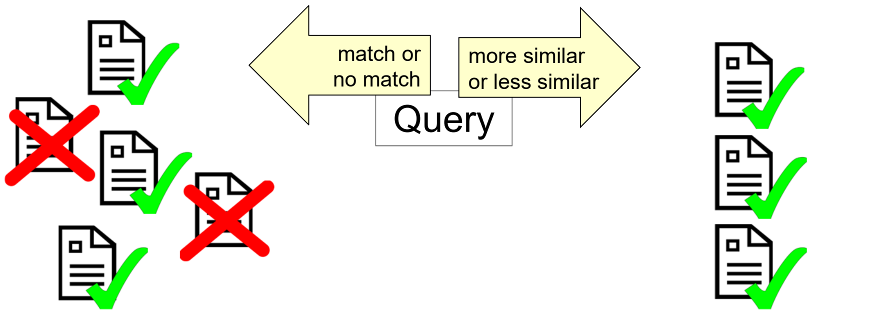
Basics of Information Retrieval

- Today
  - Vector-space retrieval model – the most common retrieval model
    - Similarity between documents
    - Measuring importance of a word
  - Brief mention of the probabilistic retrieval model
  - Evaluation of retrieval
    - Retrieval performance measures

Boolean Retrieval is not Sufficient



Boolean retrieval **CAN** do this

Boolean retrieval **CANNOT** do this

In the vector-space retrieval model, we calculate similarity as a number. A bigger number means more similarity, which enables relevance ranking

Words Define the Content

The John Ripley Forbes **Big Trees Forest** Preserve is a beautiful and unique 30-acre Fulton County **Tree**, Plant and Wildlife Sanctuary

The **Forest** is open sunrise to sunset, se... week. There is ample free parking. Dogs ... welcomed in the **Forest**. The **Forest** is a smoke-free environment. Donations are appreciated and used wisely to promote the **Forest** mission.

The **Forest** is located about 4 miles north of I-285 in the northern metro – Atlanta community of Sandy Springs, Georgia, next to the North Fulton County Government Annex building at 7645 Roswell Road.This previously threatened urban **forest**, one of the last in the highly developed area of Sandy Springs, was assembled in three purchases beginnin... Southeast Land Preserva... Fulton County and the Sta... Georgia owns 10 acres a... owns 20 acres that was d... 2006.

Query

Document

Like no other place on Earth, the **Giant Forest** in Sequoia National Park is alive with mystery and wonder. We see it ... they know what to expect here. ...e **trees**, but just as often, they ...ch, much larger.

At the heart of the park, in the shade of towering sequoias and redwood groves, the **Giant Forest** is home to half of the Earth's largest and longest-living **trees**. Named in 1875 by John Muir, the **forest** is a stand of more than 8000 colossal sequoia **trees** – many standing just as Muir found them.

The undisputed King of the **Forest**, the General Sherman **tree** is not only the largest living **tree** in the world, but the largest living organism, by volume, on the planet. A **giant** ...um), General Sherman

	t_1 (big)	t_2 (tree)	t_3 (forest)	t_4 (giant)	…	t_n
q	1	1	1	0	…	q_n
d	0	1	1	1	…	d_n

Vector-space retrieval model uses algebraic operations between vectors to calculate similarity between documents

- Where are the vectors?
- Which algebraic operations?

Document Vector (a.k.a. Term Vector)

- d_a, d_b, … , d_z are documents, q is also a document
- t_1, t_2, … , t_n are terms present in those documents

n dimensions of the vector space

vectors in the *n*-dimensional space

	t_1 (big)	t_2 (tree)	t_3 (forest)	t_4 (giant)	…	t_n
q	1	1	1	0	…	q_n
d_a	0	1	1	1	…	d_n
d_b	d_{b1}	d_{b2}	d_{b3}	d_{b4}	…	d_{bn}
d_c	d_{c1}	d_{c2}	d_{c3}	d_{c4}	…	d_{cn}
…	…	…	…	…	…	…
d_z	d_{z1}	d_{z2}	d_{z3}	d_{z4}	…	d_{zn}

table cells contain vector coordinates

- $q = <q_1, q_2, \ldots, q_n>$ and $d = <d_1, d_2, \ldots, d_n>$ are document vectors. q_i and d_i values are 0 or 1

