

Natural Language Processing
SoSe 2015



Information Retrieval

Dr. Mariana Neves

June 22nd, 2015

(based on the slides of Dr. Saeedeh Momtazi)

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Information Retrieval

- The most popular usages of computer (and Internet)
 - Search, Communication
- Information Retrieval (IR)
 - The field of computer science that is mostly involved with R&D for search

A search input field with a light grey border. The text "Search Google or type URL" is displayed in a light grey font inside the field. On the right side of the field, there is a small microphone icon.

Information Retrieval

- Primary focus of IR is on text and documents
 - Mostly textual content
 - A bit structured data
 - Papers: title, author, date, publisher
 - Email: subject, sender, receiver, date

IR Dimensions

- Web search
 - Search engines



IR Dimensions

- Vertical search (Restricted domain/topic)
 - Books, movies, suppliers



IR Dimensions

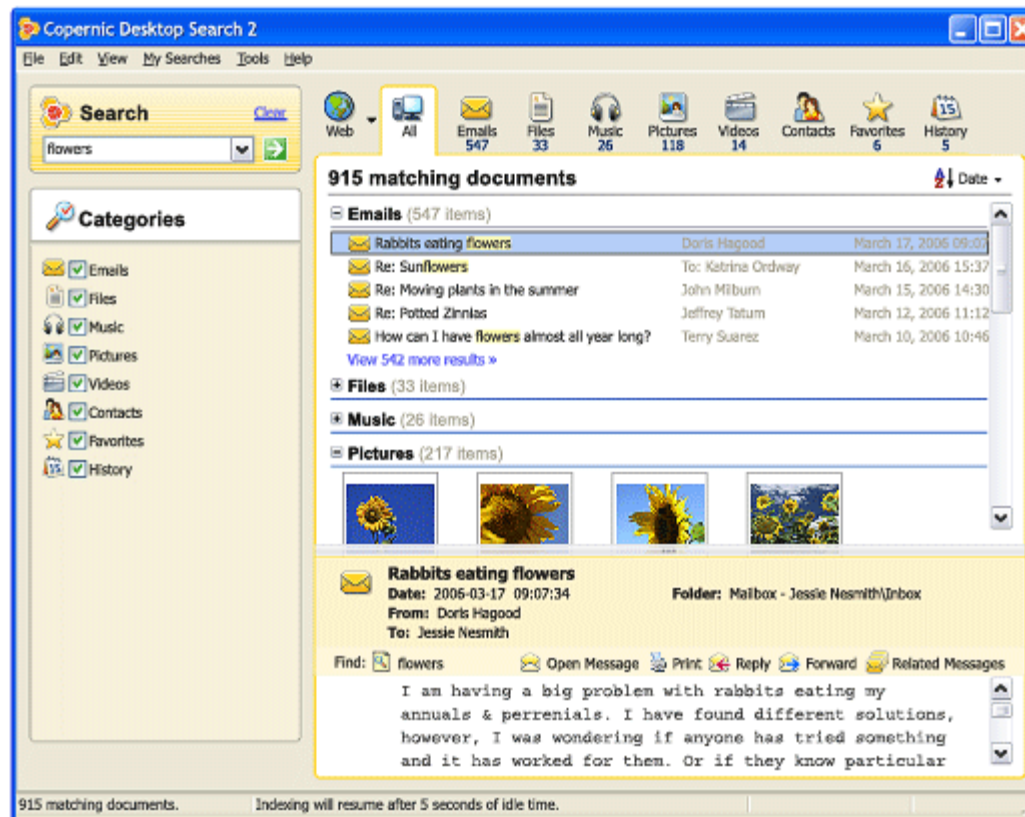
- Enterprise search, Corporate intranet
 - Emails, web pages, documentations, codes, wikis, tags, directories, presentations, spreadsheets



(<http://www.n-axis.in/practices-moss.php>)

IR Dimensions

- Desktop search
 - Personal enterprise search



(<http://www.heise.de/download/copernic-desktop-search-120141912.html>)

IR Dimensions

- P2P search (no centralized control)
 - File sharing, shared locality
- Forum search

IR Architecture

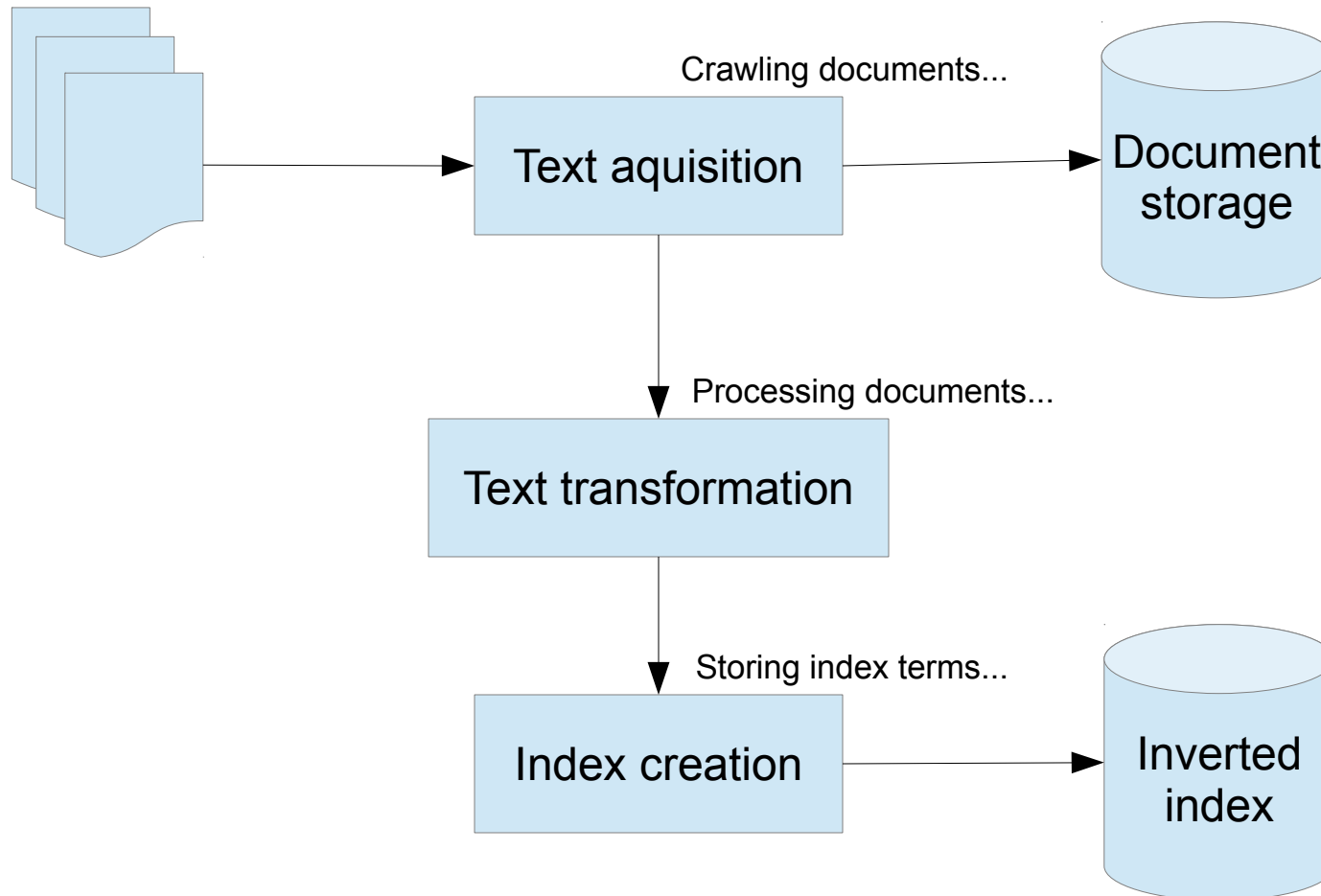
- Indexing
 - Text Acquisition
 - Text Transformation
 - Index Creation

