Maschinelle Sprachverarbeitung

Introduction to Information Retrieval

Ulf Leser

SHK Stelle bei WBI (~15h/Monat, 2 Jahre)

- Aufgabenbereich
  - Mitarbeit in einem Forschungsprojekt zur Analyse von Tabellen in wissenschaftlichen Texten
  - Erforschung und Entwicklung automatischer Verfahren zur Umstrukturierung tabellarischer Information
  - Entwicklung von Lernverfahren zur Ähnlichkeitsbewertung tabellarischer Daten
- Voraussetzungen
  - Studium der Informatik oder eines angrenzenden Bereichs
  - Sehr gute Kenntnisse in der Programmierung, vorzugsweise in Java oder Python
  - Gute Englischkenntnisse, mündlich und schriftlich
  - Kenntnisse in Verfahren der statistischen Sprachverarbeitung, der statistischen Datenanalyse, und des Maschinellen Lernens

Content of this Lecture

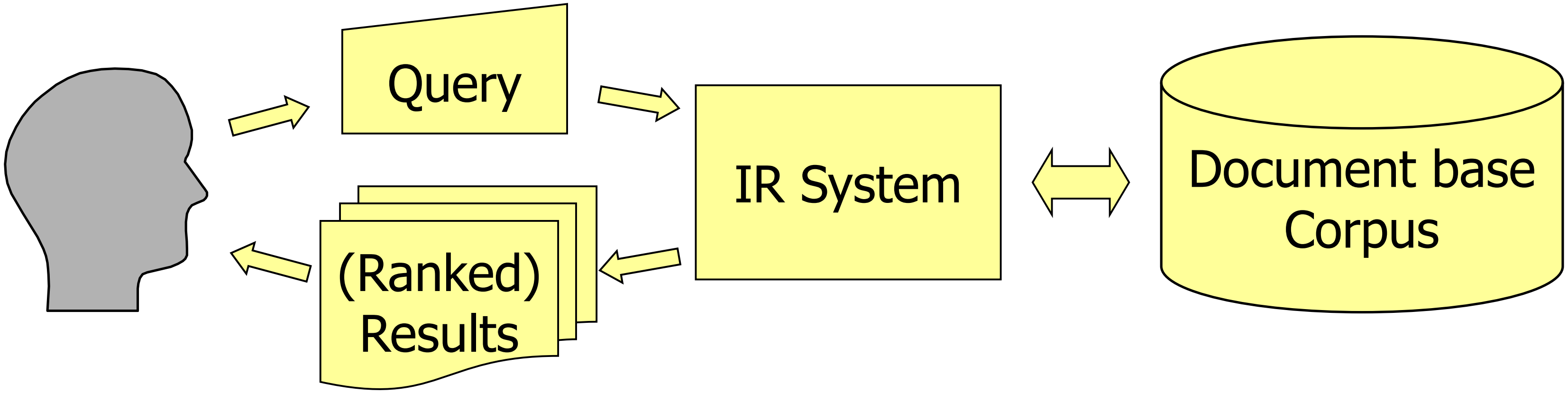
- What is Information Retrieval
- Documents & Queries
- Text Preprocessing
- Evaluating IR Systems

Information Retrieval (aka “Search”)

- Naïve: Find all documents containing the following words
- Advanced: „Leading the user to those documents that will best enable him/her to satisfy his/her need for information“[Robertson 1981]
  - A user wants to know something
  - The user needs to tell the machine what he wants to know: query
  - Posing exact queries is difficult: room for interpretation
  - Machine interprets query to compute the (hopefully) best answer
  - Goodness of answer (relevance) depends on original intention of user, not on the query
  - “Leading”: Sensible ranking of all potentially relevant docs

The Problem

- Help user in quickly finding the requested information from a given set of documents
  - Documents: Corpus, library, collection, …
  - Quickly: Few queries, fast responses, simple interfaces, …
  - Requested: The "best-fitting" documents; the "right" passages; the most "relevant" content



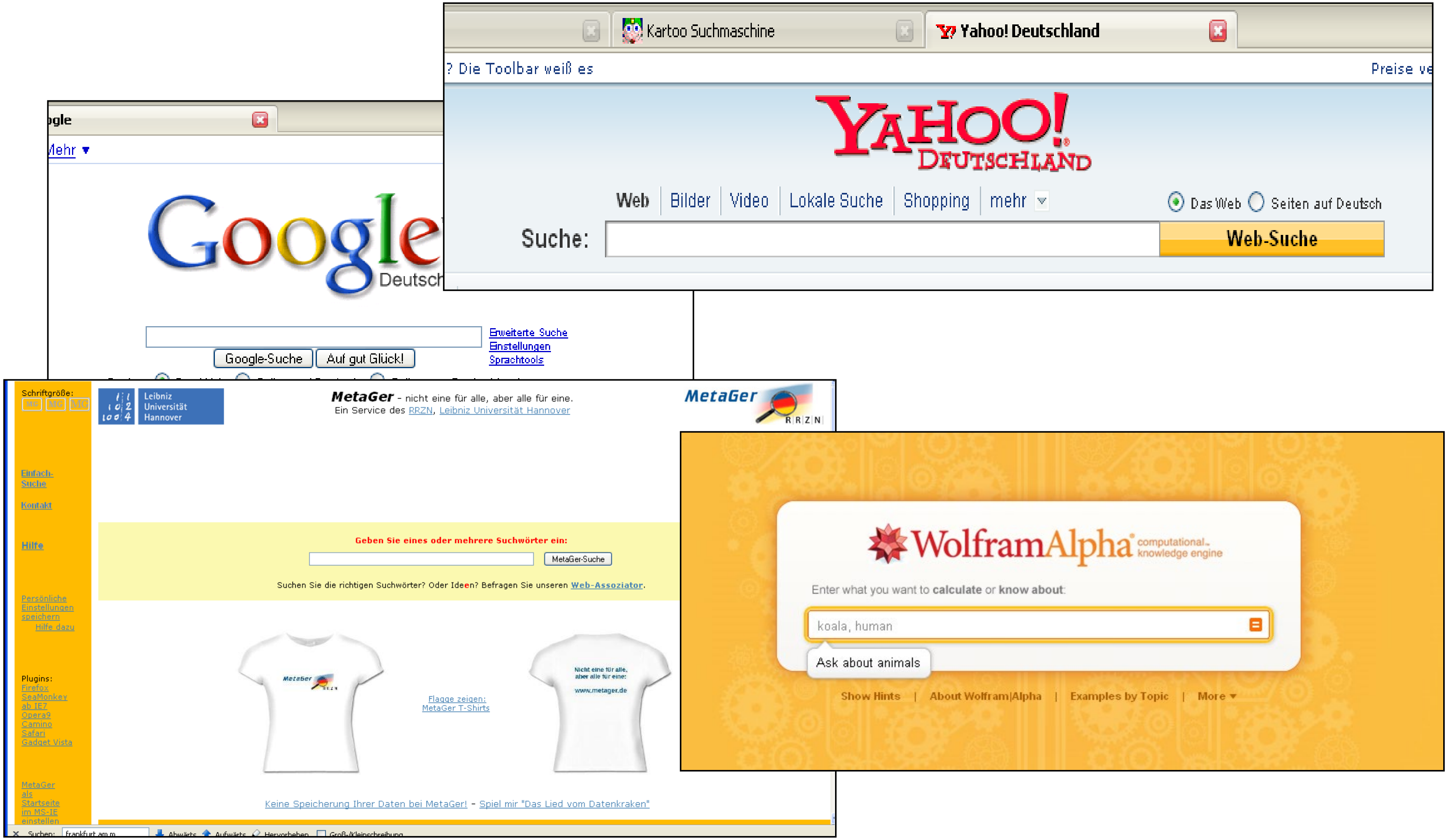
Why is it hard?

