DAT630

Search Engines

Search Engine Architecture and Indexing

22/09/2017

Krisztian Balog | University of Stavanger

Information Retrieval

Information Retrieval (IR)

> "Information retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information."
>
> (Salton, 1968)

Modern definition

> "Making the right information available to the right person at the right time."

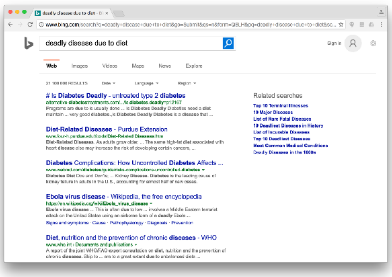
Searching in Databases

- Query: records with balance > $50,000 in branches located in Amherst, MA.

Name	Branch	Balance
Sam I. Am	Amherst, MA	$95,342.11
Patty MacPatty	Amherst, MA	$23,023.23
Bobby de West	Amherst, NY	$78,000.00
Xing O'Boston	Boston, MA	$50,000.01

Searching in Text

- Query: *deadly disease due to diet*
- Which are relevant?


deadly disease due to diet - B
www.bing.com/search?q=deadly+disease+due+to+diet&go=Submit&qs=n&form=QBLH&pq=deadly+disease+due+to+diet&sc...
deadly disease due to diet
Sign in
Web
Images
Videos
Maps
News
Explore
21 100 000 RESULTS
Date
Language
Region

# Is Diabetes Deadly - untreated type 2 diabetes
alternative-diabetestreatments.com/.../Is.diabetes.deadly=p12167
Programs are due to is usually done ... Is Diabetes Deadly Diabetics need a diet maintain ... very good diabetes.,Is Diabetes Deadly Diabetes is a disease that ...

Diet-Related Diseases - Purdue Extension
www.four-h.purdue.edu/foods/Diet-Related Diseases.htm
Diet-Related Diseases. As adults grow older, ... The same high-fat diet associated with heart disease also may increase the risk of developing certain cancers, ...

Diabetes Complications: How Uncontrolled Diabetes Affects ...
www.webmd.com/diabetes/guide/risks-complications-uncontrolled-diabetes
Diabetes Diet Dos and Don'ts; ... Kidney Disease. Diabetes is the leading cause of kidney failure in adults in the U.S., accounting for almost half of new cases.

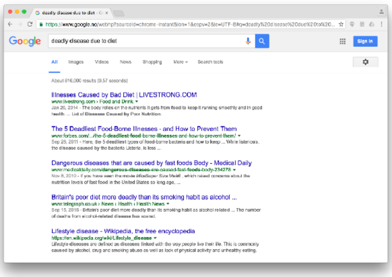
Ebola virus disease - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Ebola_virus_disease
Ebola virus disease ... This is often due to low ... involves a Middle Eastern terrorist attack on the United States using an airborne form of a deadly Ebola ...
Signs and symptoms · Cause · Pathophysiology · Diagnosis · Prevention

Diet, nutrition and the prevention of chronic diseases - WHO
www.who.int › Documents and publications
A report of the joint WHO/FAO expert consultation on diet, nutrition and the prevention of chronic diseases. Skip to ... are to a great extent due to unbalanced diets ...

Related searches
- Top 10 Terminal Illnesses
- 10 Major Diseases
- List of Rare Fatal Diseases
- 10 Deadliest Diseases in History
- List of Incurable Diseases
- Top 10 Deadliest Diseases
- Most Common Medical Conditions
- Deadly Diseases in the 1800s

Searching in Text

- Query: *deadly disease due to diet*
- Which are relevant?


deadly disease due to diet - G
https://www.google.no/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=deadly%20disease%20due%20to%20...
Google
deadly disease due to diet
Sign in
All
Images
Videos
News
Shopping
More
Search tools

About 816,000 results (0.57 seconds)

Illnesses Caused by Bad Diet | LIVESTRONG.COM
www.livestrong.com › Food and Drink
Jan 26, 2014 - The body relies on the nutrients it gets from food to keep it running smoothly and in good health. ... List of Diseases Caused by Poor Nutrition.

The 5 Deadliest Food-Borne Illnesses - and How to Prevent Them
www.forbes.com/.../the-5-deadliest-food-borne-illnesses-and-how-to-prevent-them/
Sep 28, 2011 - Here, the 5 deadliest types of food-borne bacteria and how to keep ... While listeriosis, the disease caused by the bacteria Listeria, is less ...

Dangerous diseases that are caused by fast foods Body - Medical Daily
www.medicaldaily.com/dangerous-diseases-are-caused-fast-foods-body-234278
Nov 8, 2010 - If you have seen the movie âSuper Size Meâ, which raised concerns about the nutrition levels of fast food in the United States so long ago, ...

Britain's poor diet more deadly than its smoking habit as alcohol ...
www.telegraph.co.uk › News › Health › Health News
Sep 15, 2016 - Britain's poor diet more deadly than its smoking habit as alcohol related ... The number of deaths from alcohol-related disease has soared.

