Information Retrieval

Dhruv Gupta

dhruv.gupta@ntnu.no

30-August-2022

NTNU | Norwegian University of Science and Technology

1. Administrative
   - Announcements
   - References
2. Recap
3. The IR Problem
4. The Boolean Model
   - Modeling Documents
   - Term Selection
   - The Model
5. The Vector Space Model
   - Term Frequency
   - Inverse Document Frequency
   - TF-IDF Weighting
   - Length Normalization
   - The Model
   - Example
6. Summary

Announcements

- The first assignment will be made available next week.
- Information regarding the tutorial sessions will be made by the TAs soon.

References

- Text and diagrams of some slides are based on the material from the book: Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval", Second Edition. Pearson Education Limited, 2011.

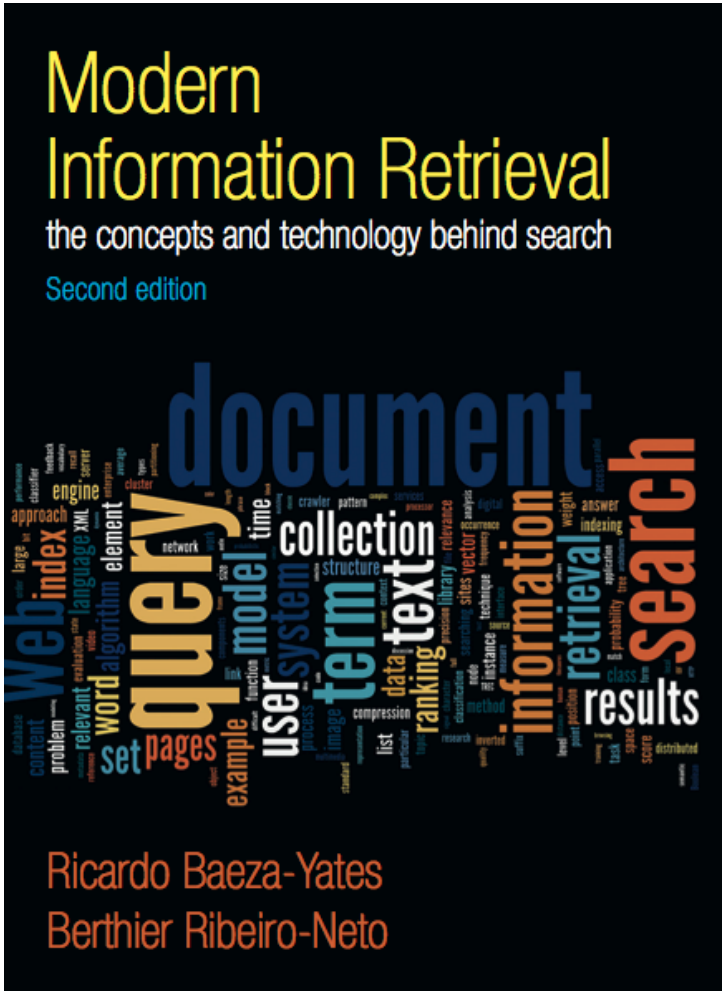


Image Credit: http://grupoweb.upf.es/mir2ed/

References

- Text and diagrams of some slides are based on the material from the book: Manning et al., "Introduction to Information Retrieval", First Edition. Cambridge University Press, 2008.

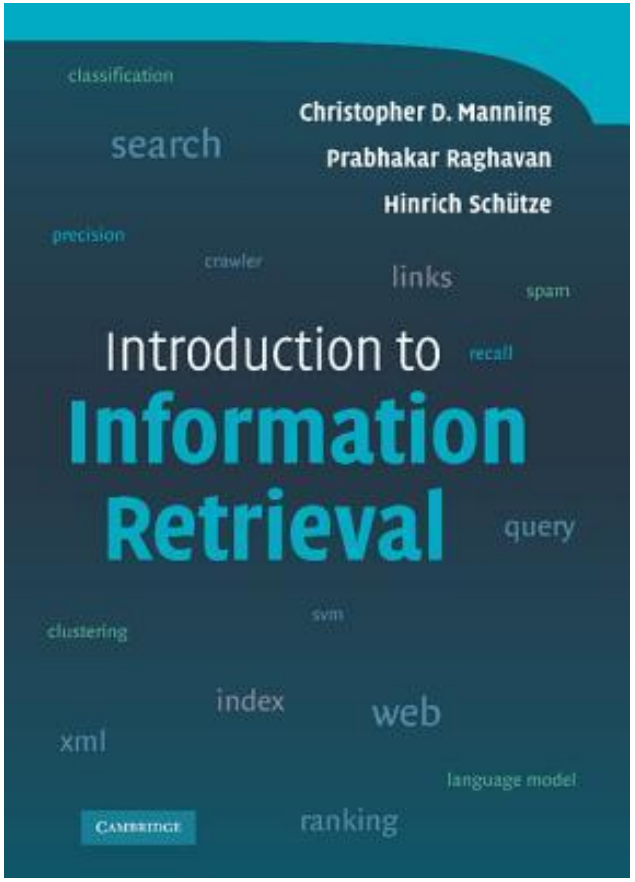


Image Credit: https://www.goodreads.com/book/show/3278309-introduction-to-information-retrieval

References

- Text and diagrams of some slides are based on the material from the book: Büttcher et al., "Information Retrieval," MIT Press, 2010.

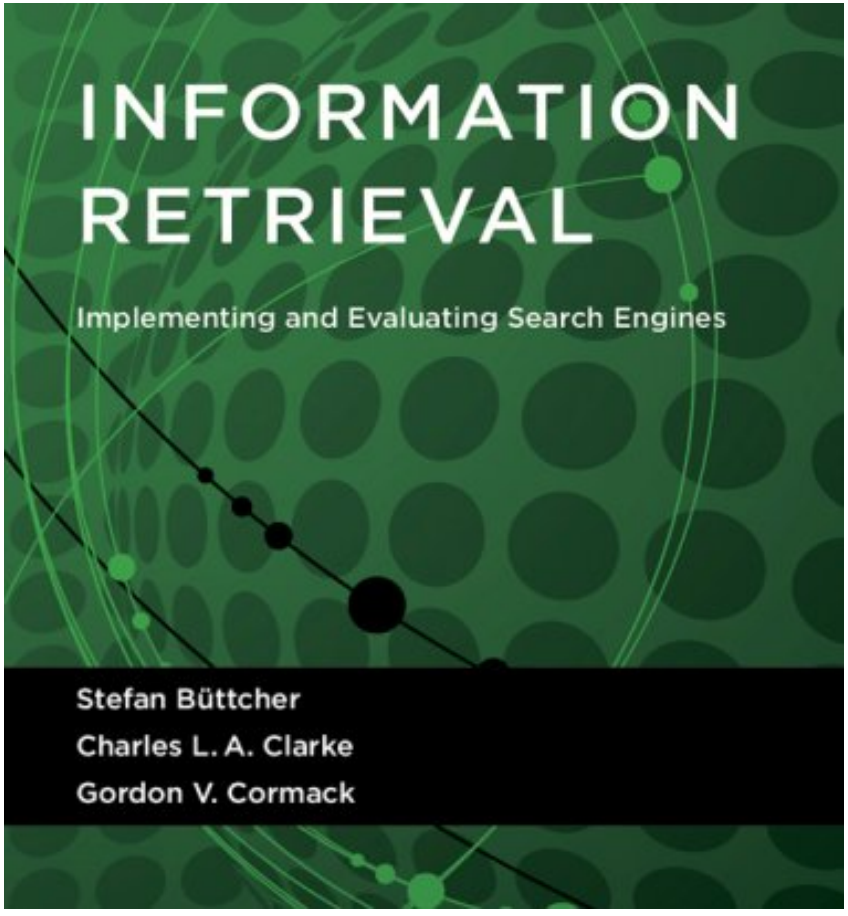


Image Credit: http://www.ir.uwaterloo.ca/book/

1. Administrative
   - Announcements
   - References
2. Recap
3. The IR Problem
4. The Boolean Model
   - Modeling Documents
   - Term Selection
   - The Model
5. The Vector Space Model
   - Term Frequency
   - Inverse Document Frequency
   - TF-IDF Weighting
   - Length Normalization
   - The Model
   - Example
6. Summary

Search in Libraries

- Information Retrieval ≡ Search.
- Search before digitization: libraries.

https://commons.wikimedia.org/wiki/File:Bookshelf_at_Yale.jpg

Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.

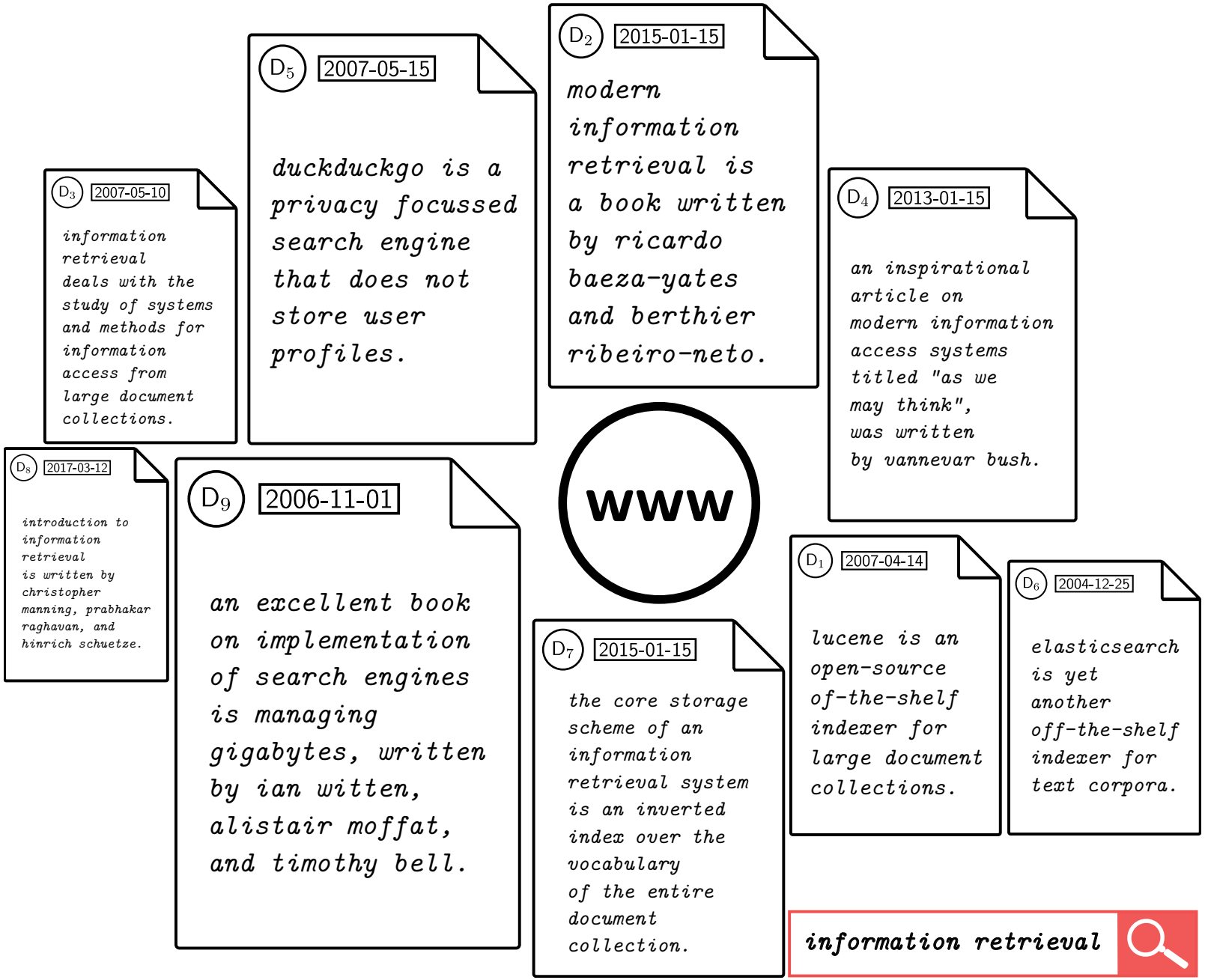
ID	Author	Book Name	ISBN	Subject
10014	Ricardo Baeza-Yates	Modern Information Retrieval	9780321416919	Computer Science

```
SELECT *
FROM   DIGITAL_LIBRARY
WHERE  BOOK_NAME LIKE '%Information Retreival%'
```

