

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

Entities

The image displays three overlapping browser windows from a Google search engine, illustrating the concept of entities in information retrieval. The windows show search results for different queries, with the rightmost window being the most prominent.

Leftmost Window: Search for "der pate". Results include links to Wikipedia and IMDb for "Der Pate (F...)" and "Der Pate -- Teil II".

Middle Window: Search for "Francis Ford Coppola". Results include links to Wikipedia and IMDb for "Francis Ford Coppola" and "Der letzte Kün...".

Rightmost Window: Search for "Detroit". Results include links to the official website, Wikipedia, and various news articles. The Wikipedia entry for "Detroit" is highlighted, showing details about the city's location, population, and area.

Wikipedia Entry for Detroit:

- Detroit**
Großstadt, Michigan
- Detroit ist eine Großstadt im Südosten des US-Bundesstaates Michigan. Sie liegt direkt an der kanadischen Grenze, am Detroit River zwischen dem Lake St. Clair und dem Eriesee. [Wikipedia](#)
- 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)
- Arbeitslosenrate:** 10,2% (Apr. 2015)
- Kommende Veranstaltungen**

Mo., 24. Okt. 19:00	Green Day The Fillmore Detroit
Sa., 29. Okt.	Luke Bryan Ford Field
Sa., 22. Okt.	The Stylistics
- Über 25 weitere ansehen
- Interessante Orte** 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 **knowledge base**
 - Despite homonyms, synonyms, colocations, 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, ...

GeneView

Color GeneView

WBI

Search database by keywords:

Sort result by: Classical keyword score Page size: 20 Only results with: Genes SNPs Genes from list Personal Genes

Go

View abstract

View at PubMed

View at PubMedCentral

Annotations:

Hint: Click on a highlighted entity in the text to view additional information for it.

Genes

Mus musculus (Mouse) ☒

Homo sapiens (Human) ☒

Gene (species)	Entrez	Count
E-cadherin (Human)	999	50
E-cadherin (Mouse)	12550	6
axin (Human)	8312	1
conductin (Human)	8313	1
beta-catenin (Mouse)	12387	1
adenomatous polyposis coli (Human)	324	1
dishevelled (Human)	1855	1
Tcf (Human)	3172	1
frizzled (Human)	8321	1

Adjust ranking Export as tsv

GeneView

WBI

Search database by keywords:

Sort result by: Classical keyword score Page size: 20 Only results with: Genes SNPs Genes from list Personal Genes

Go

View abstract

View at PubMed

View at PubMedCentral

Annotations:

Hint: Click on a highlighted entity in the text to view additional information for it.

Genes

Mus musculus (Mouse) ☒

Homo sapiens (Human) ☒

Gene (species)	Entrez	Count
E-cadherin (Human)	999	50
E-cadherin (Mouse)	12550	6
axin (Human)	8312	1
conductin (Human)	8313	1
beta-catenin (Mouse)	12387	1
adenomatous polyposis coli (Human)	324	1
dishevelled (Human)	1855	1
Tcf (Human)	3172	1
frizzled (Human)	8321	1

Adjust ranking Export as tsv

E-cadherin

Species: Homo sapiens

Epithelial cadherin

UniProt: [P12830](#)

Entrez Gene: [999](#)

Pathways:

- Reactome: 3
- Kegg: 8

Annotations:

- InterPro: 9
- GO: 27

Interactions: 76 distinct

- DIP: 2
- IntAct: 7
- MINT: 4
- Intact Disease: 4
- HPRD: 72
- MIPS: 2

[Search for articles containing this gene/protein](#)

Alterations

Chen Huiping
BMC Cancer - 2

Abstract

BackgroundThe maintenance of of intercellular studied the alter heterozygosity immunohistochem E-cadherin locus likely to contrib mutation was a mutation in 280 ovary, testis, ki E-cadherin mRNA the normal mRNA cell-cell adhesio β-catenin was id exhibited positive cytoplasmic staining for β-catenin IHC. An association was found between reduced expression of E-cadherin and β-catenin. Moreover, an association was detected between reduced expression of E-cadherin and diffuse histotype.ConclusionOur results support the hypothesis that alterations of E-cadherin and β-catenin play a role in the initiation and progression of gastric cancer.

Background

E-cadherin (120 kDa; chromosome 16q) is a classical cadherin and forms the key functional component of adherence junctions between epithelial cells [1]. It is bound via a series of undercoat proteins, the catenins (α, β and γ) to the actin cytoskeleton [2]. This linkage between transmembranous cadherins and actin

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.