Lifestyle disease - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Lifestyle_disease
Lifestyle diseases are defined as diseases linked with the way people live their life. This is commonly caused by alcohol, drug and smoking abuse as well as lack of physical activity and unhealthy eating.

Comparing Text

- Comparing the query text to the document text and determining what is a good match is the core issue of information retrieval
- Exact matching of words is not enough
  - Many different ways to write the same thing in a "natural language" like English
  - E.g., does a news story containing the text "fatal illnesses caused by your menu" match the query?
  - Some documents will be better matches than others

Dimensions of IR

- IR is more than just text, and more than just web search
  - Although these are central
- **Content**
  - Text
  - Images
  - Video
  - Audio
  - Scanned documents

Dimensions of IR

- **Applications**
  - Web search
  - Vertical search
  - Enterprise search
  - Mobile search
  - Social search
  - Desktop search
  - Patent search
  - …

Dimensions of IR

- **Tasks**
  - Ad-hoc search
  - Filtering
  - Question answering

Core issues in IR

- **Relevance**
  - Simple (and simplistic) definition: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  - Many factors influence a person's decision about what is relevant: e.g., task, context, novelty
  - *Topical relevance* (same topic) vs. *user relevance* (everything else)

Core issues in IR

- **Relevance**
  - *Retrieval models* define a view of relevance
  - *Ranking algorithms* used in search engines are based on retrieval models
  - Most models based on statistical properties of text rather than linguistic
    - I.e., counting simple text features such as words instead of parsing and analyzing the sentences

Core issues in IR

- **Evaluation**
  - Experimental procedures and measures for comparing system output with user expectations
  - Typically use test collection of documents, queries, and relevance judgments
  - *Recall* and *precision* are two examples of effectiveness measures

Core issues in IR

- **Information needs**
  - Keyword queries are often poor descriptions of actual information needs
  - Interaction and context are important for understanding user intent
  - Query refinement techniques such as query expansion, query suggestion, relevance feedback improve ranking

Search Engines in Operational Environments

- Performance
  - Response time, indexing speed, etc.
- Incorporating new data
  - Coverage and freshness
- Scalability
  - Growing with data and users
- Adaptibility
  - Tuning for specific applications

Search Engine Architecture

Search Engine Architecture

- A software architecture consists of software components, the interfaces provided by those components, and the relationships between them
  - Describes a system at a particular level of abstraction
- Architecture of a search engine determined by 2 requirements
  - Effectiveness (quality of results)
  - Efficiency (response time and throughput)

Indexing Process

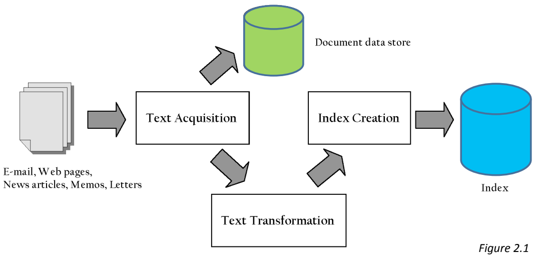
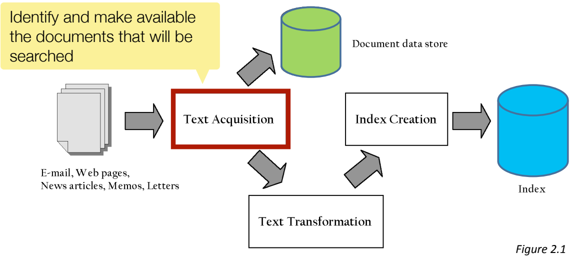


Figure 2.1

Indexing Process



Identify and make available the documents that will be searched

Figure 2.1

Text Acquisition

- **Crawler**
  - Identifies and acquires documents for search engine
  - Many types: web, enterprise, desktop, etc.
  - Web crawlers follow links to find documents
    - Must efficiently find huge numbers of web pages (*coverage*) and keep them up-to-date (*freshness*)
    - Single site crawlers for site search
    - *Topical* or *focused* crawlers for vertical search
  - Document crawlers for enterprise and desktop search
    - Follow links and scan directories

Text Acquisition

- **Feeds**
  - Real-time streams of documents
  - E.g., web feeds for news, blogs, video, radio, TV
  - RSS is common standard
  - RSS "reader" can provide new XML documents to search engine

Text Acquisition

- Documents need to be **converted** into a consistent text plus metadata format
  - E.g. HTML, XML, Word, PDF, etc. → XML
- Convert text encoding for different languages
  - Using a Unicode standard like UTF-8

Indexing Process

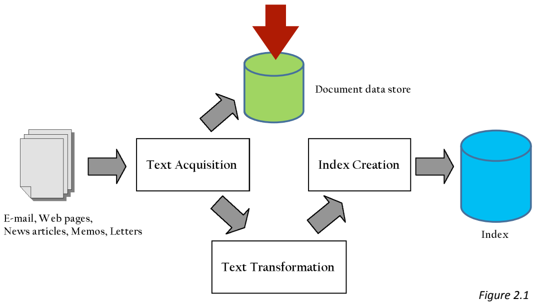


Figure 2.1

Document Data Store

- Stores text, metadata, and other related content for documents
  - Metadata is information about document such as type and creation date
  - Other content includes links, anchor text
- Provides fast access to document contents for search engine components
  - E.g. result list generation
- Could use relational database system
  - More typically, a simpler, more efficient storage system is used due to huge numbers of documents