- Querying
 - User Interaction
 - Ranking

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Indexing Architecture



Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Web Crawler

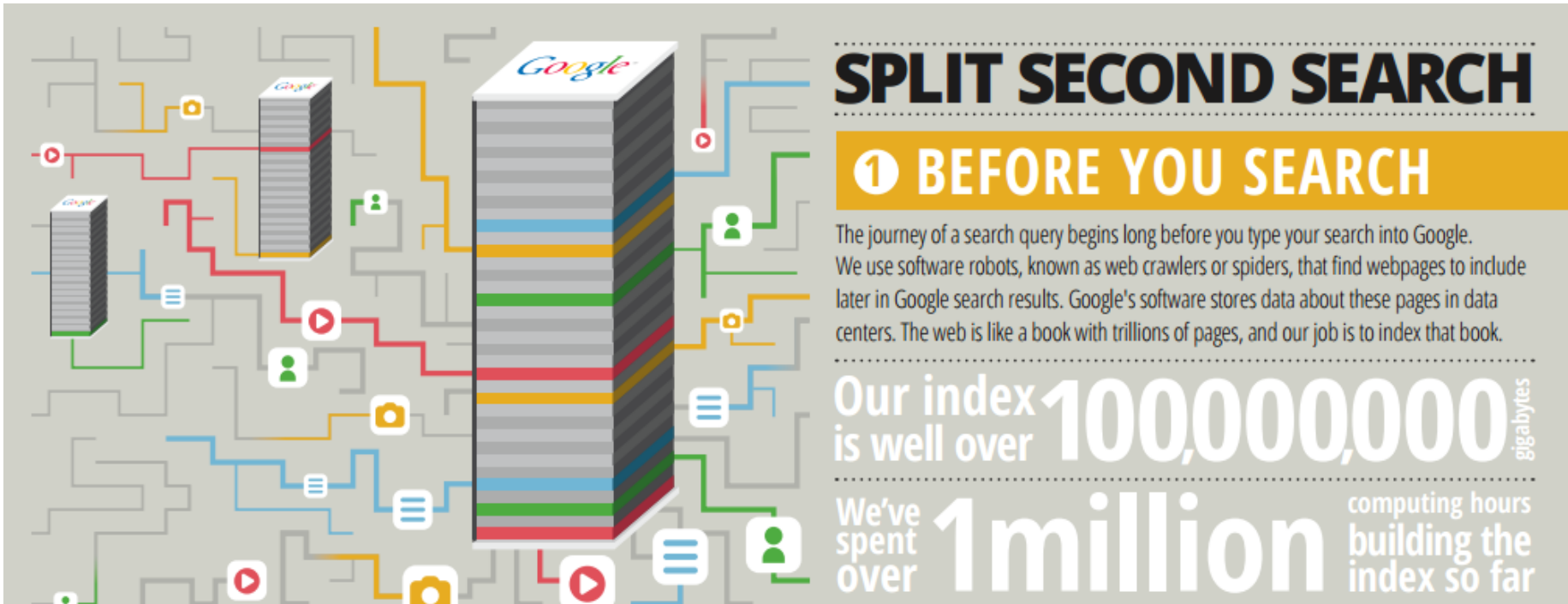
- Identifying and acquiring documents for the search engine
- Following links to find documents
 - Finding huge numbers of web pages efficiently (coverage)
 - Keeping the crawled data up-to-date (freshness)