- Homonyms (context)
  - Fenster (Glas, Computer, Brief, …), Berlin (BRD, USA, …), Boden (Dach, Fussboden, Ende von etwas, …), …
- Synonyms
  - Computer, PC, Rechner, Server, Kiste, Bolide, Knoten, Desktop, …
- Specific queries (subfield Question Answering)
  - What was the score of Bayern München versus Stuttgart in the DFB Pokal finals in 1998? Who scored the first goal for Stuttgart?
  - How many hours of sunshine on average has a day in Crete in May?
- Typical web queries have 1,6 terms
- “Information broker” was (is?) a profession

Quickly

- Time to execute a query
  - Indexing, parallelization, compression, …
- Time to answer the request (may involve multiple queries)
  - Understand request, find best matches
  - Success of search engines: Better results (and fast!)
  - Process-orientation: Exploit user feedback, query history, …
- Information overload
  - If the corpus is large, ranking is a must
  - Result summarization, result clustering
  - Different search modes: What's new? What's certain?

Prominent Systems: Web Search Engines



MetaGer - nicht eine für alle, aber alle für eine.
Ein Service des RRZN, Leibniz Universität Hannover

Geben Sie eines oder mehrere Suchwörter ein:

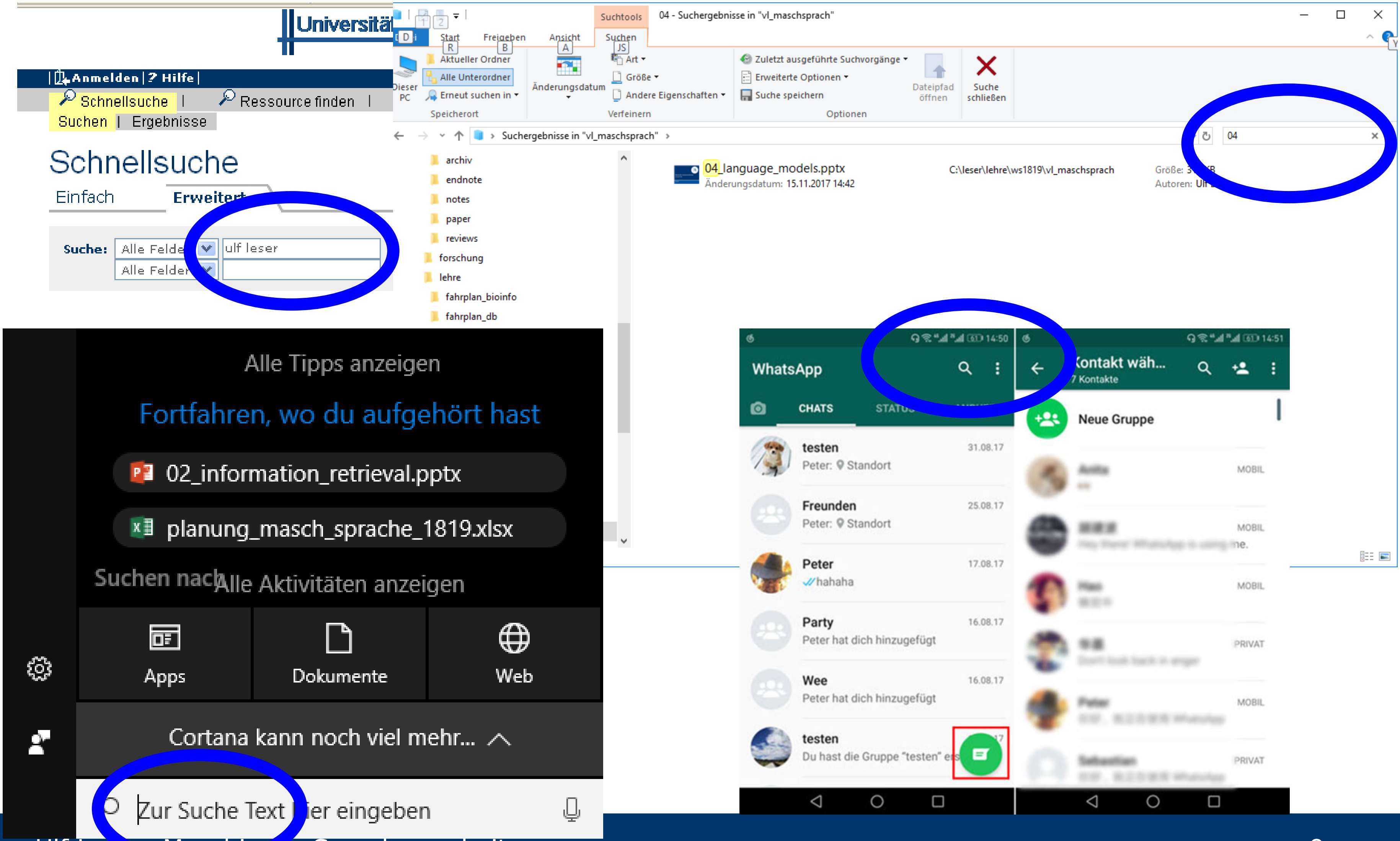
Suchen Sie die richtigen Suchwörter? Oder Ideen? Befragen Sie unseren Web-Assoziator.