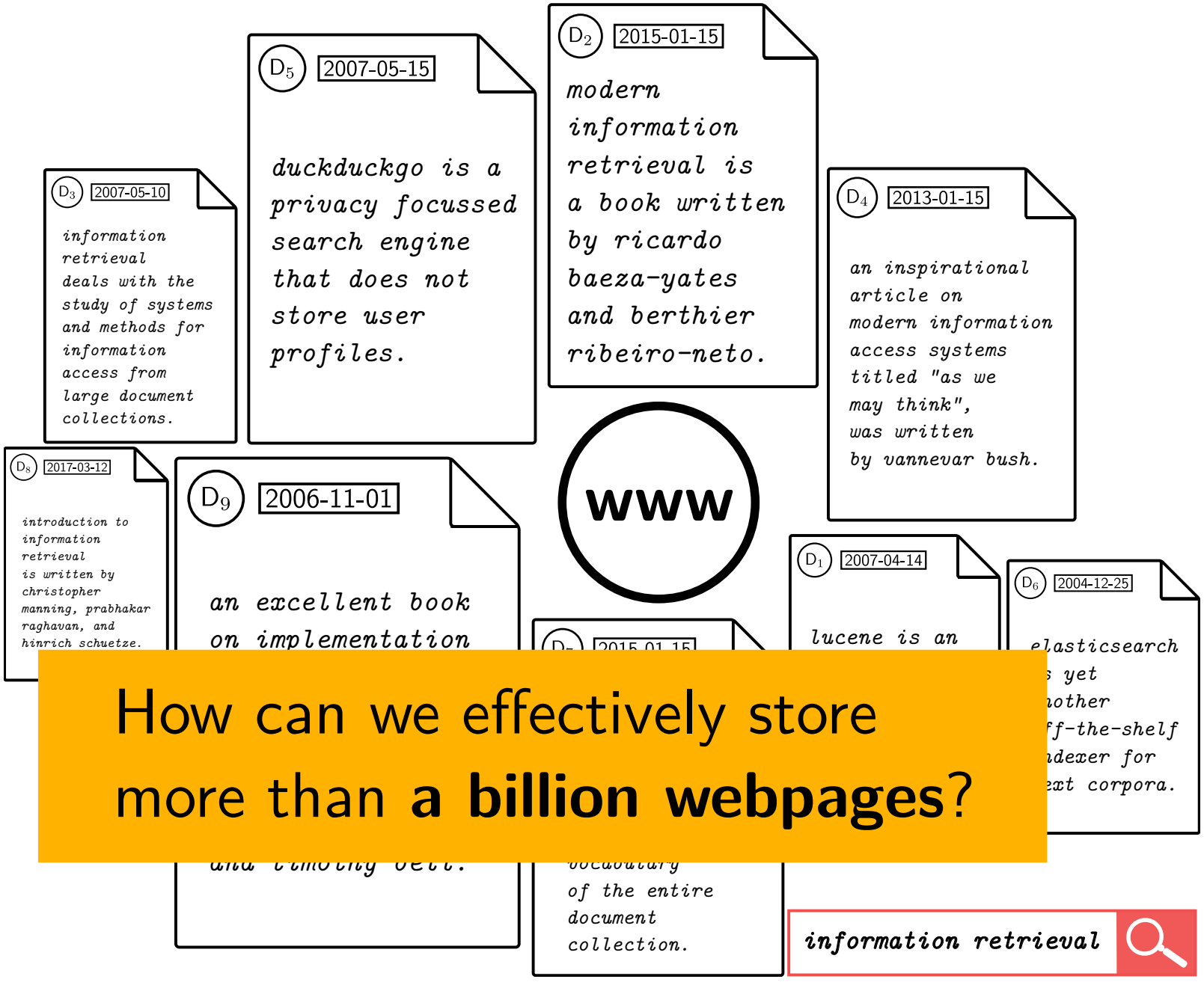
Search in Digitial Libraries

- Since, most catalogs are in structured format we can leverage a database solution.
- Naïve approach: map existing manual catalogs to digital database schema.
- Drawbacks:
  - Unstructured text within documents is in-accessible
  - Presentation of results
  - Determination of which is the best book
  - Slow
- Can not cast "Information Retrieval" as "Data Retrieval" task!

Searching the Web — Document Collection



Searching the Web — Challenge of Size



How can we effectively store more than **a billion webpages**?

Searching the Web — Users, Queries, and Interface

information retrieval

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

`https://en.wikipedia.org/wiki/Information_retrieval`

Searching the Web — Queries and Information Needs

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection

Web Search - Wikipedia

https://en.wikipedia.org/wiki/Search_engine

Wikipedia is a public portal fo

Wikipedia entry of the topic w

Modern Information Re

http://grupoweb.upf.es/mir2

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Users express their **information needs** using natural language expressions.

Searching the Web — Queries

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection

Web Search - Wikipedia

https://en.wikipedia.org/wiki/Search_engine

Wikipedia is a public portal fo

Wikipedia entry of the topic w

Modern Information Re

http://grupoweb.upf.es/mir2

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Information need can be expressed as:
keywords, phrases, and many more!

Searching the Web — Query Operators

Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection

Web Search - Wikipedia

https://en.wikipedia.org/wiki/Search_engine

Wikipedia is a public portal fo

Wikipedia entry of the topic w

Boolean Operators further enable **conjunctive, disjunctive and negation queries**!

Modern Information Re

http://grupoweb.upf.es/mir2

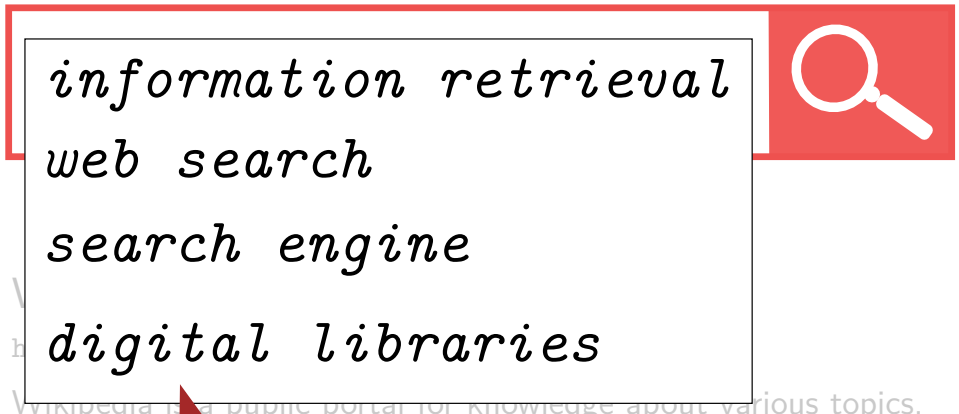
Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Query Reformulations



Information Retrieval

" Information retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

https://en.wikipedia.org/wiki/Information_retrieval

Wikipedia is a public portal for knowledge about various topics.
Wikipedia entry of the topic web search.

Modern Information Retrieval

http://grupoweb.upf.es/mir2ed/

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Query reformulations can be identified using **pseudo-relevance feedback**!

Searching the Web — Full-Text Search

information retrieval

Search results are identified by **full-text search** through document contents.

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

of obtaining information system resources that are relevant to an information need from a collection of those resources. Searches can be based on full-text or other content-based indexing."

- Wikipedia

https://en.wikipedia.org/wiki/Information_retrieval

Searching the Web — Synonym Discovery

information retrieval

Search results are also identified by **synonym discovery!**

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Relevance

information retrieval

Search results are **ranked by relevance** to the end-user!

Web Search - Wikipedia

`https://en.wikipedia.org/wiki/Search_engine`

Wikipedia is a public portal for knowledge about various topics.

Wikipedia entry of the topic web search.

Modern Information Retrieval

`http://grupoweb.upf.es/mir2ed/`

Modern information retrieval is a book written by Ricardo Baeza-Yates and Berthier Ribeiro-Neto.

Introduction to Information Retrieval

`https://nlp.stanford.edu/IR-book/information-retrieval-book.html`

Introduction to Information Retrieval is written by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze,

Searching the Web — Summary

- Information retrieval: making knowledge accessible to the user.
- Objectives:
  - Fast (millisecond response time) and fault tolerant.
  - Full-text search.
  - Ordering of documents according to relevance to the user.
  - User-centric presentation of results.

Information Retrieval — Definition

"Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)." [1]

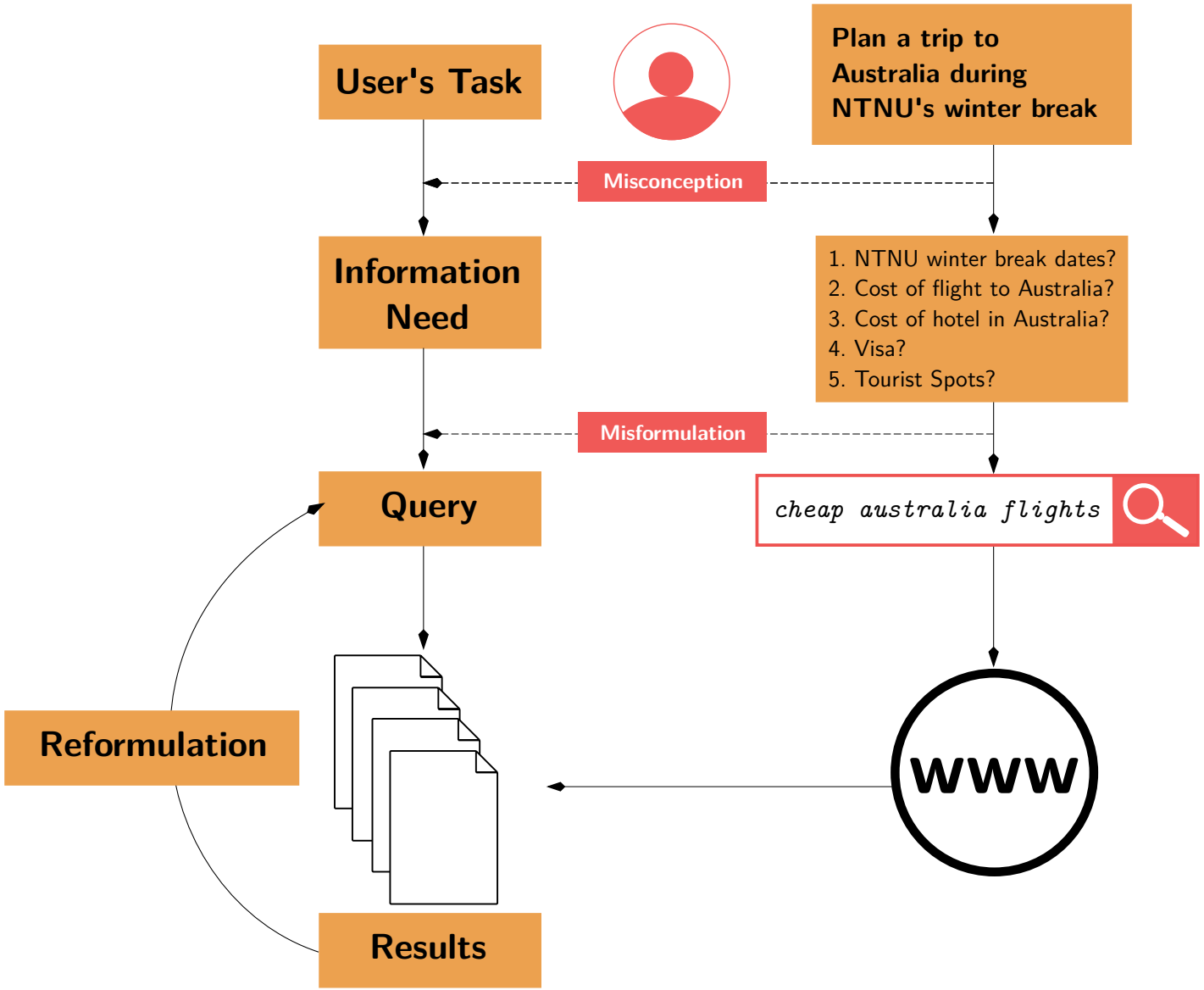
[1] Manning et al., "Introduction to Information "Retrieval", First Edition. Cambridge University Press, 2008.

Information Retrieval — Formalization

- An IR Model can be defined by a quadruple $\langle \mathcal{D}, \mathcal{Q}, \mathcal{F}, R(q, d) \rangle$, where
  - $\mathcal{D} = \{d_1, d_2, \ldots, d_N\} \equiv$ document collection,
  - $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\} \equiv$ query collection reflecting user's information needs,
  - $\mathcal{F} \equiv$ framework for modeling documents d, queries q, and their relationships.
  - $R(q, d) \equiv$ ranking function.

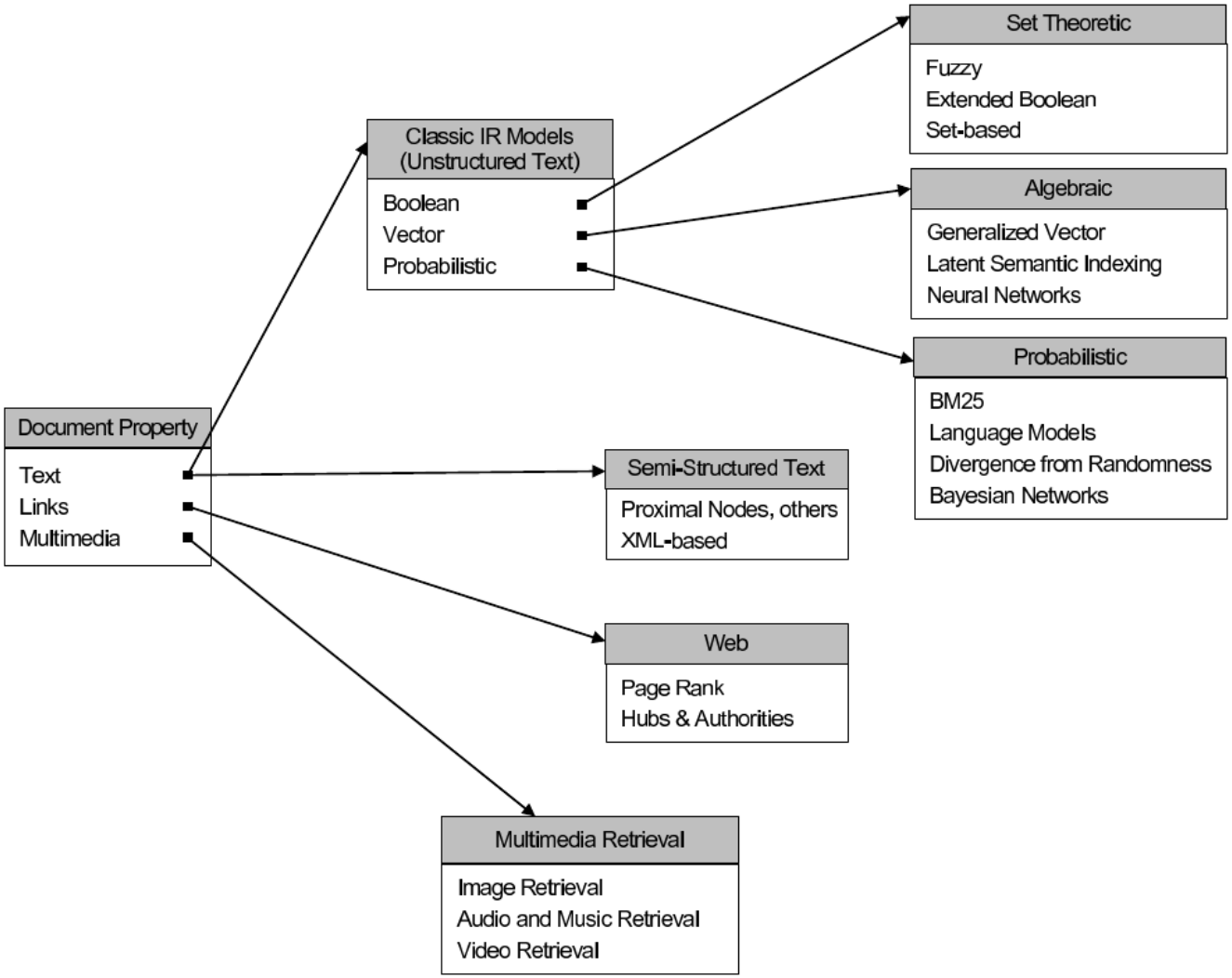
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Information Retrieval — Formalization