Googlebot



bingbot

Googlebot

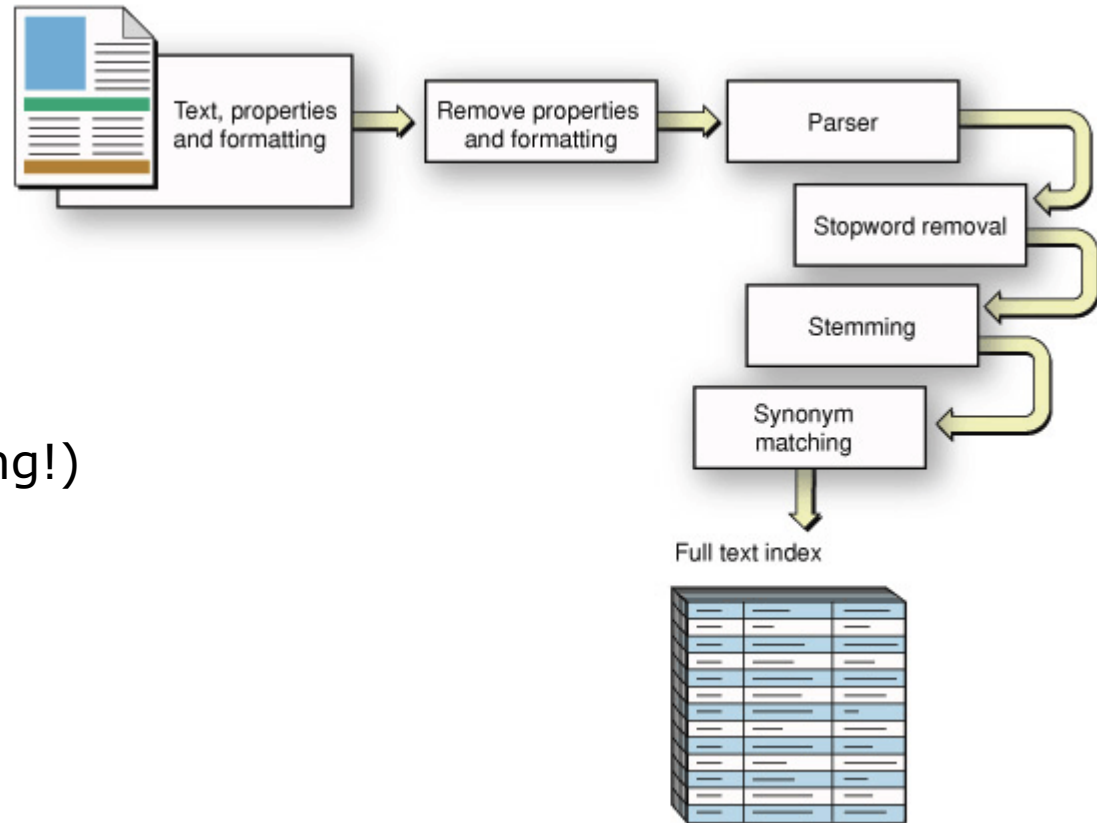


(<http://static.googleusercontent.com/media/www.google.com/en//insidesearch/howsearchworks/assets/searchInfographic.pdf>)

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Text Processing



- Document parsing
(not syntactic parsing!)
- Tokenizing
- Stopword filtering
- Stemming

Parsing

- Recognizing structural elements
 - Titles
 - Links
 - Headings
 - ...

- Using the syntax of markup languages to identify structures



(<http://htmlparser.sourceforge.net/>)

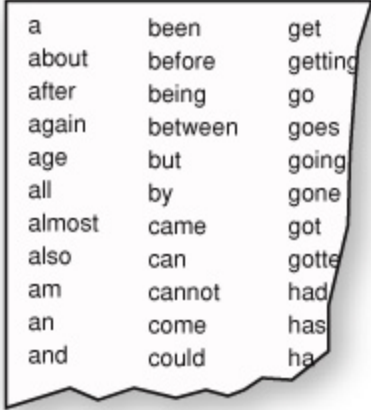
Tokenization

- Basic model
 - Considering white-space as delimiter
- Main issues that should be considered and normalized
 - Capitalization
 - apple vs. Apple
 - Apostrophes
 - O'Conner vs. owner's
 - Hyphens
 - Non-alpha characters
 - Word segmentation (e.g., in Chinese or German)

Stopwords filtering

- Removing the stop words from documents
- Stopwords: the most common words in a language
 - and, or, the, in
 - Around 400 stop words for English

Stopword list



a	been	get
about	before	getting
after	being	go
again	between	goes
age	but	going
all	by	gone
almost	came	got
also	can	gotte
am	cannot	had
an	come	has
and	could	ha

Stopwords filtering

- Advantages
 - Effective
 - Do not consider as important query terms to be searched
 - Efficient
 - Reduce storage size and time complexity
- Disadvantages
 - Problem with queries with higher impact of stop words
 - „To be or not to be“

Stemming

- Grouping words derived from a common stem
 - computer, computers, computing, compute
 - fish, fishing, fisherman



(<http://amysorr.cmswiki.wikispaces.net/Stem%20Contract%20&%20Word%20Lists>)

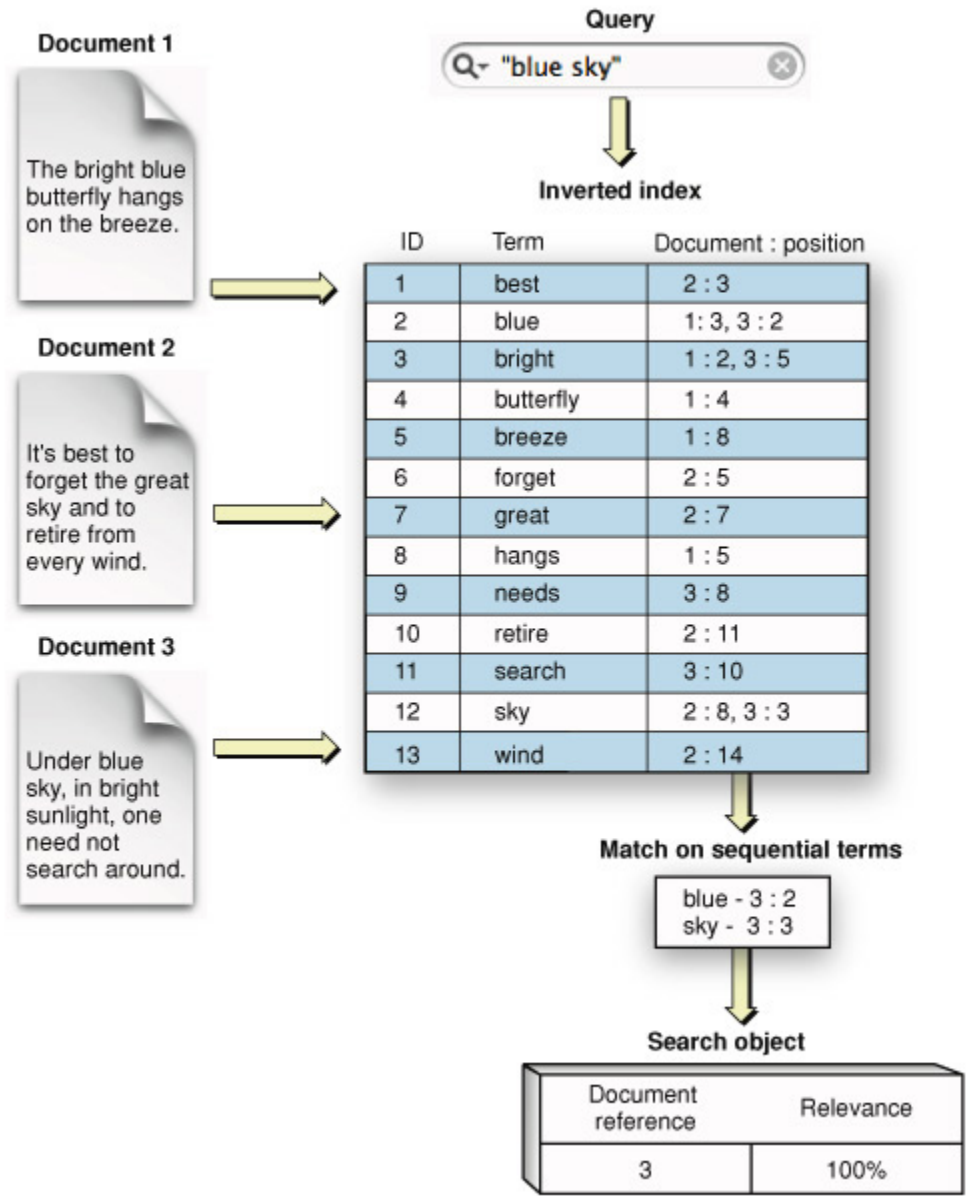
Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Index Storing