Document Vector

dimension

dimension

vector coordinate

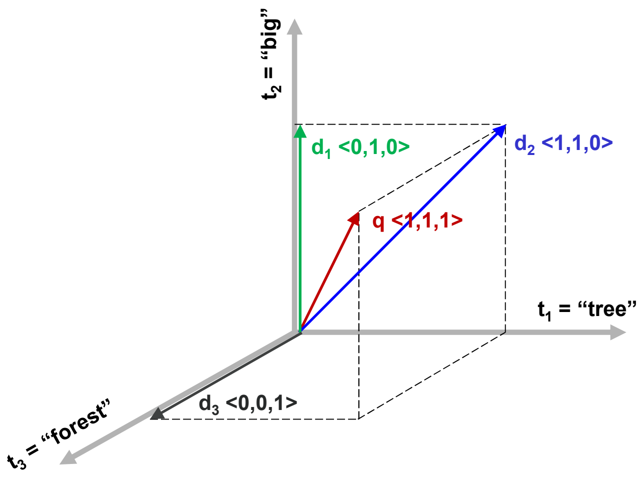
document vector

document vector

	t_1 (big)	t_2 (tree)	t_3 (forest)	t_4 (giant)	…	t_n
q	1	1	1	0	…	q_n
d_a	0	1	1	1	…	d_n
d_b	d_{b1}	d_{b2}	d_{b3}	d_{b4}	…	d_{bn}
d_c	d_{c1}	d_{c2}	d_{c3}	d_{c4}	…	d_{cn}
…	…	…	…	…	…	…
d_z	[illegible]	d_{z2}	d_{z3}	d_{z4}	…	d_{zn}

vector coordinate

Document Vector



	t_1	t_2	t_3
q	1	1	1
d_1	0	1	0
d_2	1	1	0
d_3	0	0	1

Document is a Bag of Words

- Document is a vector, each term-dimension is independent, therefore…
  - … a document is a *bag of words* where most of the information about the document structure is lost
- Vector-space retrieval model works best if both the query and the document have more than a few words
  - Best for document clustering and categorization, and finding similar documents, not for keyword-based search

Similarity as the Number of Common Terms

- Straightforward document similarity – count the number of terms that *q* and *d* have in common

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1
d	1	1	1	0	1

$$sim = 1\cdot 1 + 1\cdot 1 + 0\cdot 1 + 0\cdot 0 + 1\cdot 1$$
$$sim = q_1\cdot d_1 + q_2\cdot d_2 + q_3\cdot d_3 + q_4\cdot d_4 + q_5\cdot d_5$$

- Scalar product of the query vector and the document vector

$$q \bullet d = \sum_{i=1}^{n} q_i \cdot d_i$$

Examples of Scalar Product Similarity

- Similarity measure is a number:

$$q \bullet d = \sum_{i=1}^{5} q_i \cdot d_i$$

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

	t_1	t_2	t_3	t_4	t_5	$q \bullet d$
d_a	1	0	0	0	1	2
d_b	0	1	1	1	1	2
d_c	0	0	1	1	0	0
d_d	1	1	1	0	1	3

- Relevance:
  1. d_d
  2. d_a or d_b
  3. d_c - *non-relevant*

Similarity and Document Length

- Measuring only the scalar product has disadvantages:
  - Longer documents are more likely to be relevant because they are more likely to contain matching terms
  - If two documents have the same score, we would prefer the shorter one because it is more focused on the information need
- Conclusion: The length of the document should be integrated in the similarity score