Nicht eine für alle, aber alle für eine: www.metager.de

Keine Speicherung Ihrer Daten bei MetaGer! - Spiel mir "Das Lied vom Datenkraken"

Enter what you want to calculate or know about:

Other



Universität

| Anmelden | ? Hilfe |

Schnellsuche | Ressource finden |

Suchen | Ergebnisse

Schnellsuche

Einfach | Erweitert

Suche: Alle Felder ulf leser

Alle Felder

Suchtools | 04 - Suchergebnisse in "vl_maschsprach"

Start | Freigeben | Ansicht | Suchen

Dieser PC | Aktueller Ordner | Alle Unterordner | Erneut suchen in ▾ | Speicherort

Änderungsdatum ▾ | Art ▾ | Größe ▾ | Andere Eigenschaften ▾ | Verfeinern

Zuletzt ausgeführte Suchvorgänge ▾ | Erweiterte Optionen ▾ | Suche speichern | Optionen

Dateipfad öffnen | Suche schließen

› Suchergebnisse in "vl_maschsprach" ›

04

- archiv
- endnote
- notes
- paper
- reviews
- forschung
- lehre
- fahrplan_bioinfo
- fahrplan_db

04_language_models.pptx
Änderungsdatum: 15.11.2017 14:42
C:\leser\lehre\ws1819\vl_maschsprach
Größe: 3 KB
Autoren: Ulf

Alle Tipps anzeigen

Fortfahren, wo du aufgehört hast

02_information_retrieval.pptx

planung_masch_sprache_1819.xlsx

Suchen nach Alle Aktivitäten anzeigen

Apps | Dokumente | Web

Cortana kann noch viel mehr...

Zur Suche Text hier eingeben

WhatsApp

CHATS | STATUS

testen Peter: Standort	31.08.17
Freunden Peter: Standort	25.08.17
Peter ✓✓hahaha	17.08.17
Party Peter hat dich hinzugefügt	16.08.17
Wee Peter hat dich hinzugefügt	16.08.17
testen Du hast die Gruppe "testen" er	17

Kontakt wäh...
7 Kontakte

Neue Gruppe

MOBIL

MOBIL

MOBIL

PRIVAT

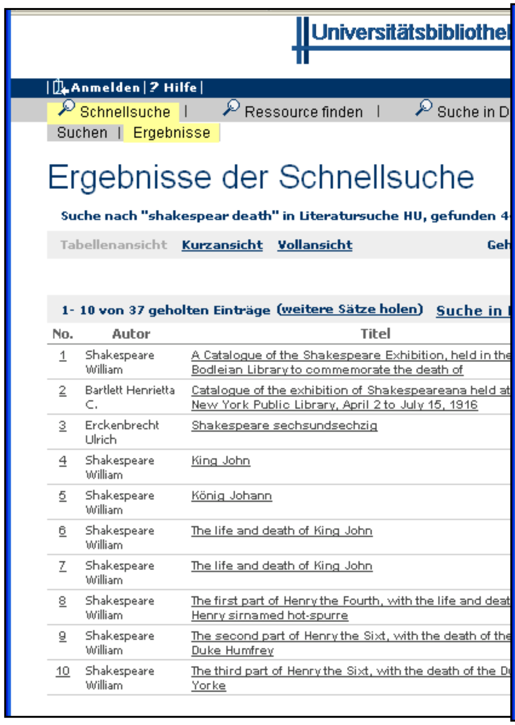
MOBIL

PRIVAT

Content of this Lecture

- What is Information Retrieval
- Documents & Queries
- Text Preprocessing
- Evaluating IR Systems

Document or Passage



Universitätsbibliothe

| Anmelden | ? Hilfe |

Schnellsuche | Ressource finden | Suche in D

Suchen | Ergebnisse

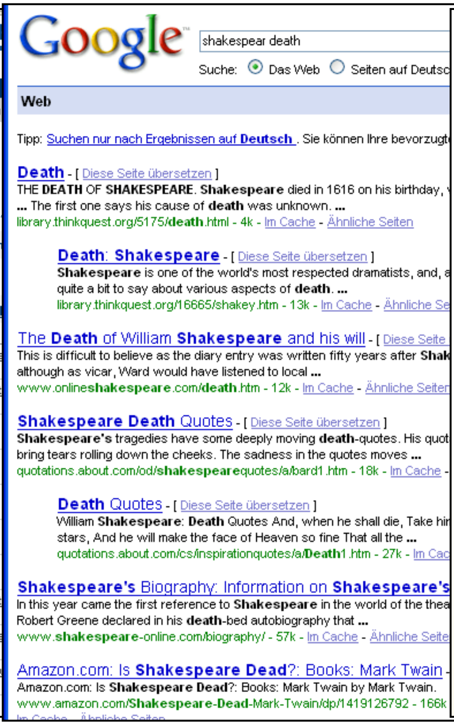
Ergebnisse der Schnellsuche

Suche nach "shakespear death" in Literatursuche HU, gefunden 4

Tabellenansicht Kurzansicht Vollansicht Geh

1- 10 von 37 geholten Einträge (weitere Sätze holen) Suche in

No.	Autor	Titel
1	Shakespeare William	A Catalogue of the Shakespeare Exhibition, held in the Bodleian Library to commemorate the death of
2	Bartlett Henrietta C.	Catalogue of the exhibition of Shakespeareana held at New York Public Library, April 2 to July 15, 1916
3	Erckenbrecht Ulrich	Shakespeare sechsundsechzig
4	Shakespeare William	King John
5	Shakespeare William	König Johann
6	Shakespeare William	The life and death of King John
7	Shakespeare William	The life and death of King John
8	Shakespeare William	The first part of Henry the Fourth, with the life and deat Henry sirnamed hot-spurre
9	Shakespeare William	The second part of Henry the Sixt, with the death of the Duke Humfrey
10	Shakespeare William	The third part of Henry the Sixt, with the death of the D Yorke



Google

shakespear death

Suche: ● Das Web ○ Seiten auf Deutsc

Web

Tipp: Suchen nur nach Ergebnissen auf **Deutsch** . Sie können Ihre bevorzugt

Death - [Diese Seite übersetzen]
THE **DEATH** OF **SHAKESPEARE**. **Shakespeare** died in 1616 on his birthday,
... The first one says his cause of **death** was unknown. **...**
library.thinkquest.org/5175/**death**.html - 4k - Im Cache - Ähnliche Seiten

Death: **Shakespeare** - [Diese Seite übersetzen]
Shakespeare is one of the world's most respected dramatists, and, a
quite a bit to say about various aspects of **death**. **...**
library.thinkquest.org/16665/shakey.htm - 13k - Im Cache - Ähnliche Se