- Storing document-words information in an inverse format
 - Converting document-term statistics to term-document for indexing
- Increasing the efficiency of the retrieval engine

Index Storing



(https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/SearchKitConcepts/searchKit_basics/searchKit_basics.html)

Index Storing

- Index-term with document ID

environments	1
fish	1 2 3 4
fishkeepers	2
found	1
fresh	2
freshwater	1 4
from	4

Index Storing

- Index-term with document ID and frequency

environments	1:1			
fish	1:2	2:3	3:2	4:2
fishkeepers	2:1			
found	1:1			
fresh	2:1			
freshwater	1:1	4:1		
from	4:1			

Index Storing

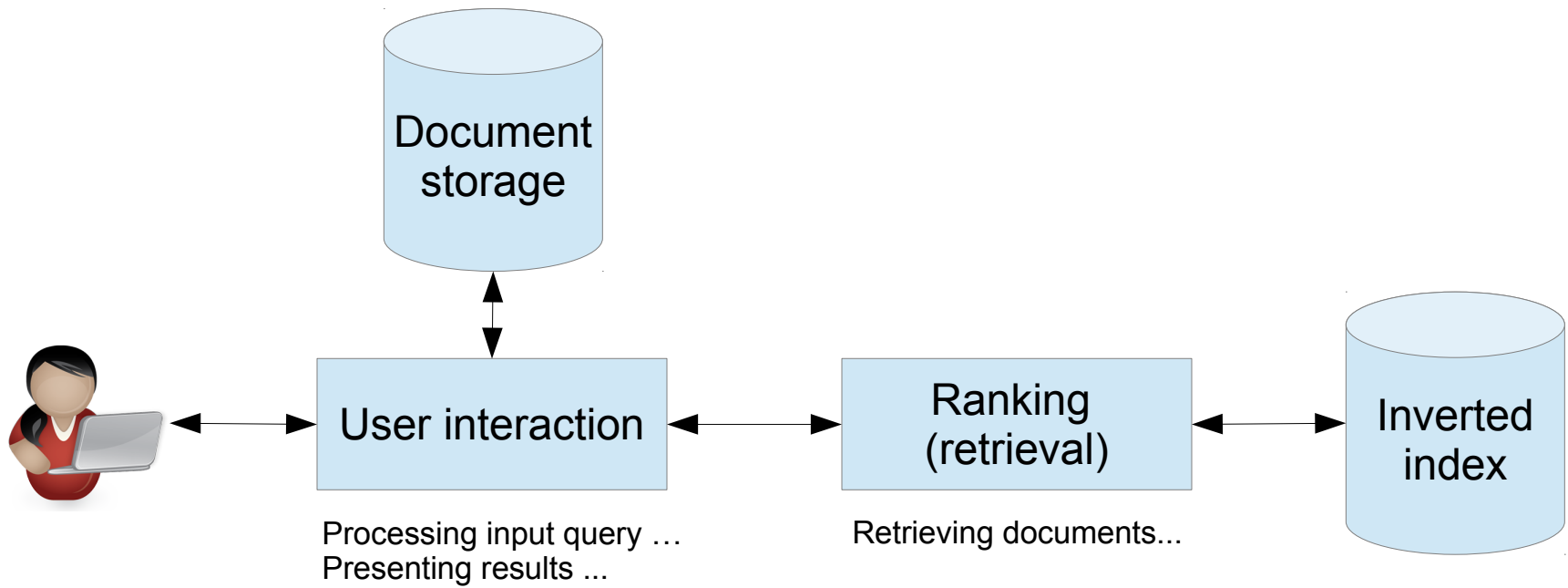
- Index-term with document ID and position

environments	1,8								
fish	1,2	1,4	2,7	2,18	2,23	3,2	3,6	4,3	4,13
fishkeepers	2,1								
found	1,5								
fresh	2,13								
freshwater	1,14	4,2							
from	4,8								

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Querying Architecture



Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- Evaluation

Query Processing

- Applying similar techniques used in text processing for documents
 - Tokenization, stopwords filtering, stemming

“the best book for natural language processing”



“best book natur languag process”

Query Processing

- Spell checking
 - Correcting the query in case of spelling errors

“the best bok for natural language procesing”



“the best book for natural language processing”

Query Processing

- Query suggestion
 - Providing alternatives words to the original query (based on query logs)

“the best book for natural language processing”



“the best book for NLP”

Query Processing

- Query expansion and relevance feedback
 - Modifying the original query with additional terms

“the best book for natural language processing”