Normalized Similarity

- Length (weight) of a document – count terms:

$$d_w = d_1 + d_2 + \ldots + d_n = \sum_{i=1}^{n} d_i$$

- Normalized document vector:

$$\frac{d}{d_w} = <\frac{d_1}{d_w}, \frac{d_2}{d_w}, \ldots, \frac{d_n}{d_w}>$$

- Normalized similarity is normalized scalar product:

$$q \bullet \frac{d}{d_w} = \sum_{i=1}^{n} q_i \cdot \frac{1}{d_w} \cdot d_i = \frac{1}{d_w} \cdot \sum_{i=1}^{n} q_i \cdot d_i = \frac{q \bullet d}{d_w}$$

Examples of Normalized Similarity

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

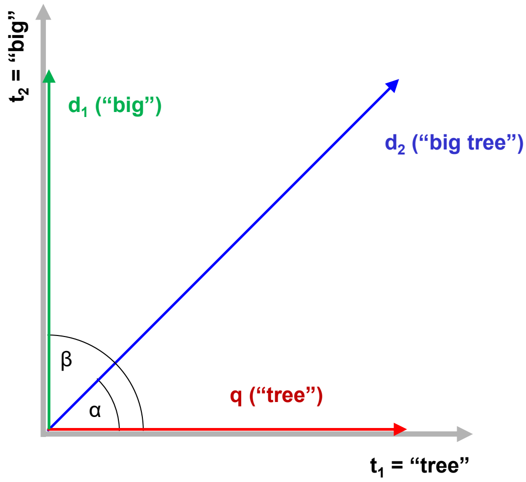
	t_1	t_2	t_3	t_4	t_5	$q \bullet d$	$(q \bullet d)/d_w$
d_a	1	0	0	0	1	2	2/2 = 1
d_b	0	1	1	1	1	2	2/4 = 0.5
d_c	0	0	1	1	0	0	0/2 = 0
d_d	1	1	1	0	1	3	3/4 = 0.75

- Relevance:
  1. d_a
  2. d_d
  3. d_b
  4. d_c

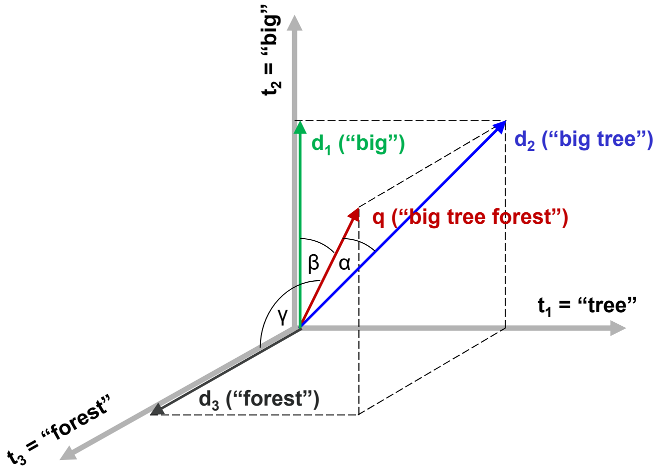
Normalized Similarity

- Problem solved:
  - Shorter and presumably more focused documents receive higher normalized similarity score than longer documents with the same matching terms
- Problem acquired:
  - Shorter documents are generally preferred over longer ones because of d_w in $\frac{q \bullet d}{d_w}$