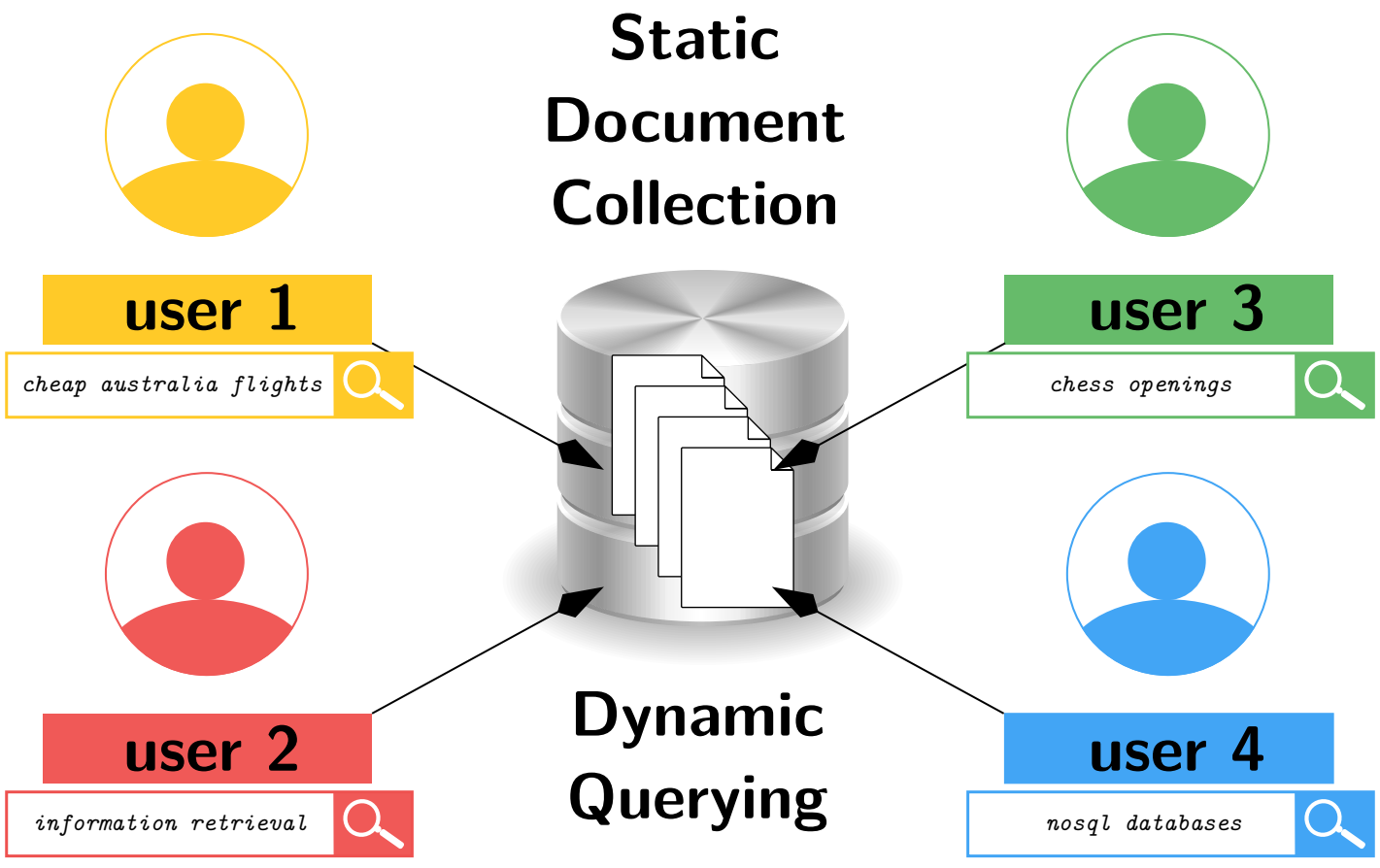
Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Information Retrieval — Formalization



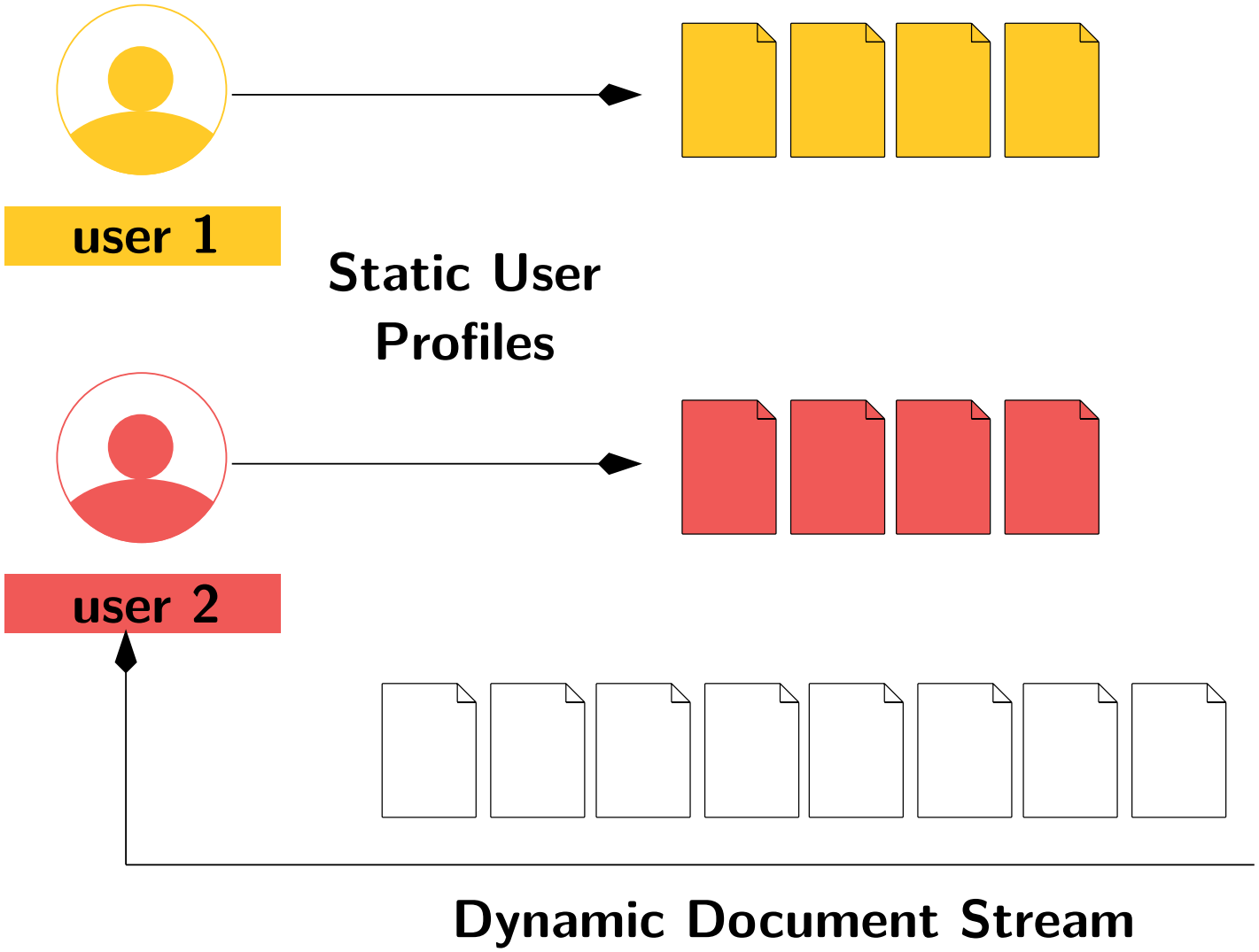
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Information Retrieval — Ad-Hoc Search



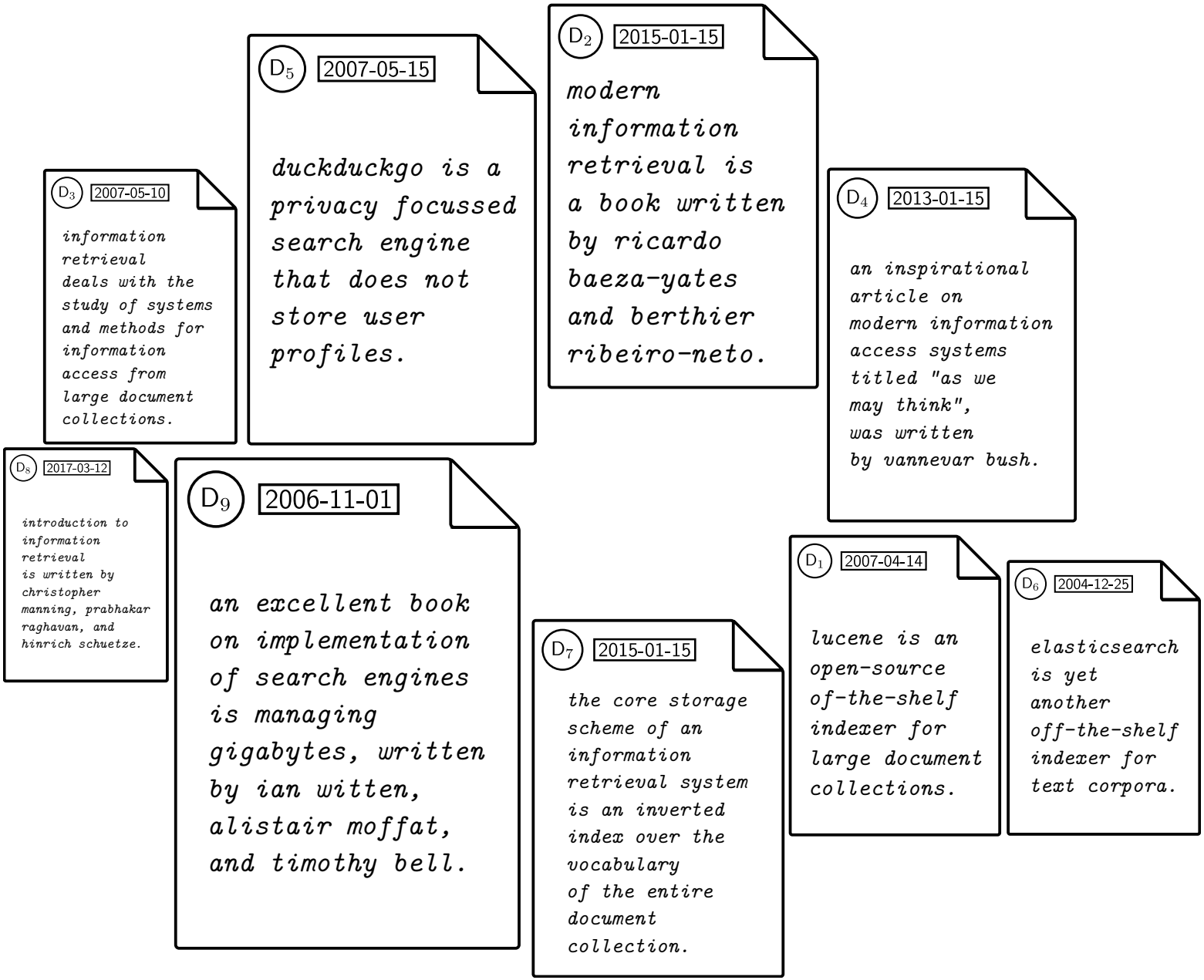
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Information Retrieval — Filtering



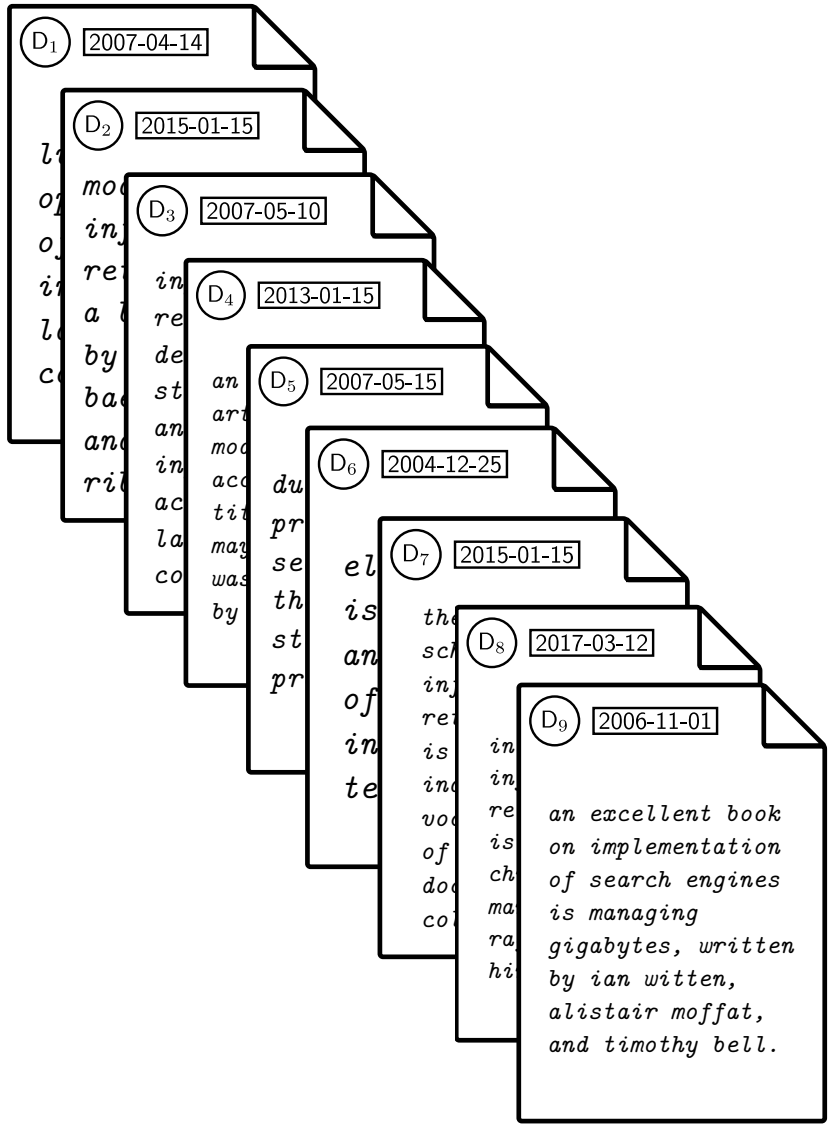
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Modeling Documents



Modeling Documents

- Document Collection, $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$.
- Document, $d = \{w_1, w_2, \ldots, w_{|d|}\}$.
- Vocabulary of Collection, $\mathcal{V} = \{w_1, w_2, \ldots, w_{|V|}\}$.
- Each term $w \in \mathcal{V}$ is distinct and occurs in $\mathcal{D}$.