“the best [book|volume] for [natural language processing|NLP]”

Query Expansion

- Short search queries are under-specified
 - 26.45% (1 word)
 - 23.66% (2 words)
 - 19.34 (3 words)
 - 13.17% (4 words)
 - 7.69% (5 words)
 - 4.12% (6 words)
 - 2.26% (7 words)
- Google:
 - „54.5% of user queries are greater than 3 words“

(<http://www.tamingthebeast.net/blog/web-marketing/search-query-lengths.htm>)
(<http://www.smallbusinesssem.com/google-query-length/3273/>)

Query Expansion

- car inventor

Nicolas-Joseph Cugnot is widely credited with building the first full-scale, self-propelled mechanical vehicle or automobile in about 1769; he created a steam-powered tricycle.^[22]

Car - Wikipedia, the free encyclopedia

<https://en.wikipedia.org/?title=Car> ▾

Axes, 1–2. Inventor, Ferdinand Verbiest ... A car is a wheeled, self-powered motor vehicle used for transportation. ... Karl Benz, the inventor of the modern car.

History of the automobile - Car (disambiguation) - Motor vehicle - Ferdinand Verbiest

Who invented the automobile? (Everyday Mysteries: Fun ...

www.loc.gov ▸ Researchers ▾

Exactly who invented the automobile is a matter of opinion. If we had to give credit to one inventor, it would probably be Karl Benz from Germany. Many suggest ...

Who invented the world's very first car? - io9

io9.com/5816040/who-invented-the-worlds-very-first-car ▾

Jun 28, 2011 - The very first car might well have been the invention of a Flemish missionary named Ferdinand Verbiest. Born in Flanders in 1623, Verbiest ...

Who Invented the Car? - LiveScience

www.livescience.com/37538-who-invented-the-car.html ▾

Jun 18, 2013 - Karl Benz patented the three-wheeled Motor Car in 1886. It was the first true, ... Karl Benz, inventor of the first practical, modern automobile.

Google search

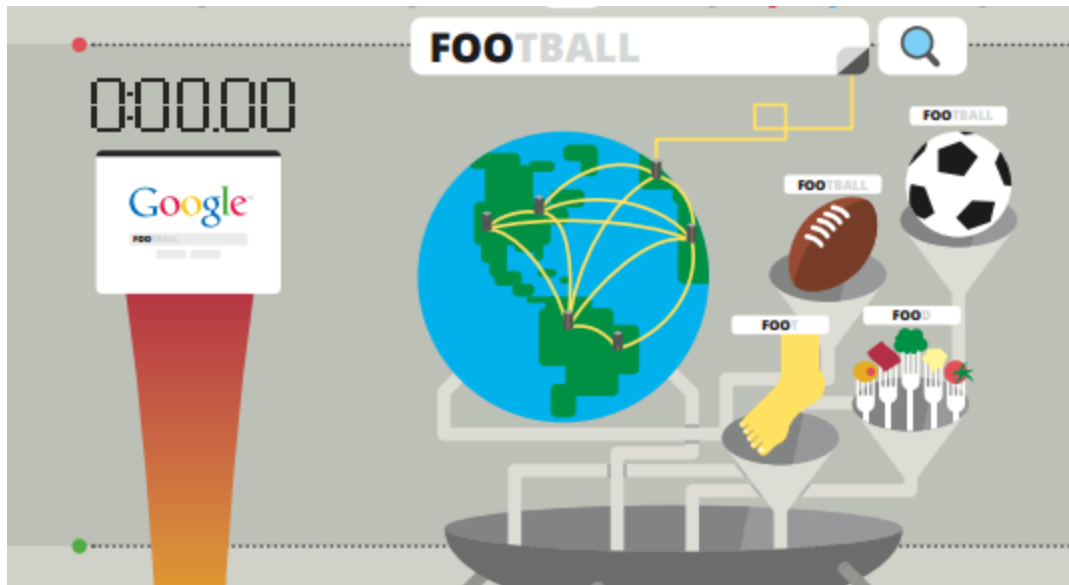
2 AS YOU SEARCH

When you start your search, that's when Google's algorithm begins to find the information you're looking for.

The search query travels on average **1,500** miles to get the answer back to you (and may hit different data centers around the world along the way), at a speed that's close to the speed of light, hundreds of millions of miles per hour.

As you type your query, you'll start seeing predictions of searches you might be looking for and results showing up, without you having to hit enter. It saves you time and gets you to your answer as quickly as possible.

This is what we call Google Instant



(<http://static.googleusercontent.com/media/www.google.com/en//insidesearch/howsearchworks/assets/searchInfographic.pdf>)

Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- **Querying Block**
 - Query Processing
 - **Retrieval Models**
 - Result Representation
- Evaluation

Retrieval Models

- Calculating the scores for documents using a ranking algorithm
 - Boolean model
 - Vector space model
 - Probabilistic model
 - Language model

Boolean Model

- Two possible outcomes for query processing
 - TRUE or FALSE
 - All matching documents are considered equally relevant
- Query usually specified using Boolean operators
 - AND, OR, NOT

Boolean Model

- Search for news articles about President Lincoln
 - lincoln
 - cars
 - places
 - people

Luxury Cars, Crossovers SUVs | The Lincoln Motor ...

www.lincoln.com/ ▾

The Lincoln Motor Company Luxury Cars, Crossovers SUVs. The official website of The Lincoln Motor Company luxury vehicles.

[2015 Lincoln MKC](#) - [2015 Lincoln Navigator](#) - [2015 Lincoln MKZ](#) - [2015 Lincoln MKX](#)

Abraham Lincoln - Wikipedia, the free encyclopedia

https://en.wikipedia.org/?title=Abraham_Lincoln ▾

Born in Hodgenville, Kentucky, Lincoln grew up on the western frontier in Kentucky and Indiana. Largely self-educated, he became a lawyer in Illinois, a Whig ...

Lincoln (2012 film) - Wikipedia, the free encyclopedia

[https://en.wikipedia.org/wiki/Lincoln_\(2012_film\)](https://en.wikipedia.org/wiki/Lincoln_(2012_film)) ▾

Lincoln is a 2012 American epic historical drama film directed by Steven Spielberg, ... Lincoln premiered on October 8, 2012 at the New York Film Festival.