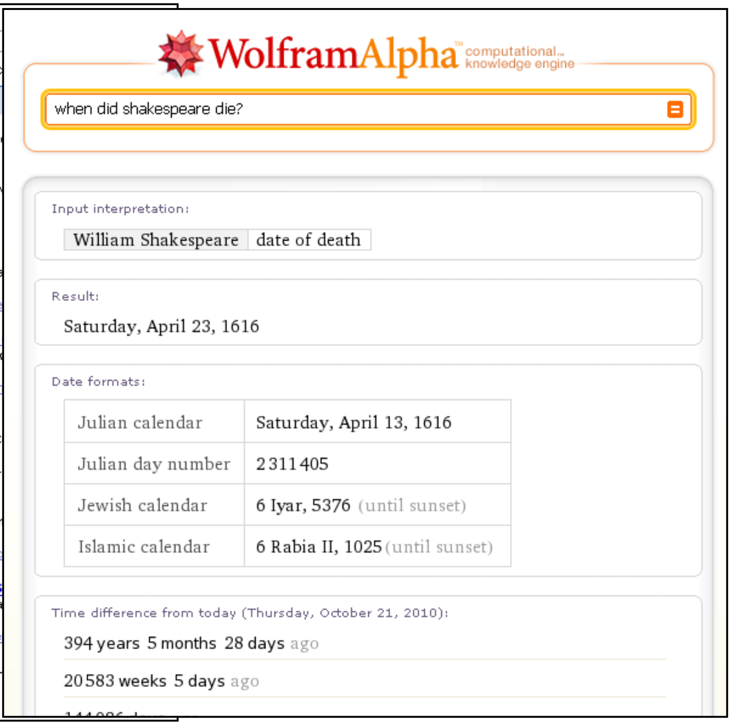
The **Death** of William **Shakespeare** and his will - [Diese Seite
This is difficult to believe as the diary entry was written fifty years after **Shak**
although as vicar, Ward would have listened to local **...**
www.online**shakespeare**.com/**death**.htm - 12k - Im Cache - Ähnliche Seiten

Shakespeare Death Quotes - [Diese Seite übersetzen]
Shakespeare's tragedies have some deeply moving **death**-quotes. His quot
bring tears rolling down the cheeks. The sadness in the quotes moves **...**
quotations.about.com/od/**shakespeare**quotes/a/bard1.htm - 18k - Im Cache -

Death Quotes - [Diese Seite übersetzen]
William **Shakespeare**: **Death** Quotes And, when he shall die, Take hir
stars, And he will make the face of Heaven so fine That all the **...**
quotations.about.com/cs/inspirationquotes/a/**Death**1.htm - 27k - Im Cac

Shakespeare's Biography: Information on **Shakespeare's**
In this year came the first reference to **Shakespeare** in the world of the thea
Robert Greene declared in his **death**-bed autobiography that **...**
www.**shakespeare**-online.com/biography/ - 57k - Im Cache - Ähnliche Seite

Amazon.com: Is **Shakespeare Dead**?: Books: Mark Twain -
Amazon.com: Is **Shakespeare Dead**?: Books: Mark Twain by Mark Twain.
www.amazon.com/**Shakespeare-Dead**-Mark-Twain/dp/1419126792 - 166k
Im Cache - Ähnliche Seiten



WolframAlpha computational knowledge engine

when did shakespeare die?

Input interpretation:

William Shakespeare | date of death

Result:

Saturday, April 23, 1616

Date formats:

Julian calendar	Saturday, April 13, 1616
Julian day number	2311405
Jewish calendar	6 Iyar, 5376 (until sunset)
Islamic calendar	6 Rabia II, 1025 (until sunset)

Time difference from today (Thursday, October 21, 2010):