Similarity as the Angle between a Query and a Document

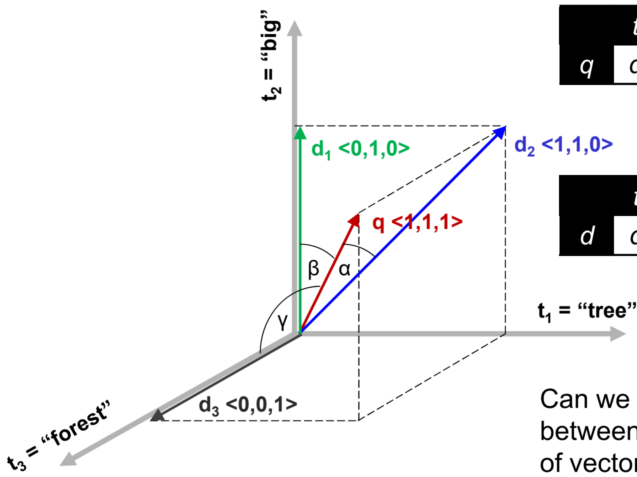


Similarity as the Angle between a Query and a Document



$\alpha < \beta$
$\beta = \gamma$

Similarity as the Angle between a Query and a Document



	t_1	t_2	t_3	…	t_n
q	q_1	q_2	q_3	…	q_n

	t_1	t_2	t_3	…	t_n
d	d_1	d_2	d_3	…	d_n

Can we express the angle between q and d as a function of vector coordinates $f(q_1, q_2, \dots, q_n, d_1, d_2, \dots, d_n)$?

Angle Expressed by Vector Coordinates

- Original definition of scalar product:

$$q \bullet d = |q| \cdot |d| \cdot \cos \alpha$$

- Angle expressed by scalar product and vector length

$$\cos \alpha = \frac{q \bullet d}{|q| \cdot |d|} \qquad q \bullet d = \sum_{i=1}^{n} q_i \cdot d_i$$

$$|q| = \sqrt{\sum_{i=1}^{n} q_i^2} \qquad |d| = \sqrt{\sum_{i=1}^{n} d_i^2}$$

cos α expressed by vector coordinates q_i and d_i

Cosine Similarity

- Query vector q and document vector d, both of length n. Cosine similarity between them is defined as:

$$sim(q,d) = \cos \alpha = \frac{q \bullet d}{|q| \cdot |d|}$$

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

- If α is 0, similarity is 1.
- Orthogonal vectors have similarity 0.

Examples of Cosine Similarity

orthogonal vectors

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

	t_1	t_2	t_3	t_4	t_5	$q \bullet d$	$(q \bullet d) / d_w$	$sim(q, d)$
d_a	1	0	0	0	1	2	$2/2 = 1$	0.816
d_b	0	1	1	1	1	2	$2/4 = 0.5$	0.577
d_c	0	0	1	1	0	0	$0/2 = 0$	0
d_d	1	1	1	0	1	3	$3/4 = 0.75$	0.866

- Relevance:

  1. d_d 2. d_a 3. d_b 4. d_c

Today

- Vector-space retrieval model
  - \+ Similarity between documents
  - **Measuring importance of a word**
- Brief mention of the probabilistic retrieval model
- Evaluation of retrieval
  - Retrieval performance measures

Binary Coordinates

- So far document vectors had binary coordinates:
  1 : term occurs in the document
  0 : term does not occur in the document

	t_1	t_2	t_3	t_4	t_5
q	1	1	0	0	1

- Binary coordinates have a shortcomings – all terms in the document and the document collection are considered being equally important

Term Weights

- Not all words in text are equally important

Johnson has repeatedly said he will take **UK** out of **EU** in October 31.
- Johnson UK EU → Brexit means Brexit
- has repeatedly said he will take out of in October 31 → not much sense

- Instead of 0 and 1, we use numeric term weights as document vector coordinates. A higher weight means a more important term:

	UK	EU	t_3	t_4	…
d_a	1	0	1	w_{14}	…
d_b	1.42	4	0	w_{24}	…
d_c	0	0.16	w_{33}	w_{34}	…
d_d	1.76	3.73	w_{43}	0	…
…	…	…	…		

How do we calculate it?

Term Frequency: tf-score

- tf_{ij} is the frequency of the j^{th} term in the i^{th} document
- In plain English – count the terms in a document

Johnson said there was "bags of time" for the **EU** to compromise on the Irish border backstop plan before the **Brexit** deadline of 31 October.

He also warned MPs not to oppose **Brexit**, and to respect the 2016 referendum result.