Lincoln (2012) - IMDb

www.imdb.com/title/tt0443272/ ▾

★★★★★ Rating: 7.4/10 - 178,527 votes

Directed by Steven Spielberg. With Daniel Day-Lewis, Sally Field, David Strathairn, Joseph Gordon-Levitt. As the Civil War continues to rage, America's ...

Boolean Model

- Search for news articles about President Lincoln
 - president AND lincoln
 - “Ford Motor Company today announced that Darryl Hazel will succeed Brian Kelley as president of Lincoln Mercury ”

Boolean Model

- Search for news articles about President Lincoln
 - president AND lincoln AND NOT (automobile OR car)
 - “President Lincoln’s body departs Washington in a nine-car funeral train.”

Boolean Model

- Search for news articles about President Lincoln
 - president AND lincoln AND (biography OR life OR birthplace OR gettysburg) AND NOT (automobile OR car)
 - “President’s Day - Holiday activities - crafts, mazes, mazes word searches, ... ‘The Life of Washington’ Read the entire searches The Washington book online! Abraham Lincoln Research Site ...”

Boolean Model

- Advantages
 - Results are predictable and relatively easy to explain
- Disadvantages
 - Relevant documents have no order
 - Complex queries are difficult to write

Vector Space Model

- Very popular model, even today
- Documents and query represented by a vector of term weights
 - t is number of index terms (i.e., very large)
 - $D_i = (d_{i1}, d_{i2}, \dots, d_{it})$
 - $Q = (q_1, q_2, \dots, q_t)$

Vector Space Model

- Document collection represented by a matrix of term weights

	<i>Term₁</i>	<i>Term₂</i>	...	<i>Term_t</i>
<i>Doc₁</i>	d_{11}	d_{12}	...	d_{1t}
<i>Doc₂</i>	d_{21}	d_{22}	...	d_{2t}
...
<i>Doc_n</i>	d_{n1}	d_{n2}	...	d_{nt}

Vector Space Model

- Ranking each document by distance between query and document
- Using cosine similarity (normalized dot product)
 - Cosine of angle between document and query vectors

$$\text{Cosine}(D_i, Q) = \frac{\sum_{j=1}^t d_{ij} \times q_j}{\sqrt{\sum_j d_{ij}^2} \sqrt{\sum_j q_j^2}}$$

- Having no explicit definition of relevance as a retrieval model
- Implicit: Closer documents are more relevant.

Vector Space Model

- Term frequency weight (tf)
 - Measuring the importance of term „k“ in document „i“
- Inverse document frequency (idf)
 - Measuring importance of the term „k“ in collection

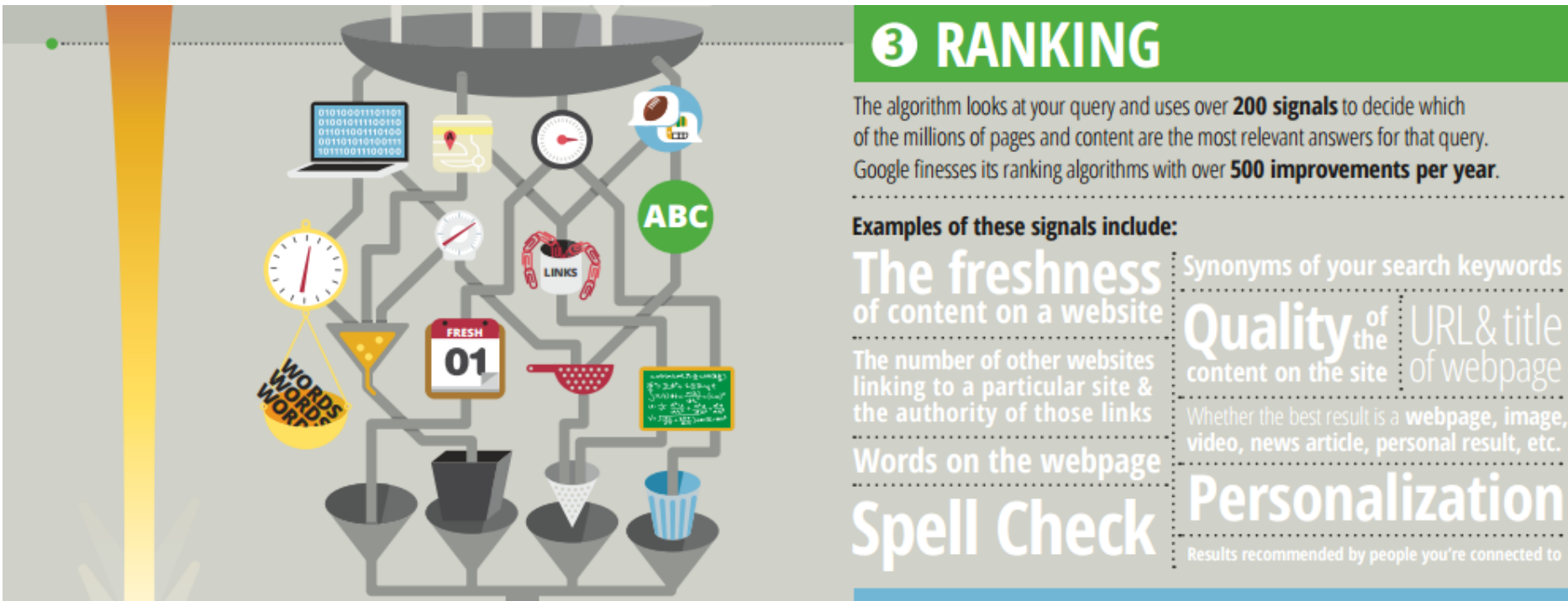
$$idf_k = \log \frac{N}{n_k}$$

- Final weighting: multiplying tf and idf, called tf.idf

Vector Space Model

- Advantages
 - Simple computational framework for ranking
 - Any similarity measure or term weighting scheme can be used
- Disadvantages
 - Assumption of term independence

Google ranking



Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- **Querying Block**
 - Query Processing
 - Retrieval Models
 - **Result Representation**
- Evaluation

Results

- Constructing the display of ranked documents for a query
- Generating snippets to show how queries match documents
- Highlighting important words and passages
- Providing clustering (if required)

Car - Wikipedia, the free encyclopedia

<https://en.wikipedia.org/?title=Car> ▾

Axles, 1–2. **Inventor, Ferdinand Verbiest** ... A **car** is a wheeled, self-powered motor vehicle used for transportation. **Karl Benz**, the **inventor** of the modern **car**.