Indexing Process

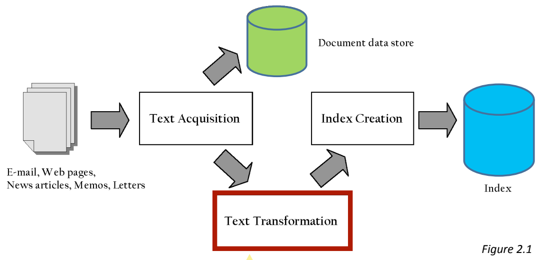


Figure 2.1

Transform documents into index terms or features

Text Transformation

- Tokenization
- Stopword removal
- Stemming
- Information extraction
  - Identify index terms that more complex than single words
    - E.g., named entity recognizers identify classes such as people, locations, companies, dates, etc
  - Important for some applications

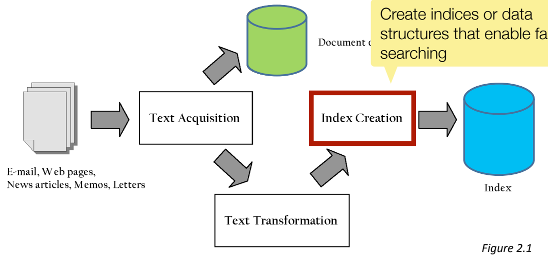
Text Transformation

- Link Analysis
  - Makes use of links and anchor text in web pages
  - Link analysis identifies popularity and community information
    - E.g., PageRank
  - Anchor text can significantly enhance the representation of pages pointed to by links
    - Significant impact on web search
    - Less importance in other applications

Text Transformation

- Classification
  - Identifies class-related metadata for documents or part of documents
    - Topics, reading levels, sentiment, genre
    - Spam vs. non-spam
    - Non-content parts of documents, e.g., advertisements
  - Use depends on application

Indexing Process



Create indices or data structures that enable fast searching

Figure 2.1

Index Creation

- Document Statistics
  - Gathers counts and positions of words and other features
  - Used in ranking algorithm
- Weighting
  - Computes weights for index terms
  - Usually reflect “importance” of term in the document
  - Used in ranking algorithm

Index Creation

- Inversion
  - Core of indexing process
  - Converts document-term information to term-document for indexing
    - Difficult for very large numbers of documents
  - Format of inverted file is designed for fast query processing
    - Must also handle updates
    - Compression used for efficiency

Index Creation

- Index Distribution
  - Distributes indexes across multiple computers and/or multiple sites
  - Essential for fast query processing with large numbers of documents
  - Many variations
    - Document distribution, term distribution, replication
  - P2P and distributed IR involve search across multiple sites

Query Process

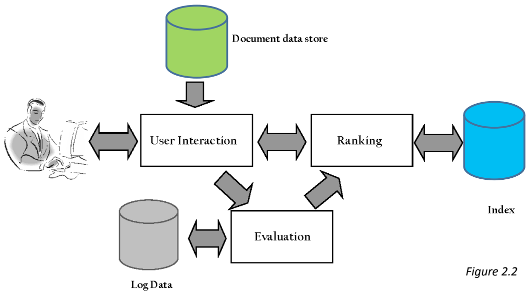


Figure 2.2

Query Process

Interface between the person doing the searching and the search engine

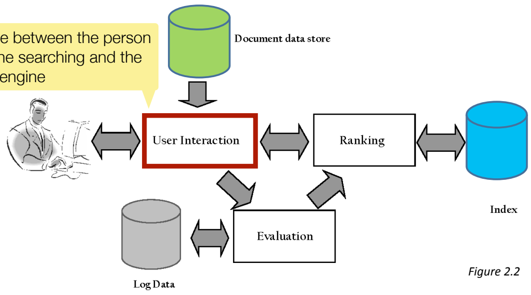


Figure 2.2

User Interaction

- Accepting the user's query and transforming it into index terms
- Taking the ranked list of documents from the search engine and organizing it into the results shown to the user
  - E.g., generating snippets to summarize documents
- Range of techniques for refining the query (so that it better represents the information need)

User Interaction

- Query input
  - Provides interface and parser for *query language*
  - Query language used to describe complex queries
    - *Operators* indicate special treatment for query text
  - Most web search query languages are very simple
    - Small number of operators
  - There are more complicated query languages
    - E.g., Boolean queries, proximity operators
    - IR query languages also allow content and structure specifications, but focus on content

User Interaction

- Query transformation
  - Improves initial query, both before and after initial search
  - Includes text transformation techniques used for documents
  - *Spell checking* and *query suggestion* provide alternatives to original query
    - Techniques often leverage query logs in web search
  - *Query expansion* and *relevance feedback* modify the original query with additional terms