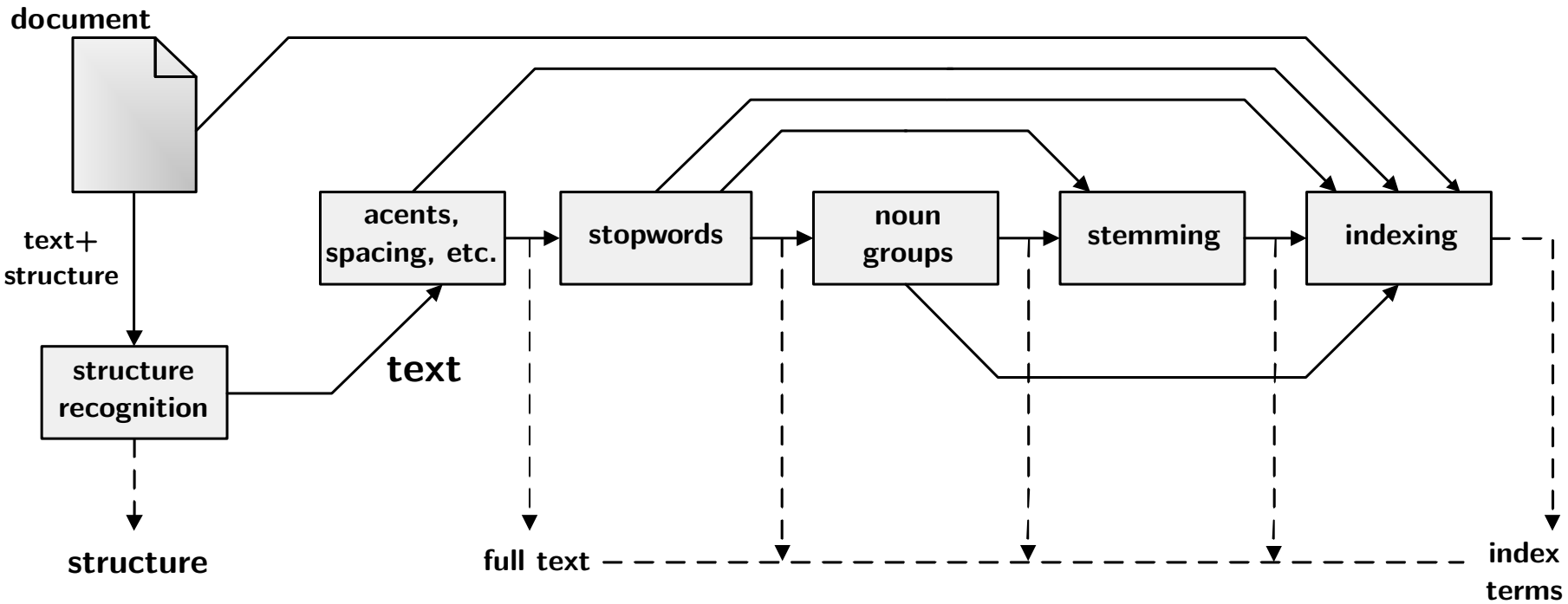


Modeling Documents — Term Document Matrix

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\lvert\mathcal{V}\rvert}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

Term Selection

- Logical view of a document: from full text to a set of index terms



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Term Selection — Punctuation and Capitalization

- Remove punctuation marks: apostrophe, periods, and commas.
- Case normalization: convert all to lowercase.

Original	**Normalized**
To be, or not to be: that is the question: Whether 'tis nobler in the mind to suffer The slings and arrows of outrageous fortune, Or to take arms against a sea of troubles, And by opposing end them? To die: to sleep; No more; and by a sleep to say we end The heart-ache and the thousand natural shocks That flesh is heir to, 'tis a consummation Devoutly to be wish'd. To die, to sleep; To sleep: perchance to dream: ay, there's the rub;	to be or not to be that is the question whether tis nobler in the mind to suffer the slings and arrows of outrageous fortune or to take arms against a sea of troubles and by opposing end them to die to sleep no more and by a sleep to say we end the heart ache and the thousand natural shocks that flesh is heir to tis a consummation devoutly to be wish d to die to sleep to sleep perchance to dream ay there s the rub

Büttcher et al., "Information Retrieval," MIT Press.

Term Selection — Stemming

- Stemming: each token is passed through the stemmer and the resulting root form is indexed.
- The Porter stemmer: one of the best-known stemmers for the English language.
- Stemmer operates by applying lists of rewrite rules organized into a sequence of steps.
- Example: the first list of rewrite rules (constituting step 1a) handles plural forms.

$$\begin{aligned} \text{sses} &\rightarrow \text{ss} \\ \text{ies} &\rightarrow \text{i} \\ \text{ss} &\rightarrow \text{ss} \\ \text{s} &\rightarrow \end{aligned}$$

Büttcher et al., "Information Retrieval," MIT Press.

Term Selection — Stemming

- Stemmer may improve recall but may also harm precision.
- Stemming is related to lemmatization from linguistics.
- Lemmatization reduces a term to a lexeme, which roughly corresponds to a word in the sense of a dictionary entry.

Normalized	Normalized & Stemmed
to be or not to be that is the question whether tis nobler in the mind to suffer the slings and arrows of outrageous fortune or to take arms against a sea of troubles and by opposing end them to die to sleep no more and by a sleep to say we end the heart ache and the thousand natural shocks that flesh is heir to tis a consummation devoutly to be wish d to die to sleep to sleep perchance to dream ay there s the rub	to be or not to be that is the question whether ti nobler in the mind to suffer the sling and arrow of outrag fortun or to take arm against a sea of troubl and by oppos end them to die to sleep no more and by a sleep to sai we end the heart ach and the thousand natur shock that flesh is heir to ti a consumm devoutli to be wish d to die to sleep to sleep perchanc to dream ay there s the rub

Büttcher et al., "Information Retrieval," MIT Press.

Term Selection — Stopword Removal

- Function words are words that have no well-defined meanings in and of themselves; rather, they modify other words or indicate grammatical relationships.
- In English, function words include: prepositions, articles, pronouns and articles, and conjunctions.
- Function words among the most frequently occurring words in any language.
- IR systems traditionally define a list of stopwords (including function words).
- Often these stopwords are removed from the query and document, and retrieval takes place on the basis of the remaining terms alone.
- Stopwords may also include single letters, digits, and other common terms, such as the state-of-being verbs.

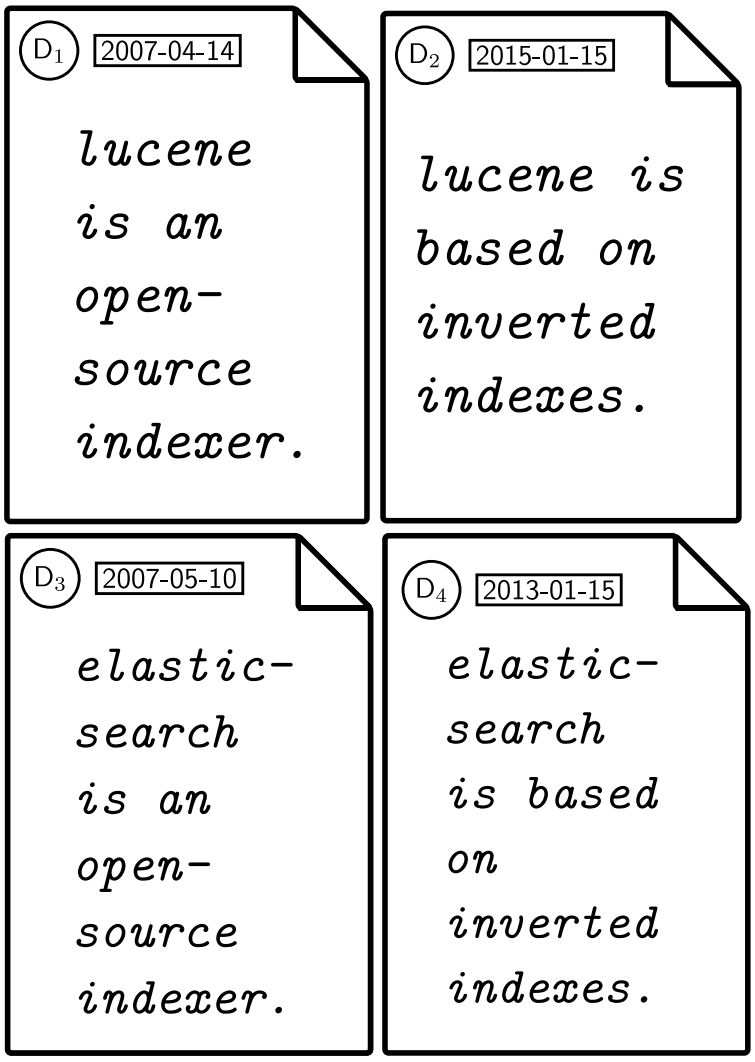
Büttcher et al., "Information Retrieval," MIT Press.

Term Selection — Stopword Removal

Normalized & Stemmed	**Normalized, Stemmed, & Stopwords Removal**
to be or not to be that is the question whether ti nobler in the mind to suffer the sling and arrow of outrag fortun or to take arm against a sea of troubl and by oppos end them to die to sleep no more and by a sleep to sai we end the heart ach and the thousand natur shock that flesh is heir to ti a consumm devoutli to be wish d to die to sleep to sleep perchanc to dream ay there s the rub	question ti nobler mind suffer sling arrow outrag fortun take arm sea troubl oppos end die sleep sleep sai end heart ach thousand natur shock flesh heir ti consumm devoutli wish die sleep sleep perchanc dream ay rub

Büttcher et al., "Information Retrieval," MIT Press.

Modeling Documents — Term Document Matrix



	lucene	elastic-search	open-source	invert	index
d_1	1	0	1	0	1
d_2	1	0	0	1	1
d_3	0	1	1	0	1
d_4	0	1	0	1	1

Boolean Retrieval

- Consider a Boolean query: $q = w_1 \wedge (w_2 \vee \neg w_3)$.
- Term vector for $w_1 = \langle 1, 1, 0, 0, 0 \rangle$.
- Term vector for $w_2 = \langle 0, 1, 0, 1, 0 \rangle$.
- Term vector for $w_3 = \langle 1, 0, 0, 0, 1 \rangle$.

	1	2	3	4	5
w_3	1	0	0	0	1
$\neg w_3$	0	1	1	1	0

	1	2	3	4	5
$\neg w_3$	0	1	1	1	0
w_2	0	1	0	1	0
OR	0	1	1	1	0

	1	2	3	4	5
$w_2 \vee \neg w_3$	0	1	1	1	0
w_1	1	1	0	0	0
AND	0	1	0	0	0

	w_1	w_2	w_3	w_4	w_5
d_1	*1*	*0*	*1*	*1*	*0*
d_2	*1*	*1*	*0*	*0*	*1*
d_3	*0*	*0*	*0*	*1*	*0*
d_4	*0*	*1*	*0*	*0*	*1*
d_5	*0*	*0*	*1*	*0*	*0*

The Boolean Model — Disadvantages

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\lvert\mathcal{V}\rvert}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

Boolean Retrieval — Disadvantages

- Consider following collection statistics:
  - $|\mathcal{D}| = 1 \cdot 10^6 = 1M$
  - $|d| = 1 \cdot 10^3 = 1K$
  - $|\mathcal{V}| = 5 \cdot 10^5 = 500K$
  - Dimensions of term-document matrix: $1M \times 500K$
  - Number of 1's in term-document matrix: $\leq 1M \cdot 1K = 1B$
  - Number of entries in term-document matrix: $5 \cdot 10^{11} = 0.5T$
  - Percentage of non-zero entries: 0.2%
  - Percentage of zero entries: 99.8%
  - nnz <<< nz.

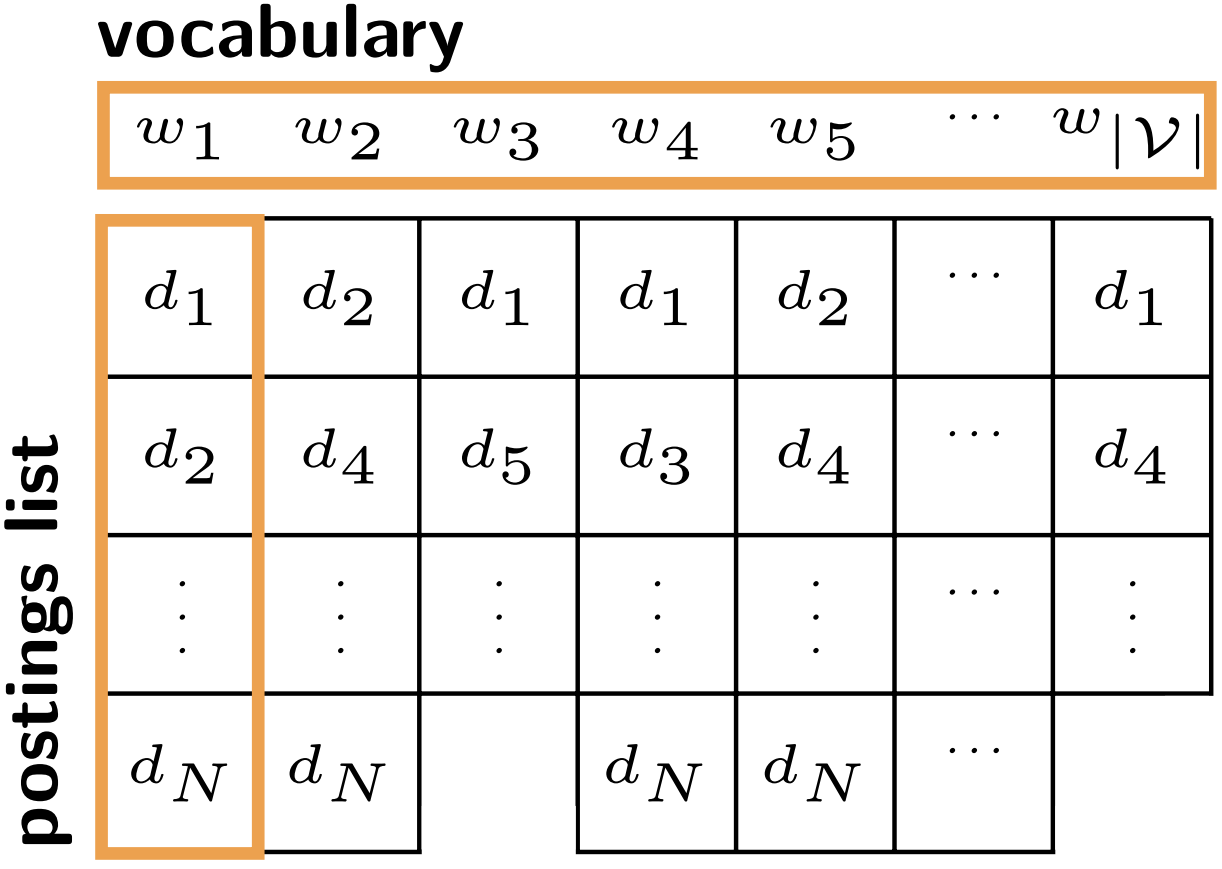
Boolean Retrieval — Disadvantages

- Consider following collection statistics:
  - $|\mathcal{D}| = 1 \cdot 10^6 = 1M$
  - $|d| = 1 \cdot 10^3 = 1K$
  - $|\mathcal{V}| = 5 \cdot 10^5 = 500K$
  - Dimensions of term-document matrix: $1M \times 500K$
  - Number of 1's in term-document matrix: $\leq 1M \cdot 1K = 1B$
  - Number of entries in term-document matrix: $5 \cdot 10^{11} = 0.5T$
  - Percentage of non-zero entries: 0.2%
  - Percentage of zero entries: 99.8%
  - nnz <<< nz.
- Term-Document Matrix is SPARSE!