[History of the automobile - Car \(disambiguation\) - Motor vehicle - Ferdinand Verbiest](#)

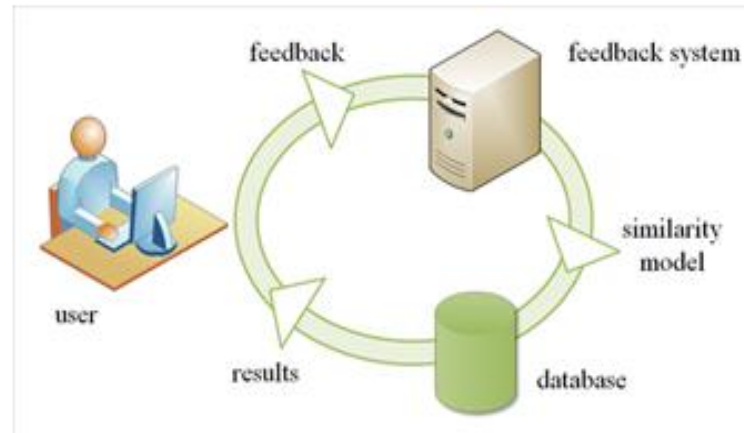
Who invented the automobile? (Everyday Mysteries: Fun ...

www.loc.gov ▸ [Researchers](#) ▾

Exactly who invented the **automobile** is a matter of opinion. If we had to give credit to one **inventor**, it would probably be **Karl Benz** from Germany. Many suggest ...

Relevance feedback

- Good results even after only one iteration



(<http://dme.rwth-aachen.de/en/research/projects/relevance-feedback>)

Relevance feedback

- Push the vector
 - Towards the relevant documents
 - Away from the irrelevant documents

$$\vec{q}_{i+1} = \vec{q}_i + \frac{\beta}{R} \sum_{j=1}^R \vec{r}_j - \frac{\gamma}{S} \sum_{k=1}^S \vec{s}_k$$

R = number of relevant documents
S = number of irrelevant document
 $\beta + \gamma = 1$; both defined experimentally

$\beta = 0.75$; $\gamma = 0.25$
(Salton and Buckley 1990)

Relevance feedback

- Evaluation only on the residual collection
 - Documents not shown to the user

Google result



Outline

- Introduction
- Indexing Block
 - Document Crawling
 - Text Processing
 - Index Storing
- Querying Block
 - Query Processing
 - Retrieval Models
 - Result Representation
- **Evaluation**

Evaluation Metrics

- Evaluation of unranked sets
 - Precision
 - Recall
 - F-measure
 - Accuracy
- Two categories: relevant or not relevant to the query

Evaluation Metrics

- Total of „T“ documents in response to a query
 - „R“ relevant documents
 - „N“ irrelevant documents
- Total „U“ relevant documents in the collection

$$Precision = \frac{R}{T}$$

$$Recall = \frac{R}{U}$$

Homonymy, Polysemy, Synonymy

- Homonymy and polysemy reduce the precision
 - Documents of the „wrong“ sense will be judged as irrelevant

Canine - Wikipedia, the free encyclopedia

<https://en.wikipedia.org/wiki/Canine> ▾

Canine may refer to: ... Canine tooth, in mammalian oral anatomy. Other uses[edit].
 Ralph Canine (1895–1969), first director of the United States National ...

Canine tooth - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Canine_tooth ▾

In mammalian oral anatomy, the canine teeth, also called cuspids, dog teeth, fangs, or (in the case of those of the upper jaw) eye teeth, are relatively long, ...

Canine | Define Canine at Dictionary.com

[dictionary.reference.com/browse/canine](https://www.dictionary.reference.com/browse/canine) ▾

of or like a dog; relating to or characteristic of dogs: canine loyalty. 2. Anatomy, Zoology. of or relating to the four pointed teeth, especially prominent in dogs, ...

Homonymy, Polysemy, Synonymy

- Synonymy and hyponymy reduce the recall
 - Miss relevant documents that do not contain the keyword

[Dog Health Center | Dog Care and Information from WebMD](https://pets.webmd.com/dogs/)
pets.webmd.com/dogs/ ▾

Welcome to the new WebMD Dog Health Center. WebMD veterinary experts provide comprehensive information about **dog** health care, offer nutrition and ...

[Dog Breed Info Center - List of All Dog Breeds by Type ...](https://dogtime.com/dog-breeds/)
dogtime.com/dog-breeds ▾

Dog breed profiles of more than 200 breeds. Includes personality, history, **dog** pictures, **dog** health info, and more. Find the **dog** breed that is right for you.

[DOG](https://www.dog.org/)
www.dog.org/ ▾ [Translate this page](#)

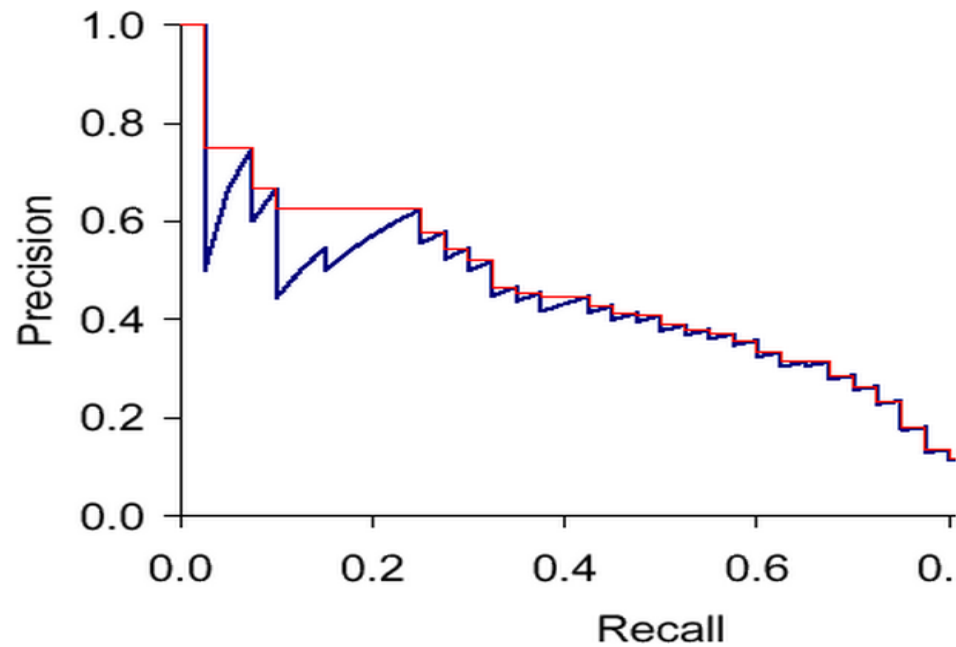
Publikationen, Termine und Patienteninformationen auf den Seiten der 1857 gegründeten Gesellschaft.