User Interaction

- Results output
  - Constructs the display of ranked documents for a query
  - Generates *snippets* to show how queries match documents
  - *Highlights* important words and passages
  - Retrieves appropriate *advertising* in many applications ("related" things)
  - May provide *clustering* and other visualization tools

Query Process

Core of the search engine: generates a ranked list of documents for the user's query

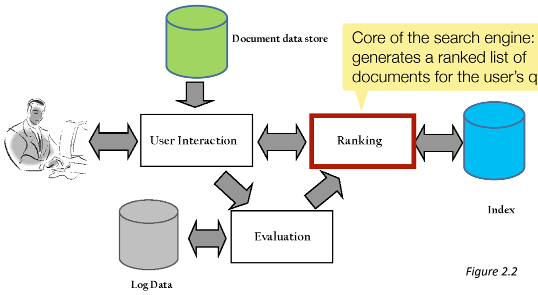


Figure 2.2

Ranking

- Scoring
  - Calculates scores for documents using a *ranking algorithm*, which is based on a *retrieval model*
  - Core component of search engine
  - Basic form of score is $\sum_i q_i d_i$
    - q_i and d_i are query and document term weights for term i
  - Many variations of ranking algorithms and retrieval models

Ranking

- Performance optimization
  - Designing ranking algorithms for efficient processing
    - *Term-at-a time* vs. *document-at-a-time* processing
    - *Safe* vs. *unsafe* optimizations
- Distribution
  - Processing queries in a distributed environment
  - *Query broker* distributes queries and assembles results
  - *Caching* is a form of distributed searching

Query Process

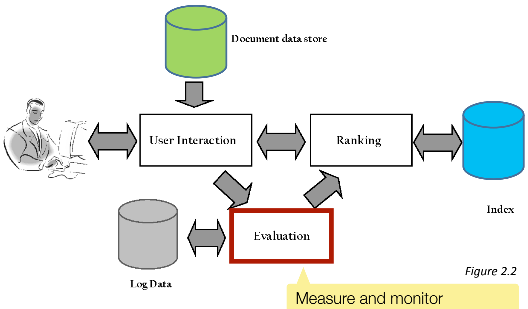


Figure 2.2

Measure and monitor effectiveness and efficiency. Record and analyze usage data

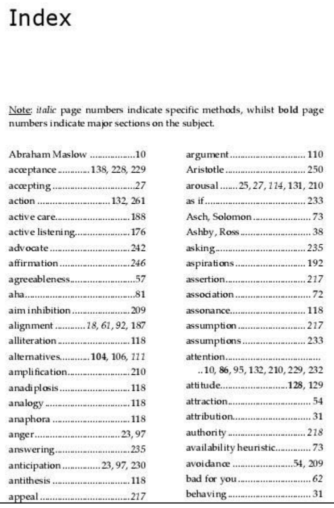
Evaluation

- Logging
  - Logging user queries and interaction is crucial for improving search effectiveness and efficiency
  - *Query logs* and *clickthrough data* used for query suggestion, spell checking, query caching, ranking, advertising search, and other components
- Ranking analysis
  - Measuring and tuning ranking effectiveness
- Performance analysis
  - Measuring and tuning system efficiency

Indexing

Indices

- Indices are data structures designed to make search faster
- Text search has unique requirements, which leads to unique data structures
- Most common data structure is the *inverted index*
  - General name for a class of structures
  - "Inverted" because documents are associated with words, rather than words with documents
    - Similar to a concordance



Index

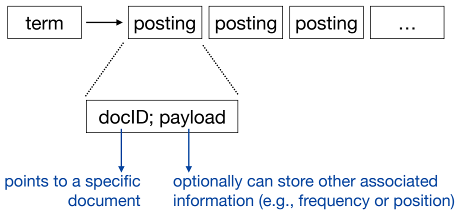
<u>Note</u>: *italic* page numbers indicate specific methods, whilst **bold** page numbers indicate major sections on the subject.

Abraham Maslow	10	argument	110
acceptance	138, 228, 229	Aristotle	250
accepting	*27*	arousal	25, *27*, *114*, 131, 210
action	132, 261	as if	233
active care	188	Asch, Solomon	*73*
active listening	176	Ashby, Ross	38
advocate	242	asking	*235*
affirmation	*246*	aspirations	192
agreeableness	57	assertion	*217*
aha	81	association	72
aim inhibition	209	assonance	118
alignment	*18*, *61*, *92*, 187	assumption	*217*
alliteration	118	assumptions	233
alternatives	**104**, 106, *111*	attention	10, 86, 95, 132, 210, 229, 232
amplification	210	attitude	**128**, 129
anadiplosis	118	attraction	54
analogy	118	attribution	31
anaphora	118	authority	*218*
anger	23, 97	availability heuristic	73
answering	*235*	avoidance	54, 209
anticipation	23, 97, 230	bad for you	*62*
antithesis	118	behaving	31
appeal	*217*		

Inverted Index

- Each index term is associated with a *postings list* (or *inverted list)*
  - Contains lists of documents, or lists of word occurrences in documents, and other information
  - Each entry is called a *posting*
  - The part of the posting that refers to a specific document or location is called a *pointer*
    - Each document in the collection is given a unique number (*docID*)
  - The posting can store additional information, called the *payload*
  - Lists are usually *document-ordered* (sorted by docID)

Inverted Index


term
posting
posting
posting
…
docID; payload
points to a specific document
optionally can store other associated information (e.g., frequency or position)

Example

S_1 Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.