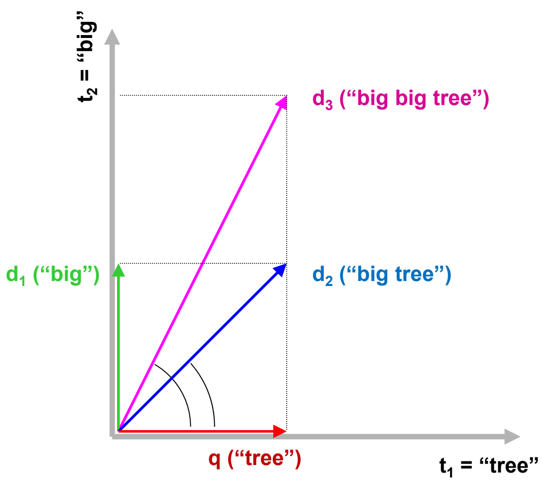
The **EU** has said repeatedly the backstop arrangements cannot be changed.

Johnson said the policy - designed to guarantee there will not be a hard Irish border after **Brexit** - would turn the **UK** into a "satellite state" of the **EU** if it came into effect.

(from BBC news)

$tf_{Brexit} = 3$ $tf_{EU} = 3$ $tf_{Johnson} = 2$ $tf_{UK} = 1$ $tf_{Asia} = 0$

Term Frequency in Vector Space



Common Words

- Only term frequency does not make words important

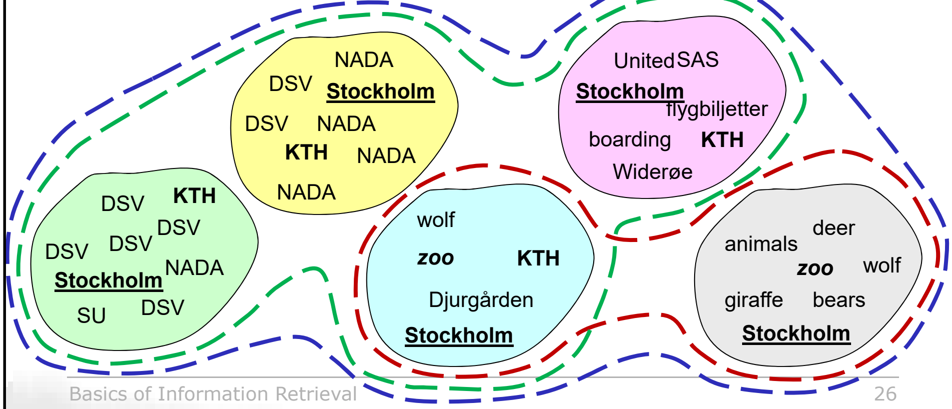
We **at** DSV **are the** department **of the** IT University **that** focuses **on** bridging **the** gap, between **on the** one hand information technology, **and on the** other hand **the** social sciences, **the** behavioral sciences **as** well **as the** humanities.

and, are, as (2x), at, between, behavioral, bridging, department, dsv, focuses, gap, hand (2x), humanities, information, it, of, on (3x), one, other, sciences (2x), social, technology, that, the (8x), university, we, well

Measuring Uniqueness of a Term in the Collection

"zoo" is in 2 of 5 documents
"KTH" is in 4 of 5 documents
"Stockholm" is in 5 of 5 documents

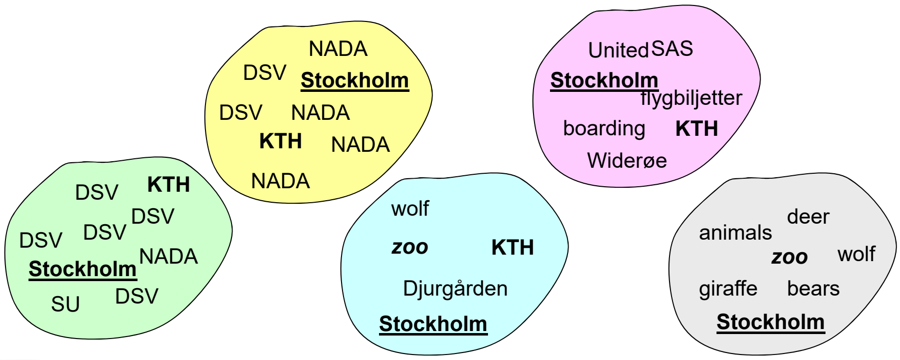
Which words distinguish a smaller subset of the documents?