Evaluation Metrics

- Evaluation of ranked sets
 - Precision-recall curve
 - Mean average precision
 - Precision at n
 - Mean reciprocal rank

Evaluation Metrics

- Precision-recall curve
 - Based on the rank of the results



(<http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html>)

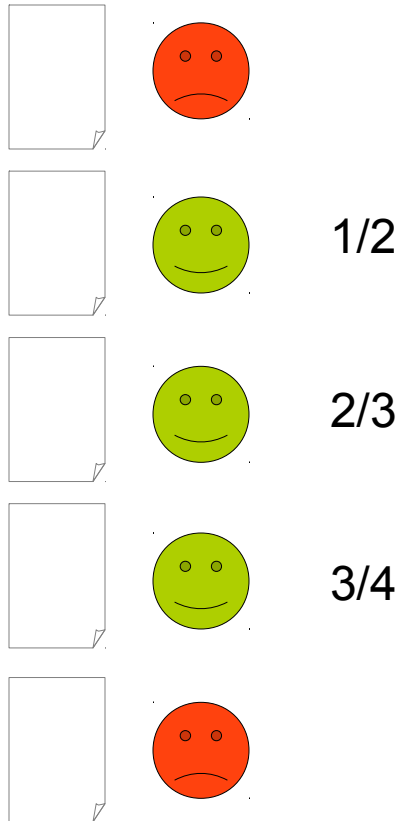
Average Precision

- Average precision
 - Calculating precision of the system after retrieving each relevant document
 - Averaging these precision values

$$\textit{AveragePrecision} = \frac{\sum_{k=1}^K P@k}{\textit{number relevant documents}}$$

(k is the rank of relevant documents in the retrieved list)

Average Precision



$$AP = \frac{\frac{1}{2} + \frac{2}{3} + \frac{3}{4}}{3} = 0.64$$

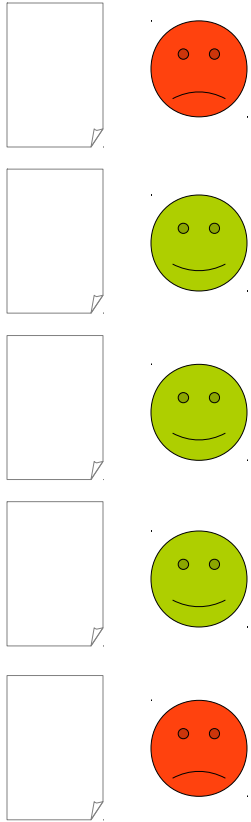
Mean average Precision

- Mean average precision
 - Reporting the mean of the average precisions over all queries in the query set

Precision at n

- Evaluate results for the first n items of the retrieved documents
- Calculating precision at n
 - Considering the top n documents only
 - Ignoring the rest of documents

Precision at 5



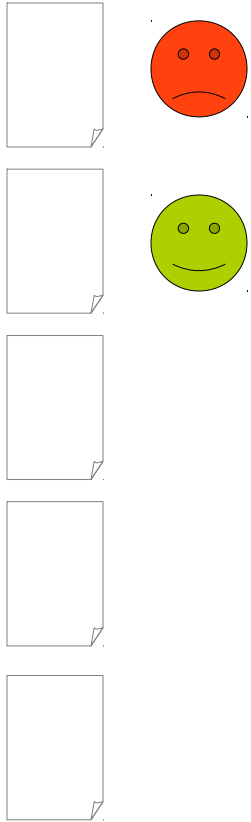
$$P @ 5 = \frac{3}{5} = 0.6$$

Reciprocal Rank

- Evaluate only one correct answer
- Reciprocal Rank
 - Inversing the score of the rank at which the first correct answer is returned

$$\textit{ReciprocalRank} = \frac{1}{R} \quad (\text{R is the position of the first correct item in the ranked list})$$

Reciprocal Rank



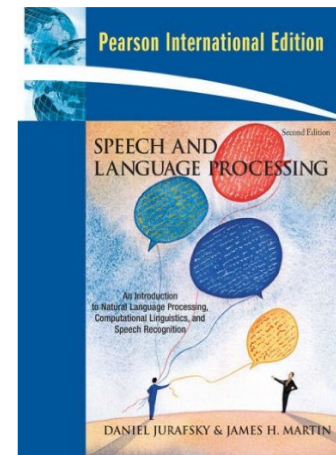
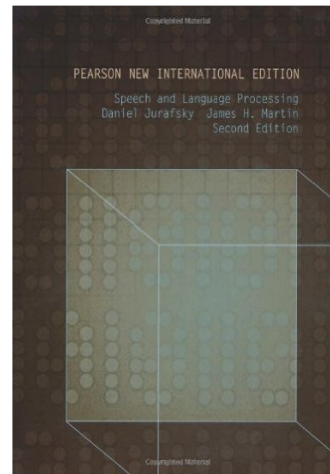
$$\textit{ReciprocalRank} = \frac{1}{2} = 0.5$$

Reciprocal Rank

- Mean Reciprocal Rank
 - Reporting the mean of the reciprocal rank over all queries in the query set

Further Reading

- Speech and Language Processing
 - Chapter 23.1



Further Reading