The Boolean Model — Insight into Inverted Indexes

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\lvert\mathcal{V}\rvert}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

The Boolean Model — Insight into Inverted Indexes



Boolean Retrieval — Disadvantages

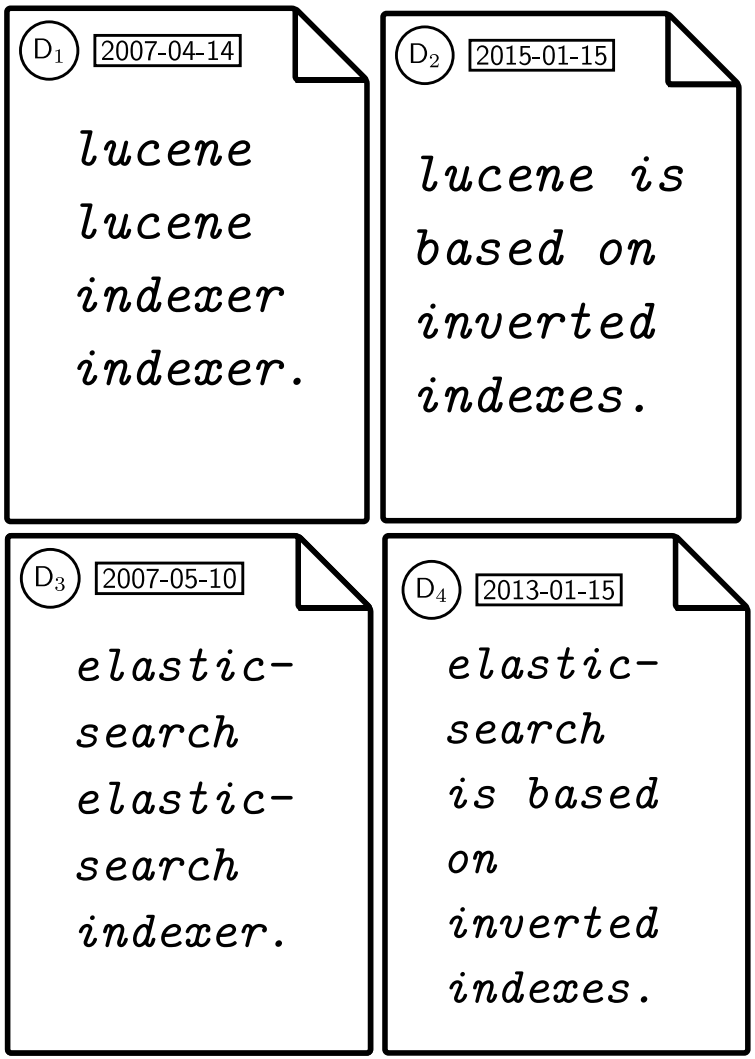
- Consider following collection statistics:
  - $|\mathcal{D}| = 1 \cdot 10^6 = 1M$
  - $|d| = 1 \cdot 10^3 = 1K$
  - $|\mathcal{V}| = 5 \cdot 10^5 = 500K$
  - Dimensions of term-document matrix: $1M \times 500K$
  - Number of 1's in term-document matrix: $\leq 1M \cdot 1K = 1B$
  - Number of entries in term-document matrix: $5 \cdot 10^{11} = 0.5T$
  - Percentage of non-zero entries: 0.2%
  - Percentage of zero entries: 99.8%
  - nnz <<< nz.
- Term-Document Matrix is SPARSE!
- There is no ranking amongst the documents in the answer set.
- Boolean operators allow only exact matches to terms.

Weighted Term Vectors — Term Frequency

- Assign weights to terms instead of Boolean values in the term-document matrix.
- Naïve Weight Assignment: frequency of word in the document.
- Weight of term t in document d is its frequency: $m_{i,j} = tf_{d,t}$.

	w_1	w_2	w_3	w_4	w_5
d_1	2	0	0	2	0
d_2	1	0	0	1	1
d_3	0	2	1	2	0
d_4	0	1	0	1	1

Weighted Term Vectors — Term Frequency



	lucene	elastic-search	open-source	invert	index
d_1	2	0	0	2	0
d_2	1	0	0	1	1
d_3	0	2	1	2	0
d_4	0	1	0	1	1

Weighted Term Vectors — Term Frequency

- A variant of TF weight used in the literature is:

$$m_{i,j} = 1 + \log\left[\mathrm{tf}_{d,t}\right].$$

- The log expression is a the preferred form because it makes them directly comparable to IDF weights, as we later discuss.

	w_1	w_2	w_3	w_4	w_5
d_1	2	0	0	2	0
d_2	1	0	0	1	1
d_3	0	2	1	2	0
d_4	0	1	0	1	1

Weighted Term Vectors — Term Frequency

- **Collection Frequency** (cf). # times term t occurs in collection $\mathcal{D}$:

$$\sum_{d \in \mathcal{D}}^{|\mathcal{D}|} tf_{d,t}.$$

- **Document Frequency** (df). # of documents d in which the term t occurs in the collection $\mathcal{D}$.
- Document frequency of a term t is always less than its collection frequency.

	w_1	w_2	w_3	w_4	w_5
d_1	2	0	0	2	0
d_2	1	0	0	1	1
d_3	0	2	1	2	0
d_4	0	1	0	1	1

Weighted Term Vectors — Term Frequency

- Document frequency of a term t is always less than its collection frequency.
- Sample term statistics from Reuter's corpus [2]

Term	cf	df
insurance	10440	3997
try	10422	8760

- Relying only on term frequency alone doesn't give us enough discriminating power for producing a good ranking.

	w_1	w_2	w_3	w_4	w_5
d_1	2	0	0	2	0
d_2	1	0	0	1	1
d_3	0	2	1	2	0
d_4	0	1	0	1	1

[2] Manning et al. "Introduction to Information Retrieval," Cambridge University Press.

Weighted Term Vectors — Inverse Document Frequency

- A term that has infrequent is more selective than a term that is frequent.
- Specificity: property of term semantics.
  - A term is more or less specific depending on its meaning.
  - Example: *beverage* is less specific than *tea* and *beer*.
  - Expect that *beverage* occurs in more documents than *tea* and *beer*.
  - Term specificity should be interpreted as a statistical rather than semantic property of the term
- Statistical term specificity: the inverse of the number of documents in which the term occurs.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Weighted Term Vectors — Inverse Document Frequency

- Commonly used model of distribution of terms in a collection is Zipf's law.
- Zipf's Law: if t_1 is the most common term in the collection, t_2 is the next most common, and so on, then the document frequency df_j of the jth most common term is proportional to $1/j$:

$$df_j \propto \frac{1}{j}$$

- Alternative formulation:

$$df_j = c \cdot j^{-k}, \text{ where, } c \text{ and } k \text{ are constants.}$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Weighted Term Vectors — Inverse Document Frequency

- Alternative formulation:

$$
\begin{aligned}
df_j &= c \cdot j^{-1}, && \text{take } k = 1 \text{ for English collections.} \\
\log(\mathrm{df}_j) &= \log(c \cdot j^{-1}), && \text{take log on both sides.} \\
\log(\mathrm{df}_j) &= \log(c) - \log(j), && \text{take log on both sides.} \\
\log(\mathrm{df}_j) &= \log(N) - \log(j), && \text{for } j = 1,\ \log(\mathrm{c}) = \mathrm{df}_1 \approx N.
\end{aligned}
$$

$$
\mathrm{idf}_j \equiv \log\left[\frac{N}{\mathrm{df}_j}\right]
$$

- idf_j is called the inverse document frequency of term j.
- IDF provides a foundation for modern term weighting schemes and is used for ranking in almost all IR systems.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Weighted Term Vectors — TF-IDF

- The best known term weighting schemes use weights that combine idf factors with term frequencies.
- TF-IDF weight of term t in document d:

$$m_{d,t} = \left[1 + \log\left[tf_{d,t}\right]\right] \cdot \left[\log\left[\frac{N}{\mathrm{df}_t}\right]\right].$$

- Relevance of a document to a query:

$$\mathrm{score}(q, d) = \sum_{t \in q \cap d} \left[1 + \log\left[tf_{d,t}\right]\right] \cdot \left[\log\left[\frac{N}{\mathrm{df}_t}\right]\right].$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Weighted Term Vectors — TF Variants

- For TF weights, five distinct variants are illustrated below:

Variant	TF Weight
Binary	$\{0, 1\}$
Raw Frequency	$\mathrm{tf}_{i,j}$
Log Normalization	$1 + \log\left[\mathrm{tf}_{i,j}\right]$
Double Normalization 0.5	$0.5 + 0.5 \cdot \frac{\mathrm{tf}_{i,j}}{\max_j \mathrm{tf}_{i,j}}$
Double Normalization K	$K + (1\text{-}K) \cdot \frac{\mathrm{tf}_{i,j}}{\max_j \mathrm{tf}_{i,j}}$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Weighted Term Vectors — IDF Variants

- For IDF weights, five distinct variants are illustrated below:

Variant	IDF Weight
Unary	$\{1\}$
Inverse Frequency	$\log \frac{N}{\mathrm{df}_j}$
Inverse Frequency Smooth	$\log\left[1 + \frac{N}{\mathrm{df}_j}\right]$
Inverse Frequency Max	$\log\left[1 + \frac{\max_j \mathrm{df}_j}{\mathrm{df}_j}\right]$
Probabilistic Inverse Frequency	$\log\left[\frac{N - \mathrm{df}_j}{\mathrm{df}_j}\right]$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

TF-IDF Variants

- For TF-IDF weights, three distinct variants are illustrated below:

Variant	Document Term Weight	Query Term Weight
1	$\left[\mathrm{tf}_{i,j}\right] \cdot \log\left[\frac{N}{\mathrm{df}_j}\right]$	$\left[0.5 + 0.5 \cdot \frac{\mathrm{tf}_{i,j}}{\max_j \mathrm{tf}_{i,j}}\right] \cdot \log\left[\frac{N}{df_j}\right]$
2	$1 + \log\left[\mathrm{tf}_{i,j}\right]$	$\log\left[1 + \frac{N}{df_j}\right]$
3	$\left[1 + \log(\mathrm{tf}_{i,j})\right] \cdot \log\left[\frac{N}{\mathrm{df}_j}\right]$	$\left[1 + \log(\mathrm{tf}_{q,j})\right] \cdot \log\left[\frac{N}{\mathrm{df}_j}\right]$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Length Normalization

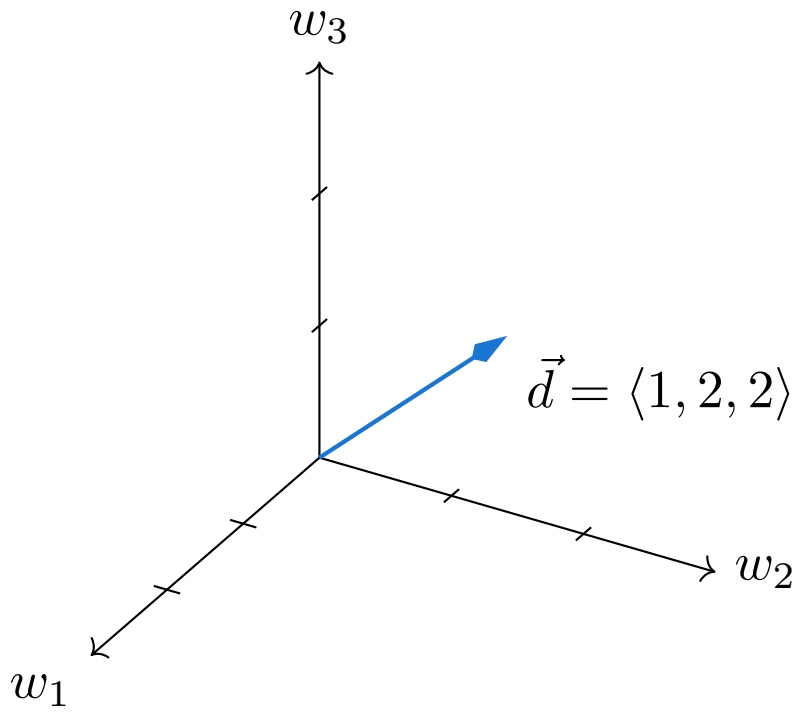
- Document sizes might vary widely.
- Problem: longer documents are more likely to be retrieved by a given query.
- Solution: document length normalization.
- Methods:
  1. Size in Bytes: consider each document is represented as a stream of bytes.
  2. Number of Words: each document is represented as a single string, and the document length is the number of words in it.
  3. Vector Norms: documents are represented as vectors of weighted terms.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Length Normalization — Vector Norm