S_2 Fishkeepers often use the term tropical fish to refer only those requiring fresh water, with saltwater tropical fish referred to as marine fish.

S_3 Tropical fish are popular aquarium fish, due to their often bright coloration.

S_4 In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Four sentences from the Wikipedia entry for *tropical fish*

Simple Inverted Index

docID

Each document that contains the term is a posting.

No additional payload.

and	1			
aquarium	3			
are	3	4		
around	1			
as	2			
both	1			
bright	3			
coloration	3	4		
derives	4			
due	3			
environments	1			
fish	1	2	3	4
fishkeepers	2			
found	1			
fresh	2			
freshwater	1	4		
from	4			
generally	4			
in	1	4		
include	1			
including	1			
iridescence	4			
marine	2			
often	2	3		
only	2			
pigmented	4			
popular	3			
refer	2			
referred	2			
requiring	2			
salt	1	4		
saltwater	2			
species	1			
term	2			
the	1	2		
their	3			
this	4			
those	2			
to	2	3		
tropical	1	2	3	
typically	4			
use	2			
water	1	2	4	
while	4			
with	2			
world	1			

Inverted Index with Counts

docID: freq

The payload is the frequency of the term in the document.

Supports better ranking algorithms.

and	1:1			
aquarium	3:1			
are	3:1	4:1		
around	1:1			
as	2:1			
both	1:1			
bright	3:1			
coloration	3:1	4:1		
derives	4:1			
due	3:1			
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			
generally	4:1			
in	1:1	4:1		
include	1:1			
including	1:1			
iridescence	4:1			
marine	2:1			
often	2:1	3:1		
only	2:1			
pigmented	4:1			
popular	3:1			
refer	2:1			
referred	2:1			
requiring	2:1			
salt	1:1	4:1		
saltwater	2:1			
species	1:1			
term	2:1			
the	1:1	2:1		
their	3:1			
this	4:1			
those	2:1			
to	2:2	3:1		
tropical	1:2	2:2	3:1	
typically	4:1			
use	2:1			
water	1:1	2:1	4:1	
while	4:1			
with	2:1			
world	1:1			

Inverted Index with Positions

docID, position

There is a separate posting for each term occurrence in the document. The payload is the term position.

Supports proximity matches.
E.g., find "tropical" within 5 words of "fish"

and	1,15				
aquarium	3,5				
are	3,3	4,14			
around	1,9				
as	2,21				
both	1,13				
bright	3,11				
coloration	3,12	4,5			
derives	4,7				
due	3,7				
environments	1,8				
fish	1,2	1,4	2,7 3,2 4,13	2,18 3,6	2,23 4,3
fishkeepers	2,1				
found	1,5				
fresh	2,13				
freshwater	1,14	4,2			
from	4,8				
generally	4,15				
in	1,6	4,1			
include	1,3				
including	1,12				
iridescence	4,9				
marine	2,22				
often	2,2				
only	2,10				
pigmented	4,16				
popular	3,4				
refer	2,9				
referred	2,19				
requiring	2,12				
salt	1,16				
saltwater	2,16				
species	1,18				
term	2,5				
the	1,10				
their	3,9				
this	4,4				
those	2,11				
to	2,8				
tropical	1,1				
typically	4,6				
use	2,3				
water	1,17				
while	4,10				
with	2,15				
world	1,11				

Issues

- Compression
  - Inverted lists are very large
  - Compression of indexes saves disk and/or memory space
- Optimization techniques to speed up search
  - Read less data from inverted lists
    - "Skipping" ahead
  - Calculate scores for fewer documents
    - Store highest-scoring documents at the beginning of each inverted list
- Distributed indexing

Exercise

- Draw the inverted index for the following document collection

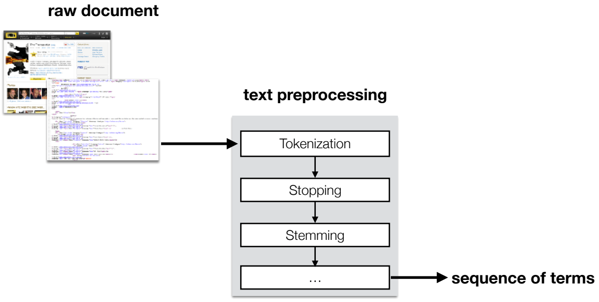
Doc 1	new home sales top forecasts
Doc 2	home sales rise in july
Doc 3	increase in home sales in july
Doc 4	july new home sales rise

Solution

new	1	4		
home	1	2	3	4
sales	1	2	3	4
top	1			
forecasts	1			
rise	2	4		
in	2	3		
july	2	3	4	
increase	3			