Measuring Uniqueness of a Term in the Collection

"zoo" is in 2 of 5 documents
"KTH" is in 4 of 5 documents
"Stockholm" is in 5 of 5 documents

"zoo": 5/2 = 2.5
"KTH": 5/4 = 1.25
"Stockholm": 5/5 = 1

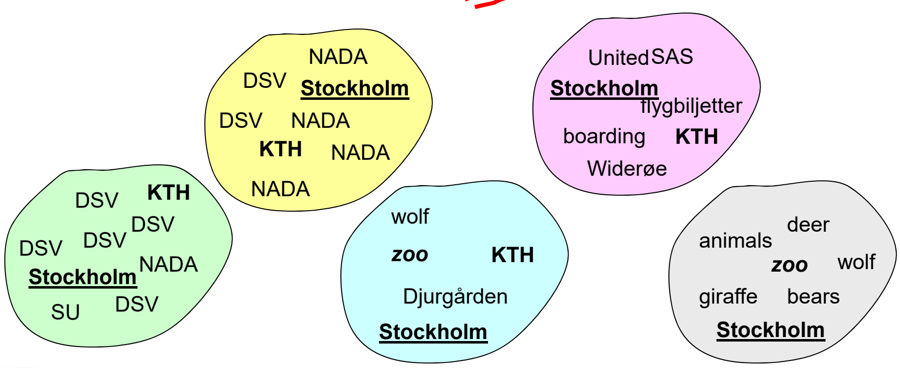


Basics of Information Retrieval 27

Inverse Document Frequency

idf_{zoo} = log(5/2) = log(2.5)
idf_{KTH} = log(5/4) = log(1.25)
$idf_{Stockholm}$ = log(5/5) = log(1) = 0

$idf_{zoo} = \log_2(5/2) \approx 1.32$
$idf_{KTH} = \log_2(5/4) \approx 0.32$



Basics of Information Retrieval 28

Inverse Document Frequency

- Unique words that appear in few documents make these few documents more related
- idf-score of the j^{th} term measures the uniqueness of the j^{th} term in the collection of documents:

$$idf_j = \log(\frac{N}{n_j})$$

  - N is the total number of documents in the collection; n_j is the number of documents that contain the j^{th} term
  - the logarithm makes idf-score ≈ 0 if $n_j \approx N$; evens out differences between large $\frac{N}{n_j}$ values

Term Weight: tf.idf-score

- tf.idf-score is the **term weight** of the j^{th} term in the i^{th} document :

$$w_{ij} = tf_{ij} \cdot idf_j$$

- tf.idf-score is high if the word is
  - frequent in the document, AND
  - occurs in few documents of the collection
- tf.idf-score is 0 if the word is
  - not present in the document, OR
  - present in all the documents of the collection