- Documents represented as vectors of weighted terms
- The document representation $\vec{d}$ is a vector composed of all its term vector components:

$$\vec{d} = \langle m_1, m_2, \ldots, m_{|\mathcal{V}|} \rangle$$

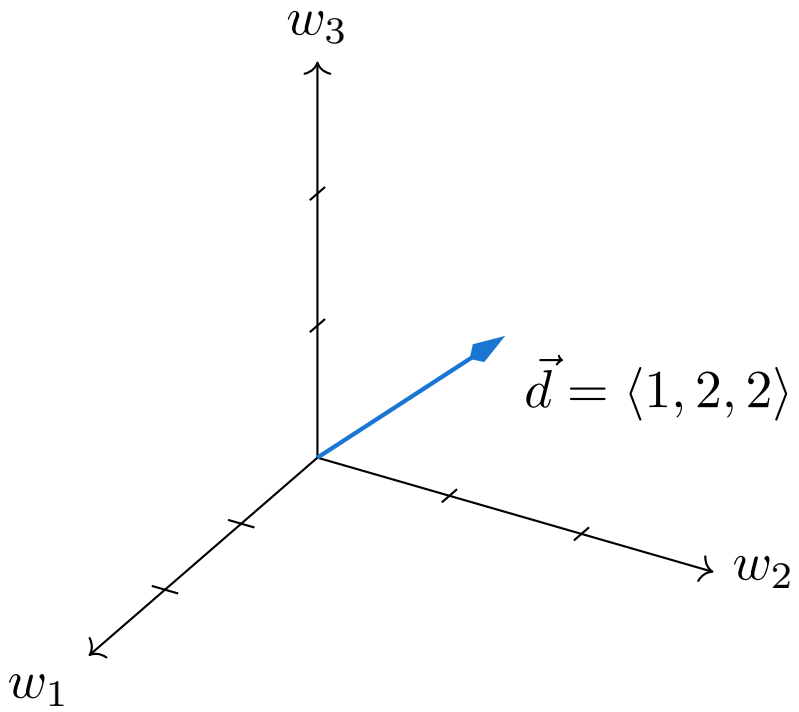


Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Length Normalization — Vector Norm

- Length normalization is essentially converting the vector to a unit vector:

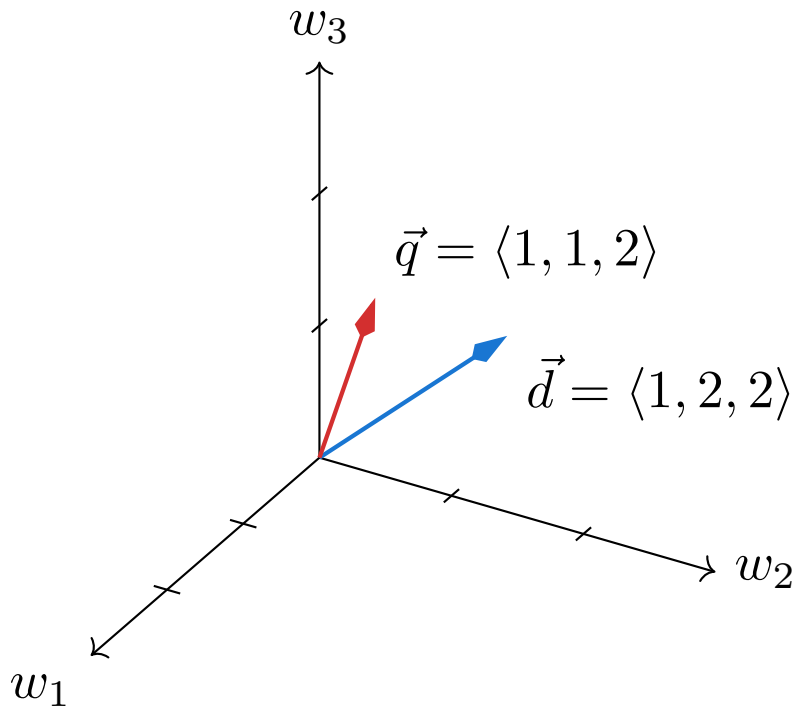
$$\hat{d} = \frac{\vec{d}}{|\vec{d}|} = \frac{\vec{d}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_j^2}}$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model

- Boolean matching and binary weights is too limiting.
- The vector model proposes a framework in which partial matching is possible.
- This is accomplished by assigning non-binary weights to index terms in queries and in documents.

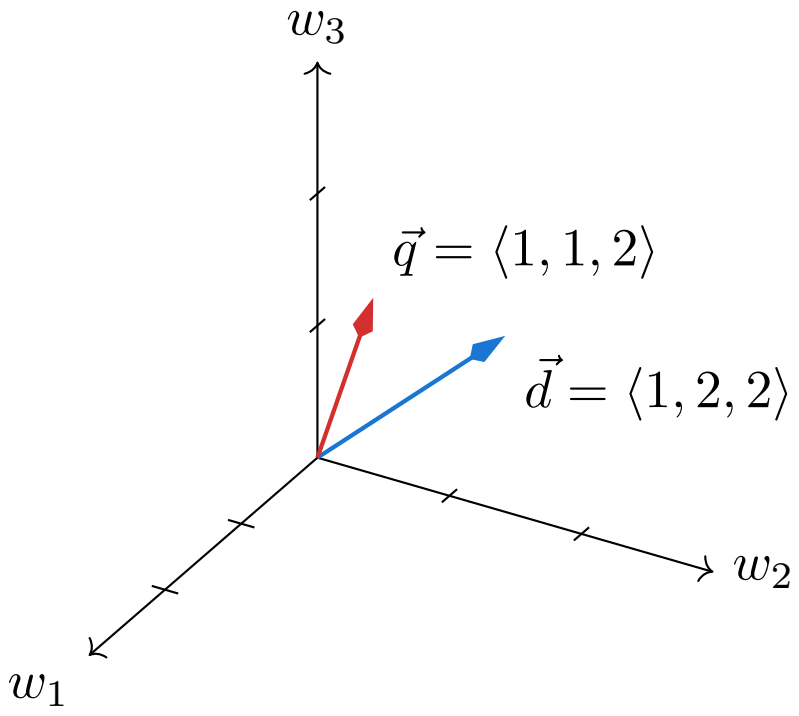


$$\vec{d}_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,|\mathcal{V}|} \rangle$$
$$\vec{q} = \langle m_{q,1}, m_{q,2}, \ldots, m_{q,|\mathcal{V}|} \rangle$$

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model

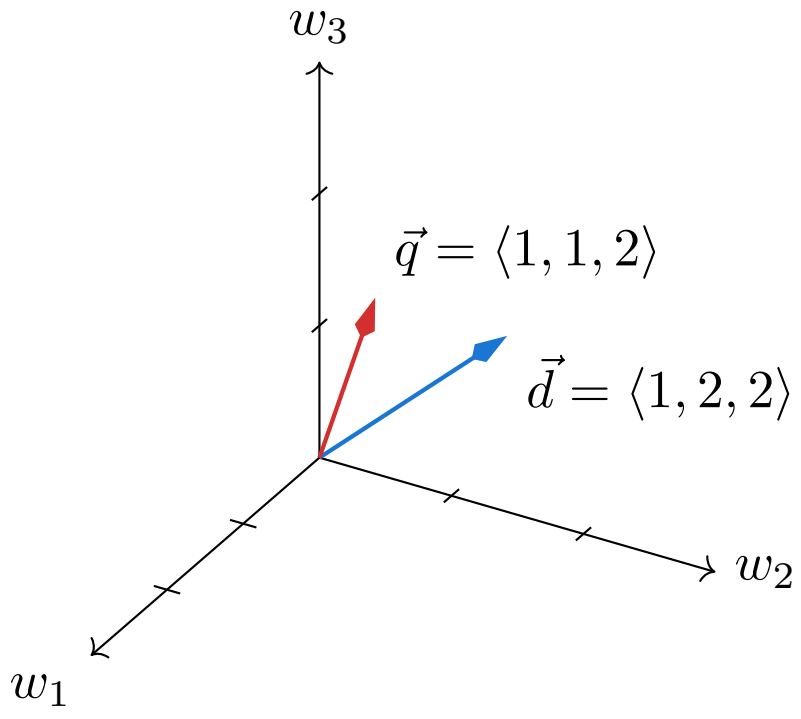
- Term weights are used to compute a degree of similarity between a query and each document.
- The documents are ranked in decreasing order of their degree of similarity.



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model

- The weight $m_{i,j}$ associated with a pair (d_i, w_j) is positive and non-binary.
- The index terms are assumed to be all mutually independent.
- They are represented as unit vectors of a $|\mathcal{V}|$-dimensional space.

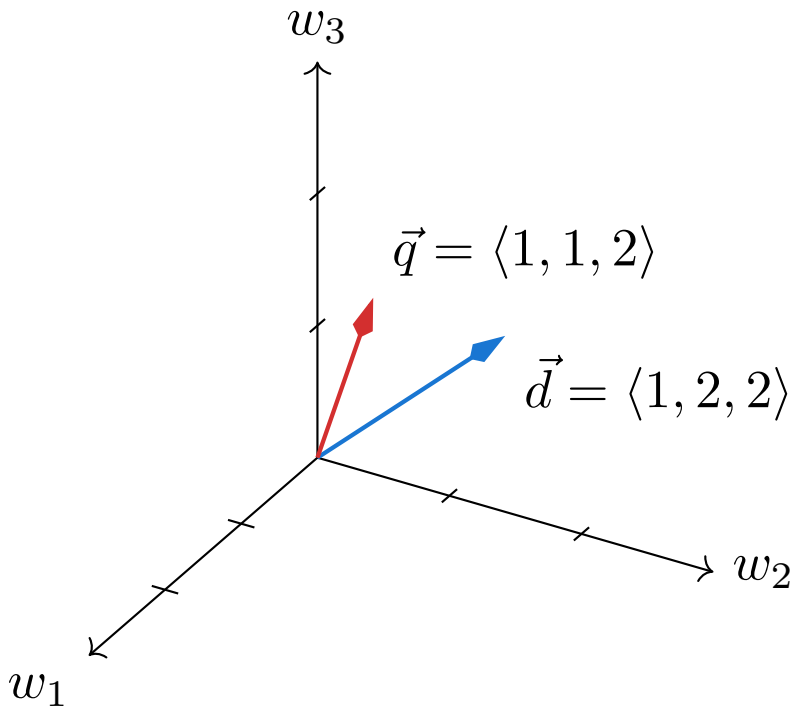


Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model

- They are represented as unit vectors of a $|\mathcal{V}|$-dimensonsal space.
- The representations of document d and query q are $|\mathcal{V}|$-dimensional vectors given by:

$$\vec{d}_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,|\mathcal{V}|} \rangle$$
$$\vec{q} = \langle m_{q,1}, m_{q,2}, \ldots, m_{q,|\mathcal{V}|} \rangle$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

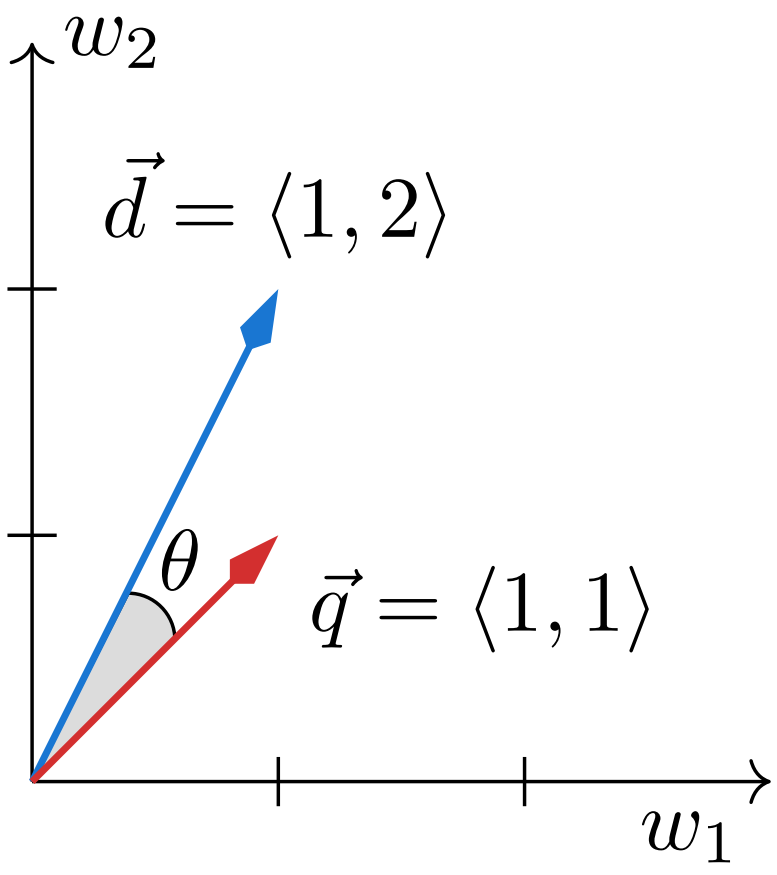
The Vector Space Model

- Similarity between a document and query $\text{sim}(\vec{d}, \vec{q})$ is equated to its cosine-similarity:

$$\vec{d}_i \cdot \vec{q} = |\vec{d}_i| \cdot |\vec{q}| \cdot \cos(\theta)$$

$$\cos(\theta) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}_i| \cdot |\vec{q}|}$$

$$\cos(\theta) = \frac{\sum_{j=1}^{|\mathcal{V}|} m_{i,j} \cdot m_{q,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{|\mathcal{V}|} m_{i,q}^2}}$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