394 years 5 months 28 days ago

20583 weeks 5 days ago

Searching only metadata

Searching tokens within documents

Interpreting natural text

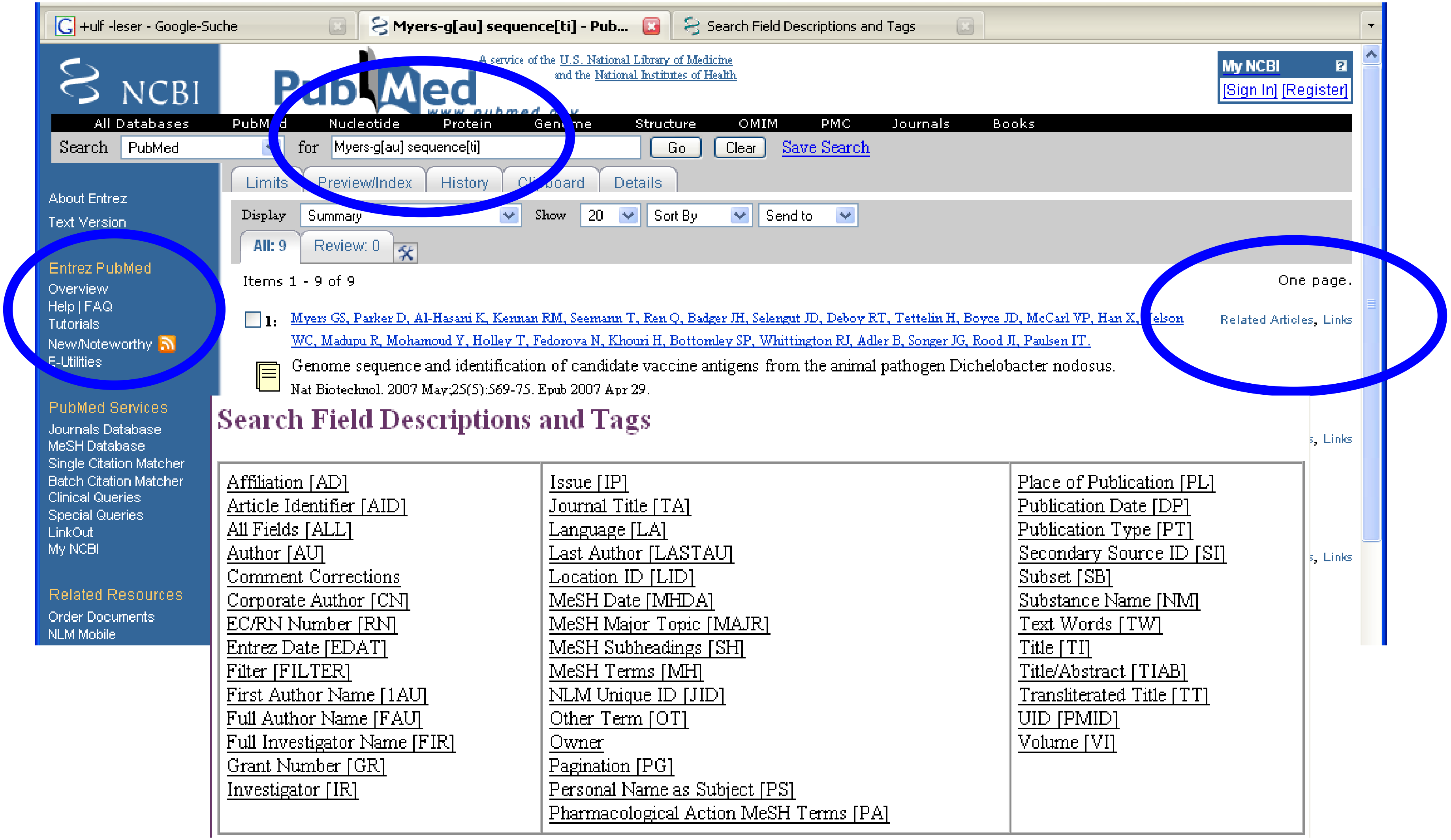
Documents

- This lecture: Natural language text
- Might be grammatically correct (books, newspapers) or not (Blogs, Twitter, spoken language)
- May have structure (title, abstract, chapters, …) or not
- May have associated (explicit or in-text) metadata or not
- May be in many different languages or even mixed
  - Foreign characters
- May refer to other documents (hyperlinks)
- May have various formats (ASCII, PDF, DOC, XML, …)

IR Queries

- Elements of IR-style queries
  - Keywords, phrases
  - Logical operations (AND, OR, NOT, ...)
    - Web search: "-ulf +leser"
  - Structured queries on metadata (author=… AND title~ …)
- Documents as queries: Find documents similar to this one
- Query refinement based on previous results
  - Find documents matching the new query within the result set of the previous search
  - Use relevant answers from previous queries to create next query

Searching with Metadata (PubMed/Medline)



Search: PubMed for Myers-g[au] sequence[ti]

Items 1 - 9 of 9

One page.

□ 1: Myers GS, Parker D, Al-Hasani K, Kennan RM, Seemann T, Ren Q, Badger JH, Selengut JD, Deboy RT, Tettelin H, Boyce JD, McCarl VP, Han X, Nelson WC, Madupu R, Mohamoud Y, Holley T, Fedorova N, Khouri H, Bottomley SP, Whittington RJ, Adler B, Songer JG, Rood JI, Paulsen IT.
Genome sequence and identification of candidate vaccine antigens from the animal pathogen Dichelobacter nodosus.
Nat Biotechnol. 2007 May;25(5):569-75. Epub 2007 Apr 29.

Search Field Descriptions and Tags

Affiliation [AD]
Article Identifier [AID]
All Fields [ALL]
Author [AU]
Comment Corrections
Corporate Author [CN]
EC/RN Number [RN]
Entrez Date [EDAT]
Filter [FILTER]
First Author Name [1AU]
Full Author Name [FAU]
Full Investigator Name [FIR]
Grant Number [GR]
Investigator [IR]

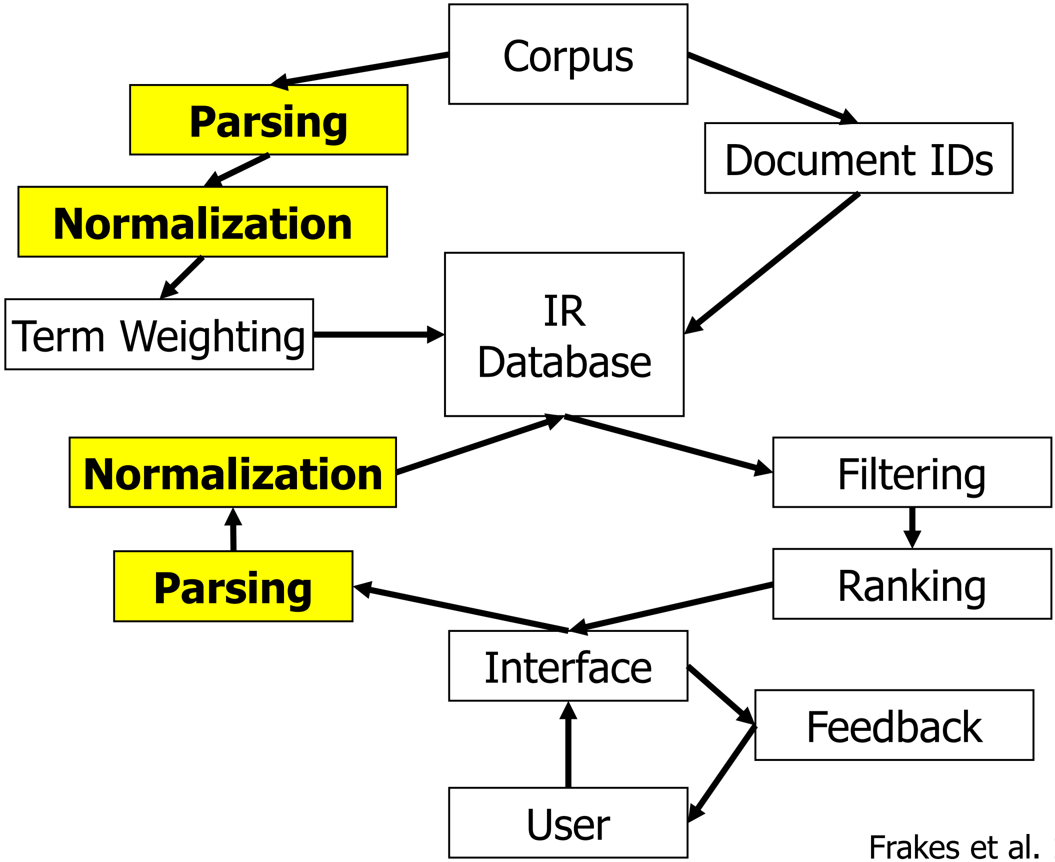
Issue [IP]
Journal Title [TA]
Language [LA]
Last Author [LASTAU]
Location ID [LID]
MeSH Date [MHDA]
MeSH Major Topic [MAJR]
MeSH Subheadings [SH]
MeSH Terms [MH]
NLM Unique ID [JID]
Other Term [OT]
Owner
Pagination [PG]
Personal Name as Subject [PS]
Pharmacological Action MeSH Terms [PA]