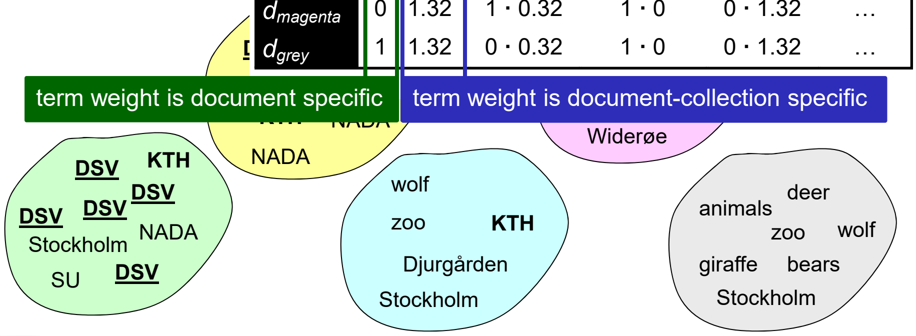
tf.idf-score in Document Vectors

$$w_{ij} = tf_{ij} \bullet \log_2\left(\frac{5}{n_j}\right)$$

	t_1 (zoo)	t_2 (KTH)	t_3 (Stockholm)	t_4 (DSV)	…
d_{green}	0 1.32	1 · 0.32	1 · 0	5 · 1.32	…
d_{cyan}	1 1.32	1 · 0.32	1 · 0	0 · 1.32	…
d_{yellow}	0 1.32	1 · 0.32	1 · 0	2 · 1.32	…
$d_{magenta}$	0 1.32	1 · 0.32	1 · 0	0 · 1.32	…
d_{grey}	1 1.32	0 · 0.32	1 · 0	0 · 1.32	…

term weight is document specific

term weight is document-collection specific



tf.idf-score in Cosine Similarity

	t_1 (zoo)	t_2 (KTH)	t_3 (Stockholm)	t_4 (DSV)	…
d_{green}	0	0.32	0	6.6	…
d_{cyan}	1.32	0.32	0	0	…
d_{yellow}	0	0.32	0	2.64	…
$d_{magenta}$	0	0.32	0	0	…
d_{grey}	1.32	0	0	0	…

$$sim(q,d) = \frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Today

- Vector-space retrieval model
  - Similarity between documents
  - Measuring importance of a word
- **Brief mention of the probabilistic retrieval model**
- Evaluation of retrieval
  - Retrieval performance measures

Short Queries

big tree

Not much of a query vector

Like no other place on Earth, the **Giant Forest** in Sequoia National Park is alive with mystery and wonder. We see it all the time – people think they know what to expect here. They come for the massive **trees**, but just as often, they leave with something much, much larger.

At the heart of the park, in the shade of towering sequoias and redwood groves, the **Giant Forest** is home to half of the Earth's largest and longest-living **trees**. Named in 1875 by John Muir, the **forest** is a stand of more than 8000 colossal sequoia **trees** – many standing just as Muir found them.

The undisputed King of the **Forest**, the General Sherman **tree** is not only the largest living **tree** in the world, but the largest living organism, by volume, on the planet. A **giant** …eum), General Sherman

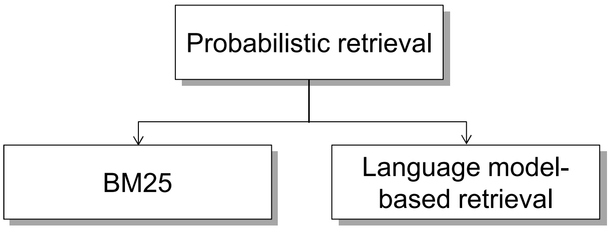
	t_1 (big)	t_2 (tree)	t_3 (forest)	t_4 (giant)	…	t_n
q	1	1	0	0	all zeros	q_n
d	0	1	1	1	…	d_n

One of the best ways to get to know the **Giant Forest** and its **trees** is to visit the **Giant Forest** Museum. Visitors flock to this park institution to:

- Learn how to identify **trees**…

Probabilistic Retrieval Model

- Designed for short queries, 2-3 words are enough



BM25

$$\mathrm{sim}(q,d) = \sum_{t \in q} \mathrm{idf}(t) \cdot \mathrm{tf}(t,d)$$

N docs in the collection

n docs have t

$$\mathrm{idf}(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$$

frequency of t in d

$$\mathrm{tf}(t,d) = \frac{f(t,d) \cdot (k_1 + 1)}{f(t,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avgdl}\right)}$$

length in words

$k_1 \in [1.2, 2.0]$

0.75

average length of a document

Cosine Similarity vs. BM25

- Cosine similarity
  - Works better than BM25 with longer queries, such as news articles, blog posts, etc.
  - Intuitive, easy to explain
- BM25
  - Works better than cosine similarity with short queries, 2-3 keywords
  - Difficult to explain without profound knowledge of probability theory