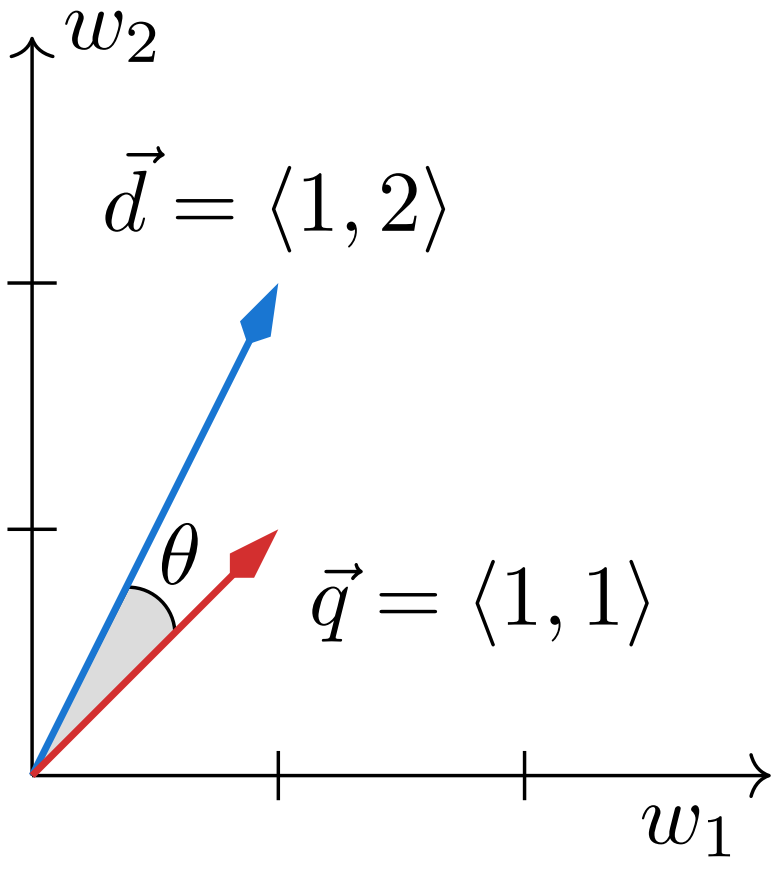
The Vector Space Model

- Weights in the Vector model are basically TF-IDF weights.

$$m_{i,j} = \left[1 + \log\left[\mathrm{tf}_{i,j}\right]\right] \times \log\left[1 + \frac{N}{\mathrm{df}_j}\right]$$

$$m_{q,j} = \left[1 + \log\left[\mathrm{tf}_{q,j}\right]\right] \times \log\left[1 + \frac{N}{\mathrm{df}_j}\right]$$

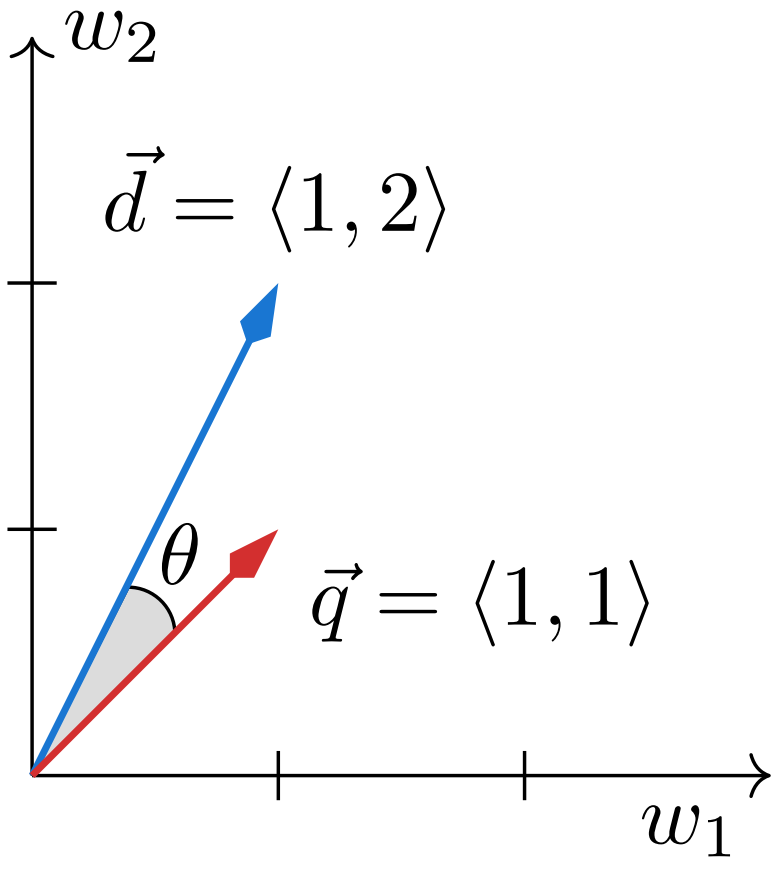
- These equations should only be applied for values of term frequency greater than zero
- If the term frequency is zero, the respective weight is also zero.



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model

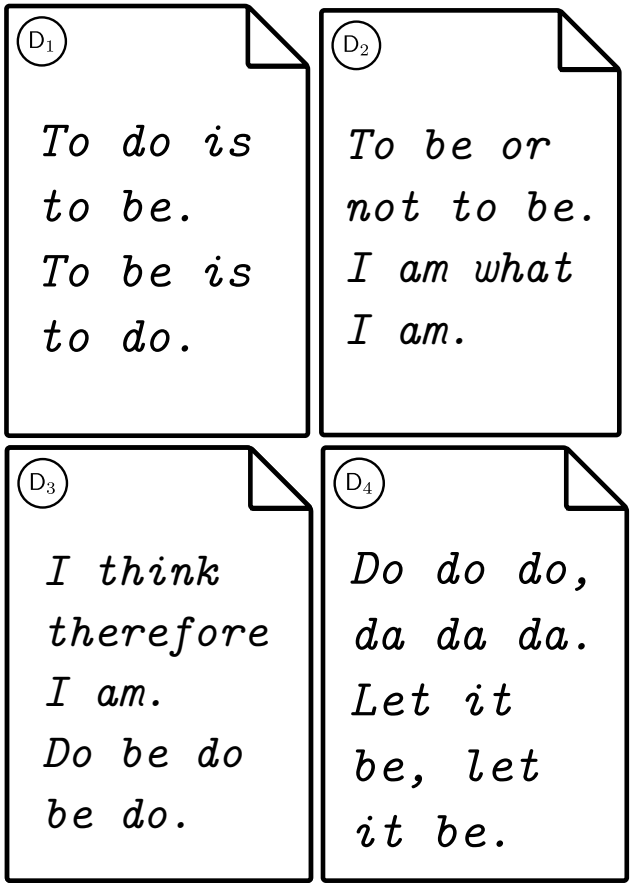
- Term-weighting improves quality of the answer set.
- Partial matching allows retrieval of documents that approximate the query conditions.
- Cosine ranking formula sorts documents according to a degree of similarity to the query.
- Document length normalization is naturally built-in into the ranking.
- Disadvantage: it assumes independence of index terms.



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model Example

$$\text{TF Weighting} : m_{i,j} = 1 + \log\left[\text{tf}_{d,t}\right]$$



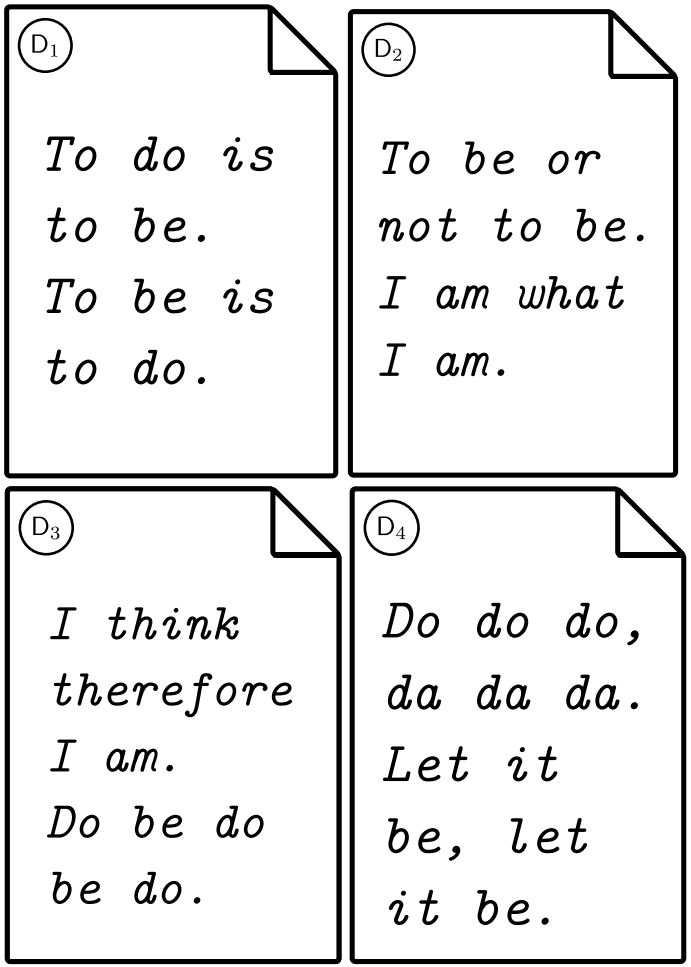
Vocabulary	
1	to
2	do
3	is
4	be
5	or
6	not
7	I
8	am
9	what
10	think
11	therefore
12	da
13	let
14	it

$tf_{1,i}$	$tf_{2,i}$	$tf_{3,i}$	$tf_{4,i}$
3	2	-	-
2	-	2.585	2.585
2	-	-	-
2	2	2	2
-	1	-	-
-	1	-	-
-	2	2	-
-	2	1	-
-	1	-	-
-	-	1	-
-	-	1	-
-	-	-	2.585
-	-	-	2
-	-	-	2

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model Example

$$\text{IDF Weighting} : \text{idf}_j \equiv \log\left[N/\text{df}_j\right]$$

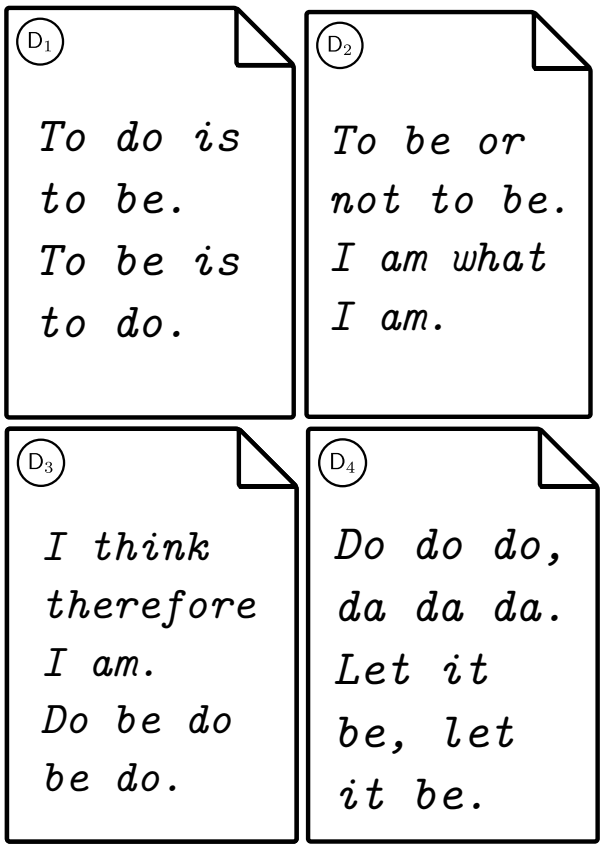


	term	df_j	$idf_i = \log(N/df_j)$
1	to	2	1
2	do	3	0.415
3	is	1	2
4	be	4	0
5	or	1	2
6	not	1	2
7	I	2	1
8	am	2	1
9	what	1	2
10	think	1	2
11	therefore	1	2
12	da	1	2
13	let	1	2
14	it	1	2

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model Example

$$m_{d,t} = \left[1 + \log\left[tf_{d,t}\right]\right] \cdot \left[\log\left[\frac{N}{\mathrm{df}_t}\right]\right].$$

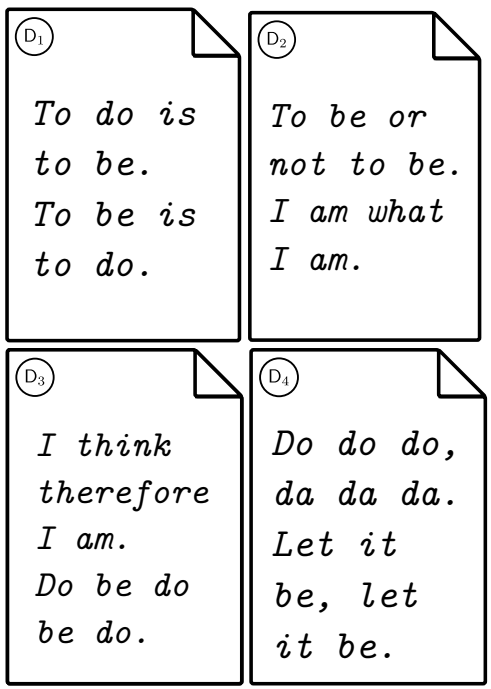


		d_1	d_2	d_3	d_4
1	to	3	2	-	-
2	do	0.830	-	1.073	1.073
3	is	4	-	-	-
4	be	-	-	-	-
5	or	-	2	-	-
6	not	-	2	-	-
7	I	-	2	2	-
8	am	-	2	1	-
9	what	-	2	-	-
10	think	-	-	2	-
11	therefore	-	-	2	-
12	da	-	-	-	5.170
13	let	-	-	-	4
14	it	-	-	-	4

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model Example

$$\hat{d} = \frac{\vec{d}}{|\vec{d}|} = \frac{\vec{d}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_j^2}}$$



	d_1	d_2	d_3	d_4
size in bytes	34	37	41	43
number of words	10	11	10	12
vector norm	5.068	4.899	3.762	7.738

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

The Vector Space Model Example

- Query: *to do*.

$$\cos(\theta) = \frac{\sum_{j=1}^{|\mathcal{V}|} m_{q,j} \cdot m_{i,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{i,j}^2} \cdot \sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{q,j}^2}} \propto \frac{\sum_{j=1}^{|\mathcal{V}|} m_{q,j} \cdot m_{i,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{i,j}^2}}$$

doc	rank computation	rank
d_1	$\frac{1*3+0.415*0.830}{5.068}$	0.660
d_2	$\frac{1*2+0.415*0}{4.899}$	0.408
d_3	$\frac{1*0+0.415*1.073}{3.762}$	0.118
d_4	$\frac{1*0+0.415*1.073}{7.738}$	0.058