Place of Publication [PL]
Publication Date [DP]
Publication Type [PT]
Secondary Source ID [SI]
Subset [SB]
Substance Name [NM]
Text Words [TW]
Title [TI]
Title/Abstract [TIAB]
Transliterated Title [TT]
UID [PMID]
Volume [VI]

Content of this Lecture

- What is Information Retrieval
- Documents & Queries
- Text Preprocessing
  - Special characters and case
  - Sentence splitting
  - Tokenization
  - Stop words
  - Zipf’s law
- Evaluating IR Systems

Processing Pipeline



Frakes et al. 1992

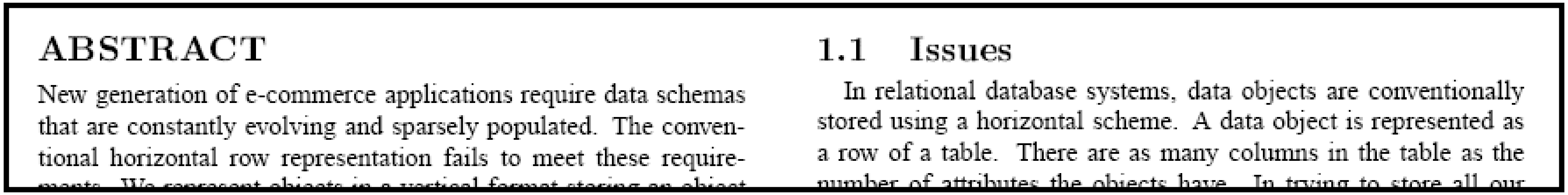
Logical View

- Definition
  - *The logical view of a document denotes its representation inside the IR system*
- Determines what users can address in a query
  - Only metadata, only title, only abstract, full text, phrases, stop words, special characters, …
- Creating the logical view involves transformations
  - Stemming, stop word removal
  - Transformation of special characters (Umlaute, Greek letters, …)
  - Removal of formatting information (HTML), tags (XML), …
  - Bag of words (BoW)
    - Arbitrary yet fixed order (e.g. sorted alphabetically)
    - See next lecture

Definitions

- Definition
  - *A document as a sequence of sentences*
  - *A sentence is a sequence of tokens*
  - *A token is the smallest unit of text (words, numbers, …)*
  - *A concept is the mental representation of a "thing"*
  - *A term is a token or a set of tokens representing a concept*
    - *"San" is a token, but not a term*
    - *"San Francisco" has two tokens but is only one term*
    - *Dictionaries usually contain terms, not tokens*
  - *A homonym is a term representing multiple concepts*
  - *A synonym is a term representing a concept which may also be represented by other terms*
  - *A syn-set is a set of synonyms representing the same concept*
- "Word" can denote either a token or a term
- We will mostly make no difference between token and terms (sadly …)

Format Conversion



ABSTRACT

New generation of e-commerce applications require data schemas that are constantly evolving and sparsely populated. The conventional horizontal row representation fails to meet these requirements. We represent objects in a vertical format storing an object

1.1 Issues

In relational database systems, data objects are conventionally stored using a horizontal scheme. A data object is represented as a row of a table. There are as many columns in the table as the number of attributes the objects have. In trying to store all our

New generation of e-commerce applications require data schemas In relational database systems, data objects are conventionally that are constantly evolving and sparsely populated. The conven- stored using a horizontal scheme. A data object is represented as tional horizontal row representation fails to meet these require- …

- Transform PDF, XML, DOC, … into ASCII / UniCode
- Problems: Formatting instruction, special characters, formulas, tables, section headers, footnotes, …
- Web: Find the net content (no ads, navigation bars, …)
- Diplomacy: To what extend can one reconstruct the original document from its logical view?

Case – A Difficult Case

- Should all text be converted to lower case letters?
- Advantages
  - Makes queries simpler
  - Decreases index size
  - Allows for some "fuzziness" in search
- Disadvantages
  - No abbreviations
  - Loss of important hints for sentence splitting
  - Loss of important hints for tokenization, NER, …
  - Loss of semantic info (proper names, Essen versus essen,…)
- Different impact in different languages (German / English)
- Often: Convert only after all other preprocessing steps

Sentence Splitting

- Most linguistic analysis works on sentence level
- Sentences group together entities and statements
- Naive approach: Reg-Exp search for "[.?!;] "
  - (note the blank!)
  - Abbreviations
    - "C. Elegans is a worm which …"; "This does not hold for the U.S."
  - Errors (due to previous normalization steps)
    - "is not clear.Conclusions.We reported on …"
  - Proper names
    - ".NET is a technique for …"
  - Direct speech
    - "By saying "It is the economy, stupid!", B. Clinton meant that …"
  - …

Tokenization

- Fundamental elements of all IR query languages are tokens
- Simple approach: search for „ „ (blanks)
  - “A state-of-the-art Z-9 Firebird was purchased on 3/12/1995.”
  - „SQL commands comprise SELECT … FROM … WHERE clauses; the latter may contain functions such as leftstr(STRING, INT).“
  - “This LCD-TV-Screen cost 3,100.99 USD.”
  - “[Bis[1,2-cyclohexanedionedioximato(1-)-O]-[1,2-cyclohexanedione dioximato(2-)-O]methyl-borato(2-)-N,N0,N00,N000,N0000,N00000)-chlorotechnetium) belongs to a family of …“
- Typical approach (but many (domain-specific) variations)
  - Treat hyphens / parentheses as blanks
  - Remove “.” (after sentence splitting)

Stop Words

- Words that are so frequent that their removal (hopefully) does not change the meaning of a doc
  - English: Top-2: 10% of all tokens; Top6: 20%; Top-50: 50%
  - English (top-10; LOB corpus): the, of, and, to, a, in, that, is, was, it
  - German(top-100): aber, als, am, an, auch, auf, aus, bei, bin, …
- Consequences
  - Removing top-100 stop words reduces a positional index by ~40%
  - Hopefully increases precision due to less spurious hits
  - Makes many phrase queries impossible
- Variations
  - Remove top 10, 100, 1000, … words
  - Language-specific, domain-specific, corpus-specific

Example

The children of obese and overweight parents have an increased risk of obesity. Subjects with two obese parents are fatter in childhood and also show a stronger pattern of tracking from childhood to adulthood. As the prevalence of parental obesity increases in the general population the extent of child to adult tracking of BMI is likely to strengthen.