Text Preprocessing

Preprocessing Pipeline



Tokenization

- Parsing a string into individual words (tokens)
- Splitting is usually done along white spaces, punctuation marks, or other types of content delimiters (e.g., HTML markup)
- Sounds easy, but can be surprisingly complex, even for English
  - Even worse for many other languages

Tokenization Issues

- Apostrophes can be a part of a word, a part of a possessive, or just a mistake
  - rosie o'donnell, can't, 80's, 1890's, men's straw hats, master's degree, …
- Capitalized words can have different meaning from lower case words
  - Bush, Apple
- Special characters are an important part of tags, URLs, email addresses, etc.
  - C++, C#, …

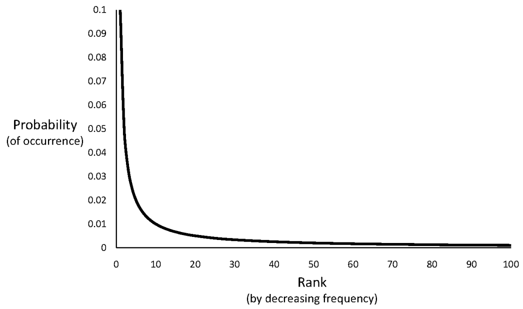
Tokenization Issues

- Numbers can be important, including decimals
  - nokia 3250, top 10 courses, united 93, quicktime 6.5 pro, 92.3 the beat, 288358
- Periods can occur in numbers, abbreviations, URLs, ends of sentences, and other situations
  - I.B.M., Ph.D., www.uis.no, F.E.A.R.

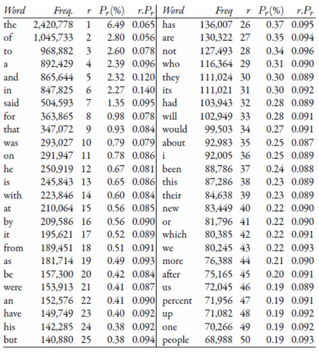
Common Practice

- First pass is focused on identifying markup or tags; second pass is done on the appropriate parts of the document structure
- Treat hyphens, apostrophes, periods, etc. like spaces
- Ignore capitalization
- Index even single characters
  - o’connor => o connor

Text Statistics



Top-50 words from AP89

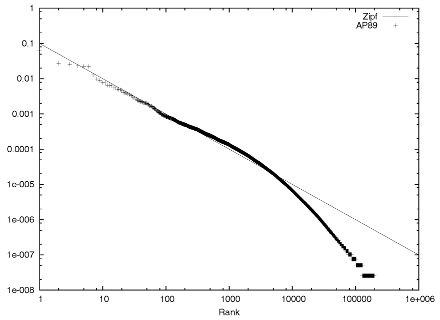


Word	*Freq.*	*r*	P_r(%)	$r.P_r$	*Word*	*Freq*	*r*	P_r(%)	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093

Zipf's Law

- Distribution of word frequencies is very *skewed*
  - A few words occur very often, many words hardly ever occur
  - E.g., two most common words ("the", "of") make up about 10% of all word occurrences in text documents
- Zipf's law:
  - Frequency of an item or event is inversely proportional to its frequency rank
  - Rank (r) of a word times its frequency (f) is approximately a constant (k): r*f~k

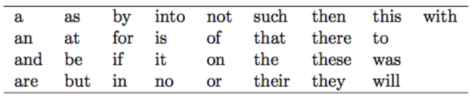
Zip's law for AP89



Stopword Removal

- Function words that have little meaning apart from other words: the, a, an, that, those, …
- These are considered *stopwords* and are removed
- A stopwords list can be constructed by taking the top n (e.g., 50) most common words in a collection
  - May be customized for certain domains or applications

Stopword Removal



a	as	by	into	not	such	then	this	with
an	at	for	is	of	that	there	to	
and	be	if	it	on	the	these	was	
are	but	in	no	or	their	they	will	

Table 2: Standard English stopwords list.

- There are problematic cases…

"to be or not to be"

Stemming

- Reduce the different forms of a word that occur to a common *stem*
  - inflectional (plurals, tenses)
  - derivational (making verbs nouns etc.)
- In most cases, these have the same or very similar meanings
- Two basic types of stemmers
  - Algorithmic
  - Dictionary-based

Stemming

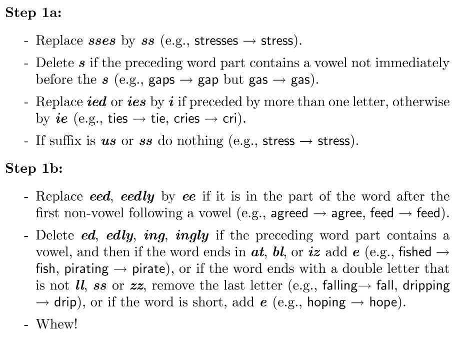
- **Suffix-s stemmer**
  - Assumes that any word ending with an s is plural
    - cakes => cake, dogs =>dog
  - Cannot detect many plural relationships (false negative)
    - centuries => century
  - In rare cases it detects a relationship where it does not exist (false positive)
    - is => i

Stemming

- **Porter stemmer**
  - Most popular algorithmic stemmer
  - Consists of 5 steps, each step containing a set of rules for removing suffixes
  - Produces stems not words
  - Makes a number of errors and difficult to modify

Porter Stemmer

- Example step (1 of 5)



Step 1a:

- Replace ***sses*** by ***ss*** (e.g., stresses → stress).
- Delete ***s*** if the preceding word part contains a vowel not immediately before the ***s*** (e.g., gaps → gap but gas → gas).
- Replace ***ied*** or ***ies*** by ***i*** if preceded by more than one letter, otherwise by ***ie*** (e.g., ties → tie, cries → cri).
- If suffix is ***us*** or ***ss*** do nothing (e.g., stress → stress).