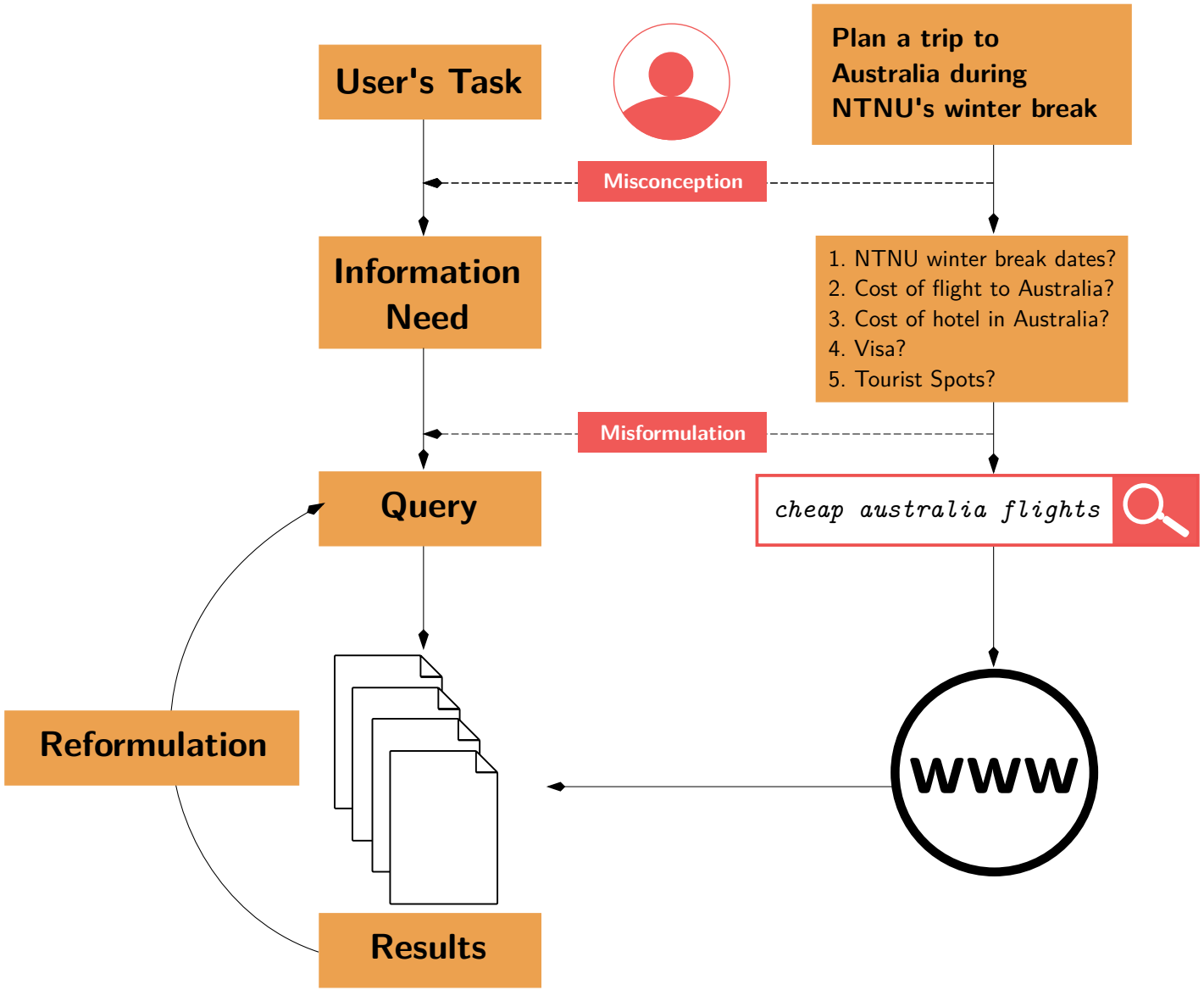
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Summary — The IR Problem

- An IR Model can be defined by a quadruple $\langle \mathcal{D}, \mathcal{Q}, \mathcal{F}, R(q, d) \rangle$, where
  - $\mathcal{D} = \{d_1, d_2, \ldots, d_N\} \equiv$ document collection,
  - $\mathcal{Q} = \{q_1, q_2, \ldots, q_M\} \equiv$ query collection reflecting user's information needs,
  - $\mathcal{F} \equiv$ framework for modeling documents d, queries q, and their relationships.
  - $R(q, d) \equiv$ ranking function.

Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Summary — The IR Problem



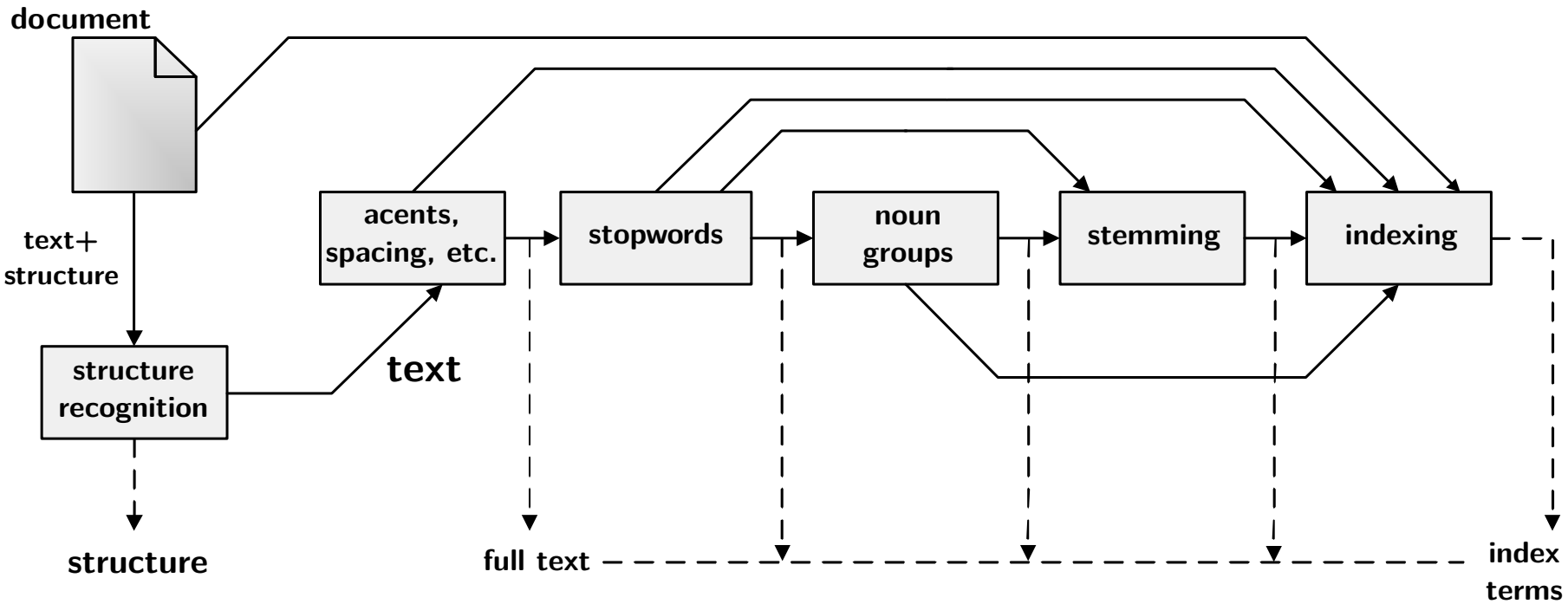
Manning et al., "Introduction to Information Retrieval," Cambridge University Press.

Summary — Modeling Documents

	w_1	w_2	w_3	w_4	w_5	$\cdots$	$w_{\lvert\mathcal{V}\rvert}$
d_1	1	0	1	1	0	$\cdots$	1
d_2	1	1	0	0	1	$\cdots$	0
d_3	0	0	0	1	0	$\cdots$	0
d_4	0	1	0	0	1	$\cdots$	1
d_5	0	0	1	0	0	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
d_N	1	1	0	1	1	$\cdots$	0

Summary — Modeling Documents

- Logical view of a document: from full text to a set of index terms



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Summary — Boolean Retrieval

- Consider a Boolean query: $q = w_1 \wedge (w_2 \vee \neg w_3)$.
- Term vector for $w_1 = \langle 1, 1, 0, 0, 0 \rangle$.
- Term vector for $w_2 = \langle 0, 1, 0, 1, 0 \rangle$.
- Term vector for $w_3 = \langle 1, 0, 0, 0, 1 \rangle$.

	1	2	3	4	5
w_3	1	0	0	0	1
$\neg w_3$	0	1	1	1	0

	1	2	3	4	5
$\neg w_3$	0	1	1	1	0
w_2	0	1	0	1	0
OR	0	1	1	1	0

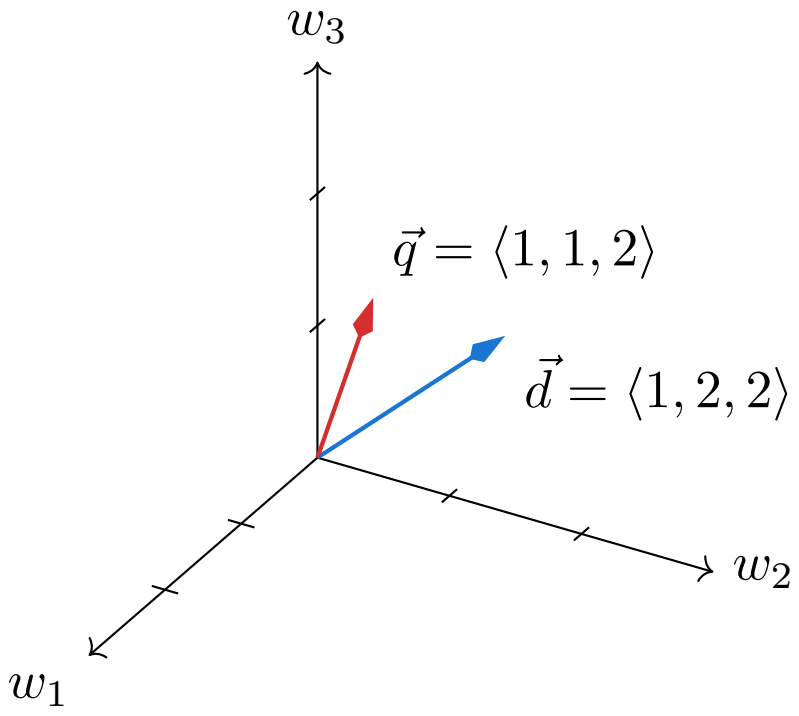
	1	2	3	4	5
$w_2 \vee \neg w_3$	0	1	1	1	0
w_1	1	1	0	0	0
AND	0	1	0	0	0

	w_1	w_2	w_3	w_4	w_5
d_1	*1*	*0*	*1*	*1*	*0*
d_2	*1*	*1*	*0*	*0*	*1*
d_3	*0*	*0*	*0*	*1*	*0*
d_4	*0*	*1*	*0*	*0*	*1*
d_5	*0*	*0*	*1*	*0*	*0*

Summary — The Vector Space Model

- They are represented as unit vectors of a $|\mathcal{V}|$-dimensonsal space.
- The representations of document d and query q are $|\mathcal{V}|$-dimensional vectors given by:

$$\vec{d}_i = \langle m_{i,1}, m_{i,2}, \ldots, m_{i,|\mathcal{V}|} \rangle$$
$$\vec{q} = \langle m_{q,1}, m_{q,2}, \ldots, m_{q,|\mathcal{V}|} \rangle$$



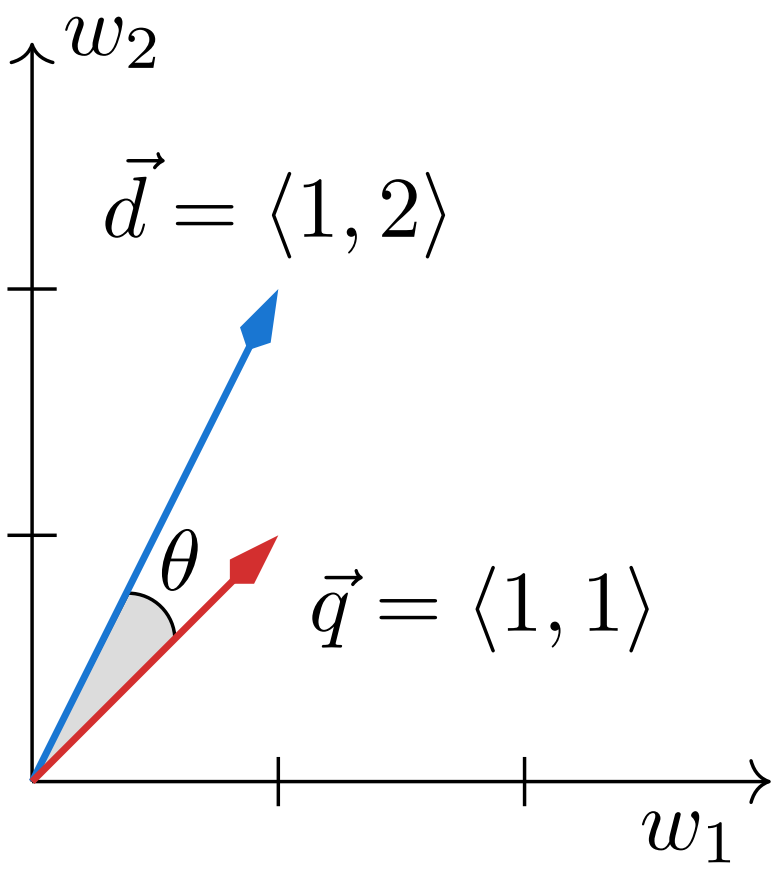
Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

Summary — The Vector Space Model

- Similarity between a document and query $\text{sim}(\vec{d}, \vec{q})$ is equated to its cosine-similarity:

$$\cos(\theta) = \frac{\vec{d}_i \cdot \vec{q}}{|\vec{d}| \cdot |\vec{q}|}$$

$$\cos(\theta) = \frac{\sum_{j=1}^{|\mathcal{V}|} m_{i,j} \cdot m_{q,j}}{\sqrt{\sum_{j=1}^{|\mathcal{V}|} m_{i,j}^2} \cdot \sqrt{\sum_{i=1}^{|\mathcal{V}|} m_{i,q}^2}}$$



Baeza-Yates and Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.