↓ 100 stop words

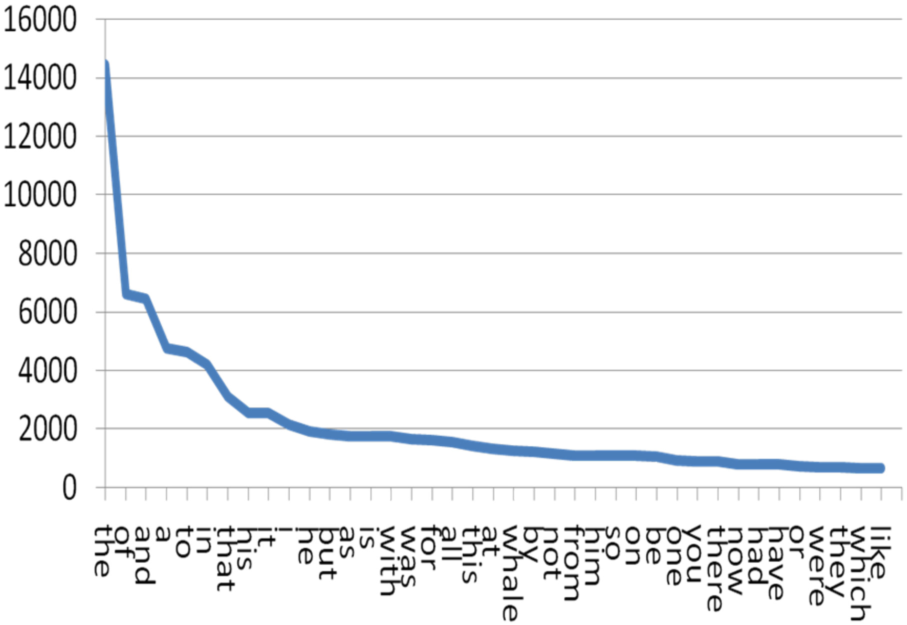
children obese overweight parents increased risk obesity. Subjects obese parents fatter childhood show stronger pattern tracking childhood adulthood. prevalence parental obesity increases general population extent child adult tracking BMI likely strengthen.

↓ 10 000 stop words

obese overweight obesity obese fatter adulthood prevalence parental obesity BMI

Zipf‘s Law

- Let f be the frequency of a word and r its rank in the list of all words sorted by frequency
- Zipf’s law: f ~ k/r for some constant k
- Example
  - Word ranks in Moby Dick
  - Good fit to Zipf’s law
  - Some domain-dependency (whale)
- Fairly good approximation for most corpora



Source: http://searchengineland.com/the-long-tail-of-search-12198

Content of this Lecture

- What is Information Retrieval
- Documents & Queries
- Text Preprocessing
- Evaluating IR Systems

Evaluation: Binary Model

- Assume that for a given query q and many docs $d \in D$, somebody determines whether d is relevant for q or not
  - An expert? An average user?
- The IR systems returns all docs it thinks that are relevant
  - No ranking for now
- Let T be the set of all truly relevant docs, X the set of all returned docs: |T|=TP+FN, |X|=TP+FP

	Truth: relevant	**Truth: not relevant**
IR: relevant	True positives	False positives
IR: not relevant	False negatives	True negatives

Precision and Recall

- Precision = TP/(TP+FP)
  - What is the fraction of relevant answers in X?
- Recall = TP/(TP+FN)
  - What is the fraction of found answers in T?
- The perfect world

	Real: Positive	Real: Negative
IR: Positive	A	0
IR: Negative	0	B

Example

- Let |D| = 10.000, |X|=15, |T|=20, |X∩T|=9

	Real: Positive	Real: Negative
IR: Positive	TP = 9	FP = 6
IR: Negative	FN = 11	TN= 9.974

  - Precision = TP/(TP+FP) = 9/15 = 60%
  - Recall = TP/(TP+FN) = 9/20 = 45%

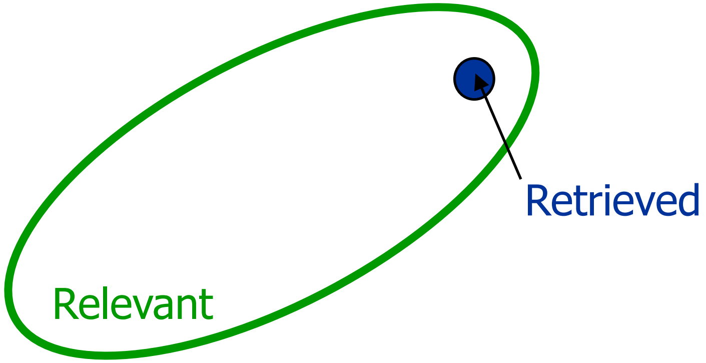
- Assume another result: |X|=10, |X∩T|=7