Step 1b:

- Replace ***eed***, ***eedly*** by ***ee*** if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed → agree, feed → feed).
- Delete ***ed***, ***edly***, ***ing***, ***ingly*** if the preceding word part contains a vowel, and then if the word ends in ***at***, ***bl***, or ***iz*** add ***e*** (e.g., fished → fish, pirating → pirate), or if the word ends with a double letter that is not ***ll***, ***ss*** or ***zz***, remove the last letter (e.g., falling→ fall, dripping → drip), or if the word is short, add ***e*** (e.g., hoping → hope).
- Whew!

Porter Stemmer

should not have the same stem

should have the same stem

False positives	*False negatives*
organization/organ	european/europe
generalization/generic	cylinder/cylindrical
numerical/numerous	matrices/matrix
policy/police	urgency/urgent
university/universe	create/creation
addition/additive	analysis/analyses
negligible/negligent	useful/usefully
execute/executive	noise/noisy
past/paste	decompose/decomposition
ignore/ignorant	sparse/sparsity
special/specialized	resolve/resolution
head/heading	triangle/triangular

Stemming

- **Krovetz stemmer**
  - Hybrid algorithmic-dictionary
  - Word checked in dictionary
    - If present, either left alone or replaced with exception stems
    - If not present, word is checked for suffixes that could be removed
  - After removal, dictionary is checked again
  - Produces words not stems

Stemmer Comparison

Original text

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals, pesticide, herbicide, fungicide, insecticide, fertilizer, predicted sales, market share, stimulate demand, price cut, volume of sales

Porter stemmer

market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem pesticid herbicid fungicid insecticid fertil predict sale stimul demand price cut volum sale

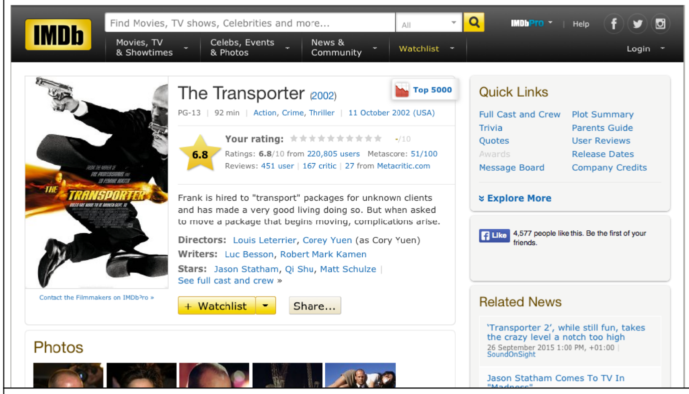
Krovetz stemmer

marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochemic pesticide herbicide fungicide insecticide fertilizer predict sale stimulate demand price cut volume sale

Stemming

- Generally a small (but significant) effectiveness improvement for English
- Can be crucial for some languages (e.g., Arabic, Russian)

Example



The Transporter (2002)

PG-13 | 92 min | Action, Crime, Thriller | 11 October 2002 (USA)

Your rating: -/10

6.8 Ratings: 6.8/10 from 220,805 users Metascore: 51/100

Reviews: 451 user | 167 critic | 27 from Metacritic.com

Frank is hired to "transport" packages for unknown clients and has made a very good living doing so. But when asked to move a package that begins moving, complications arise.

Directors: Louis Leterrier, Corey Yuen (as Cory Yuen)

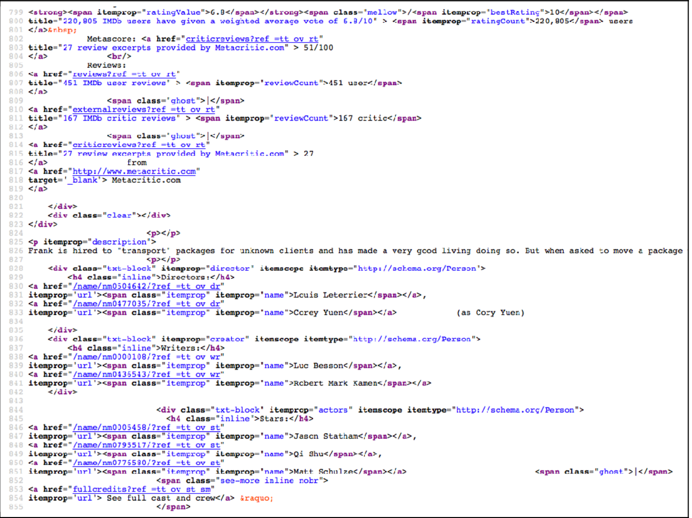
Writers: Luc Besson, Robert Mark Kamen

Stars: Jason Statham, Qi Shu, Matt Schulze | See full cast and crew »

4,577 people like this. Be the first of your friends.