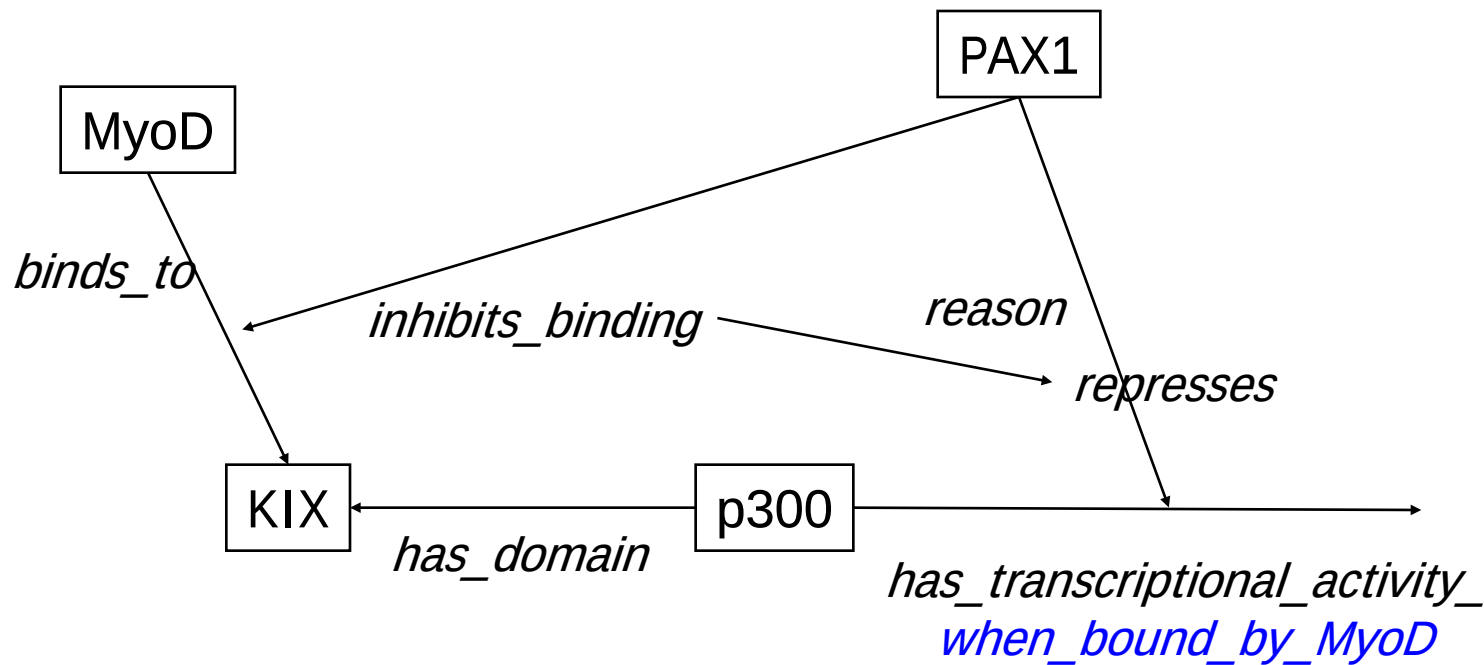
Detecting Gene Names

*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**.*

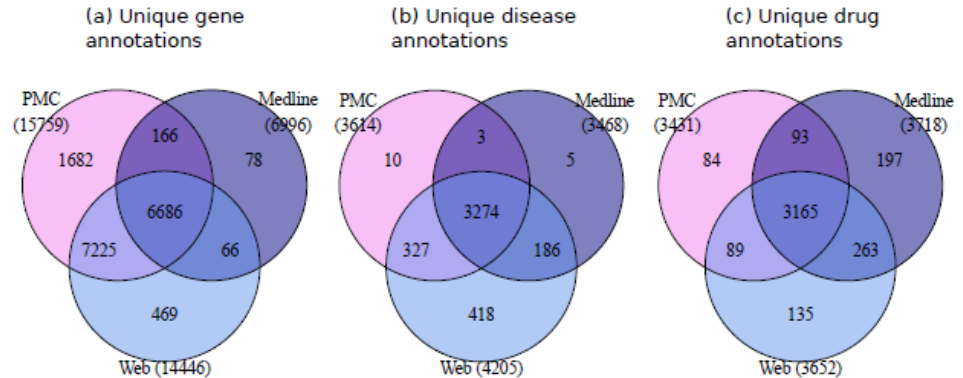
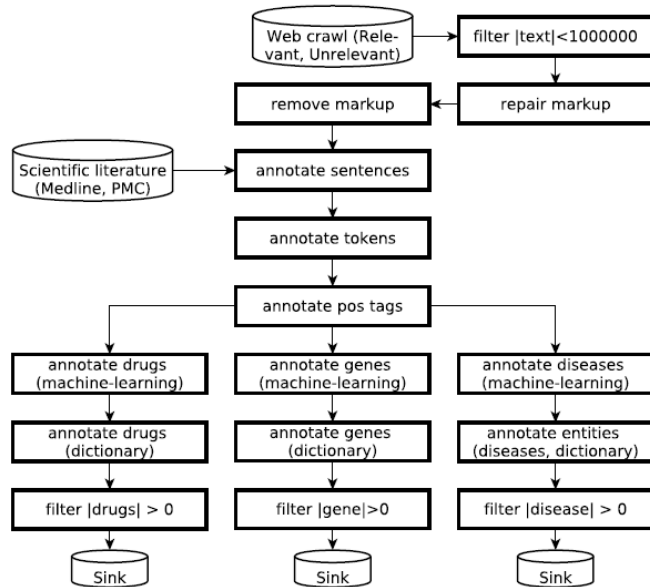
- Typical problems
 - Multi-token entities with ill-defined boundaries
 - Abbreviations
 - Synonyms, homonyms, polysemy
 - Irregular spelling, naming variations
 - ...

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 Chlormezanone 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 Sorafen 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	DB00503 ΔRT-F38	86	12	2.48	HIV