	Real: Positive	Real: Negative
IR: Positive	TP = 7	FP = 3
IR: Negative	FN = 13	

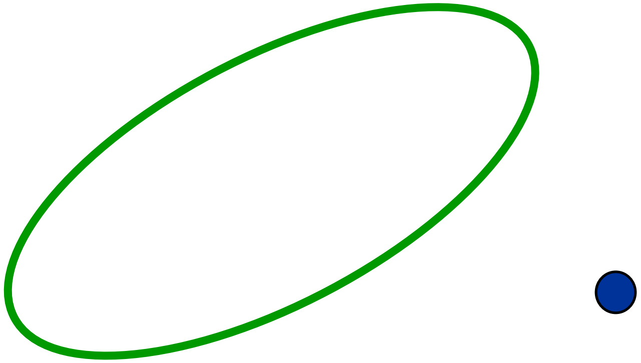
  - Precision: 70%, recall = 35%

A Different View

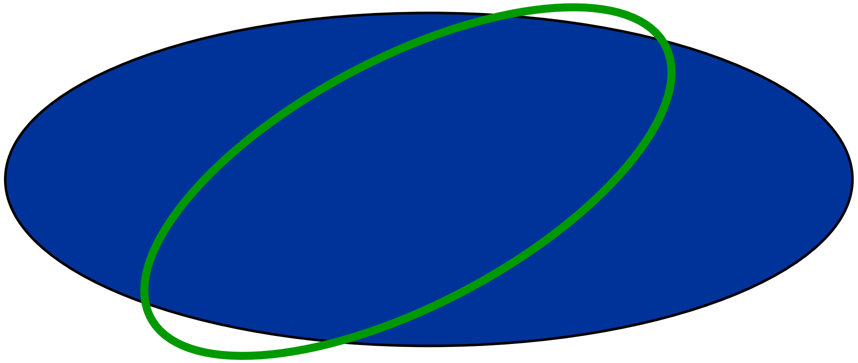
Quelle: A. Nürnberger, VL IR



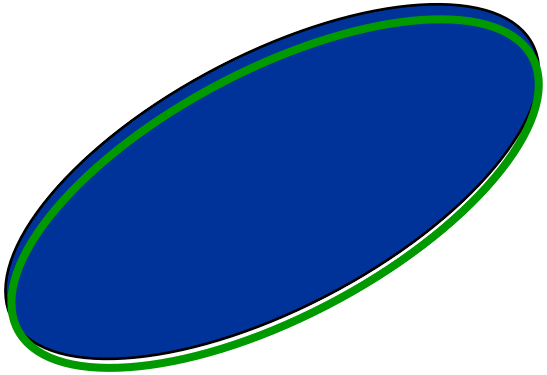
Excellent precision,
terrible recall



Terrible precision,
terrible recall



Terrible precision,
excellent recall



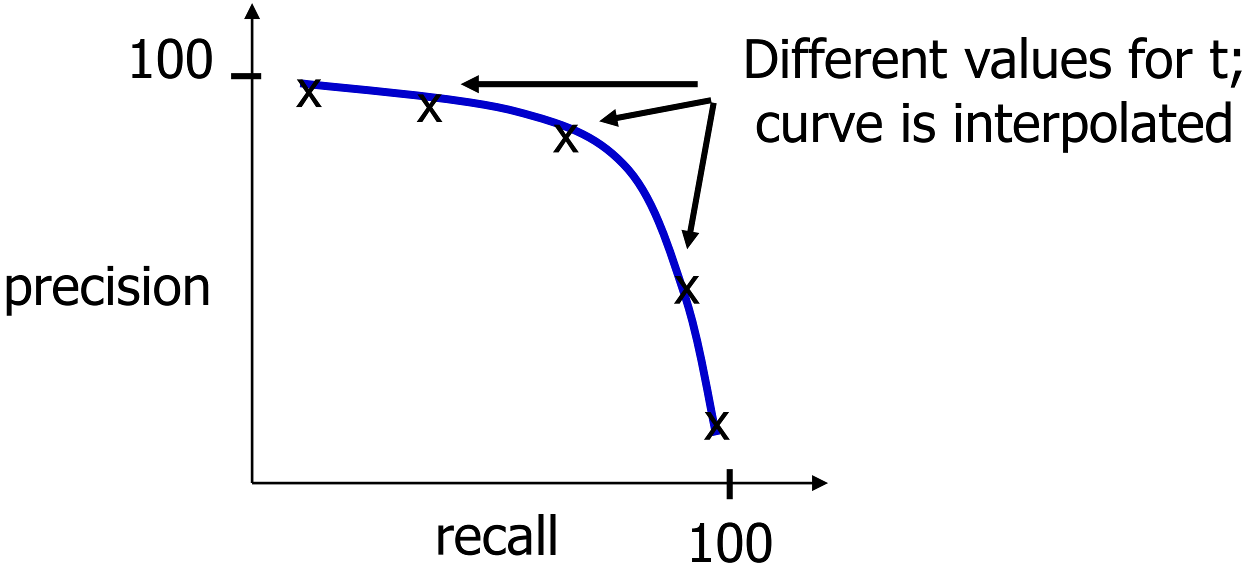
Excellent precision,
Excellent recall

Trade-off

- Inherent Trade-off between precision and recall
- Example
  - Think VSM with a threshold t to enforce a binary decision
  - Assume that docs with high relevance score are most likely also relevant for the user
  - Increase t: Less results, most of them very likely relevant
    Precision increases, recall drops
    Set t=1: P~100%, R=?
  - Decrease t: More results, some might be wrong
    Precision drops, recall increases
    Set t=0: P=?, R=100%

Precision / Recall Curve

- Sliding the threshold t gives a curve
  - Similar to Receiver-Operating-Characteristic (ROC)



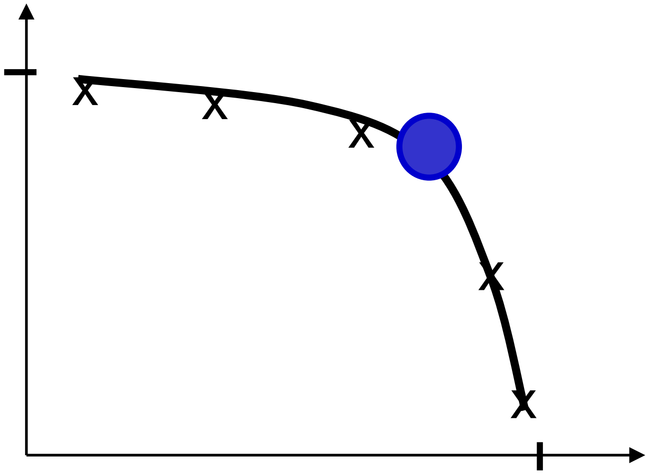
Accuracy

	Truth: relevant	**Truth: not relevant**
IR: relevant	TP	FP
IR: not relevant	FN	TN

- Sometimes, one wants one measure instead of two
    - E.g. to rank different IR-systems
- Accuracy= (TP+TN) / (TP+FP+FN+TN)
    - Which percentage of the systems decision were correct?
    - Makes only sense with small corpora and large result set
    - Typically in IR, TN >>> TP+FP+FN
    - Thus, accuracy is always good (~0,9999995)
- Used in problems with balanced sets of TN / TP

F-Measure

- F-Measure = 2*P*R / (P+R)
  - F-Measure is harmonic mean between precision and recall
  - Favors balanced P/R values



- Alternative: Area-under-the-curve (AUC)
  - AUC doesn’t help in finding the best t

From user/query to users/queries

- What if we have many queries?
  - Evaluation should always use a range of different queries
  - Compute average P/R values over all queries
  - Of course, mean and stddev are also important
- What if we have different users?
  - Different users may have different thoughts about what is relevant
  - Leads to different gold standards
  - Compute inter-annotator agreement as upper bound
- Who can judge millions of docs?
  - Evaluate on small gold standard corpus
    - But: Extrapolation is difficult: Are the properties of application/corpus really equal to properties of GS?
  - Use implicit feedback, e.g. click-through rates in top-K results

Self Assessment

- Give a definition of „Information retrieval"
- How is information retrieval different from database query evaluation?
- What are possible types of answers to a IR query?
- List 5 important steps in document preprocessing and their expected impact on precision and recall
- Give a definition of recall, precision, and accuracy
- What is the difference between micro and macro average
- Which preprocessing steps would be affected if you work with a multi language corpus?