'Transporter 2', while still fun, takes the crazy level a notch too high
26 September 2015 1:00 PM, +01:00 | SoundOnSight

Jason Statham Comes To TV In "Madness"



```
<strong><span itemprop="ratingValue">6.8</span></strong><span class="mellow">/<span itemprop="bestRating">10</span></span>
title="220,805 IMDb users have given a weighted average vote of 6.8/10" > <span itemprop="ratingCount">220,805</span> users
</a>&nbsp;
              Metascore: <a href="criticreviews?ref_=tt_ov_rt"
title="27 review excerpts provided by Metacritic.com" > 51/100
</a>          <br/>
              Reviews:
<a href="reviews?ref_=tt_ov_rt"
title="451 IMDb user reviews' > <span itemprop='reviewCount">451 user</span>
</a>
                <span class='ghost">|</span>
<a href="externalreviews?ref_=tt_ov_rt"
title="167 IMDb critic reviews' > <span itemprop="reviewCount">167 critic</span>
</a>
                <span class='ghost">|</span>
<a href="criticreviews?ref_=tt_ov_rt"
title="27 review excerpts provided by Metacritic.com" > 27
</a>                  from
<a href="http://www.metacritic.com"
target='_blank'> Metacritic.com
</a>

    </div>
    <div class="clear"></div>
</div>
                    <p></p>
<p itemprop="description">
Frank is hired to "transport' packages for unknown clients and has made a very good living doing so. But when asked to move a package
                    <p></p>
    <div class="txt-block" itemprop="director' itemscope itemtype="http://schema.org/Person'>
      <h4 class="inline">Directors:</h4>
<a href="/name/nm0504642/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop='name">Louis Leterrier</span></a>,
<a href="/name/nm0477035/?ref_=tt_ov_dr"
itemprop='url'><span class="itemprop" itemprop='name">Corey Yuen</span></a>            (as Cory Yuen)

    </div>
    <div class="txt-block" itemprop="creator" itemscope itemtype="http://schema.org/Person">
      <h4 class="inline">Writers:</h4>
<a href="/name/nm0000108/?ref_=tt_ov_wr"
itemprop='url'><span class="itemprop" itemprop='name">Luc Besson</span></a>,
<a href="/name/nm0436543/?ref_=tt_ov_wr"
itemprop='url'><span class="itemprop" itemprop='name">Robert Mark Kamen</span></a>
    </div>

                   <div class="txt-block' itemprop="actors" itemscope itemtype="http://schema.org/Person">
                    <h4 class="inline'>Stars:</h4>
<a href="/name/nm0005458/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop='name">Jason Statham</span></a>,
<a href="/name/nm0795517/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop='name">Qi Shu</span></a>,
<a href="/name/nm0776580/?ref_=tt_ov_st"
itemprop='url'><span class="itemprop" itemprop='name">Matt Schulze</span></a>                    <span class="ghost">|</span>
                   <span class="see-more inline nobr">
<a href="fullcredits?ref_=tt_ov_st_sm"
itemprop='url'> See full cast and crew</a> &raquo;
                   </span>
```

First pass extraction

The Transporter (2002)
PG-13 92 min Action, Crime, Thriller 11 October 2002 (USA)

Frank is hired to "transport" packages for unknown clients and has made a very good living doing so. But when asked to move a package that begins moving, complications arise.

Tokenization

the transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank is hired to transport packages for unknown clients and has made a very good living doing so but when asked to move a package that begins moving complications arise

Stopwords removal

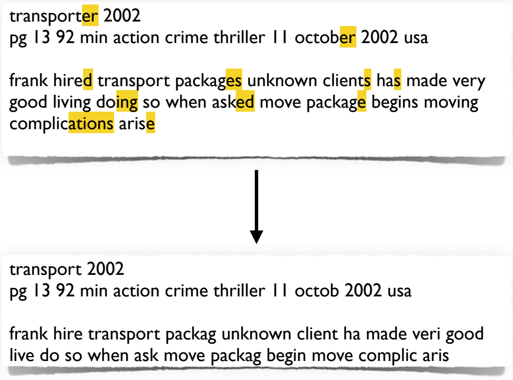
the transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank is hired to transport packages for unknown clients and has made a very good living doing so but when asked to move a package that begins moving complications arise

transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank hired transport packages unknown clients has made very good living doing so when asked move package begins moving complications arise

Stemming (Porter stemmer)



transporter 2002
pg 13 92 min action crime thriller 11 october 2002 usa

frank hired transport packages unknown clients has made very good living doing so when asked move package begins moving complications arise

transport 2002
pg 13 92 min action crime thriller 11 octob 2002 usa

frank hire transport packag unknown client ha made veri good live do so when ask move packag begin move complic aris