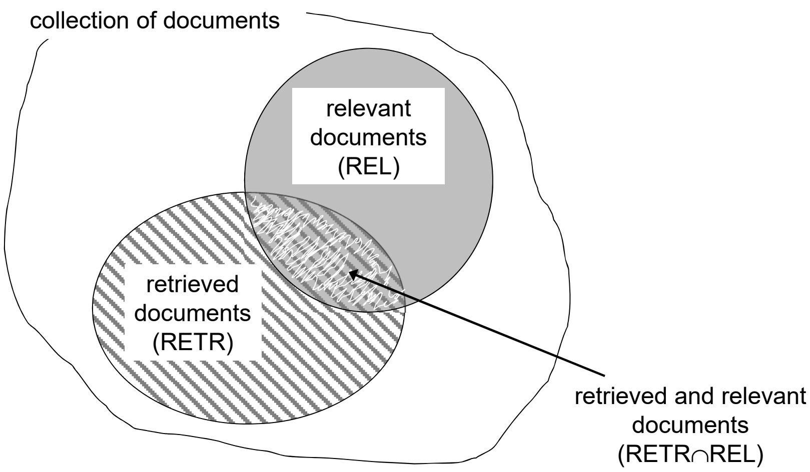
Today

+ Vector-space retrieval model
  + Similarity between documents
  + Measuring importance of a word
+ Brief mention of the probabilistic retrieval model
- Evaluation of retrieval
  - **Retrieval performance measures**

What Do We Evaluate?

- Does the system retrieve all relevant documents?
- Does the system retrieve only relevant documents?
- … and other features of the set of the retrieved documents

Retrieved, Relevant and Retrieved&Relevant Documents

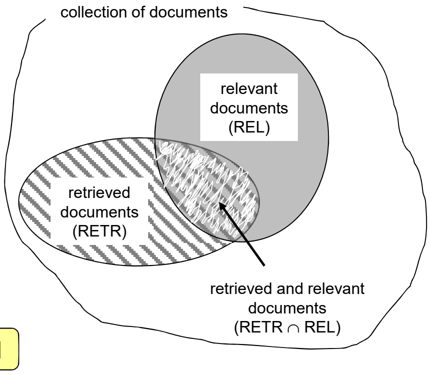


Precision of Retrieval

- **Precision** characterizes the fraction of the retrieved documents that are relevant:

$$P = \frac{|RETR \cap REL|}{|RETR|}$$

Number between 0 and 1

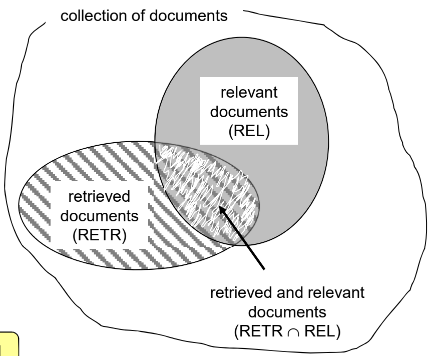


Recall of Retrieval

- **Recall** characterizes the fraction of the relevant documents that have been retrieved:

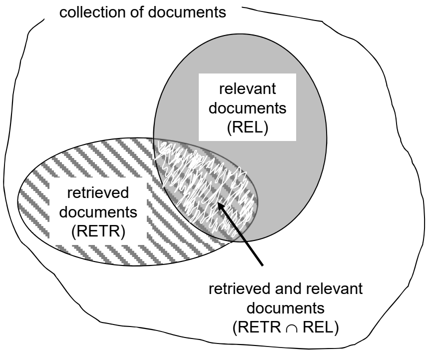
$$R = \frac{|RETR \cap REL|}{|REL|}$$

Number between 0 and 1



Recall of Retrieval

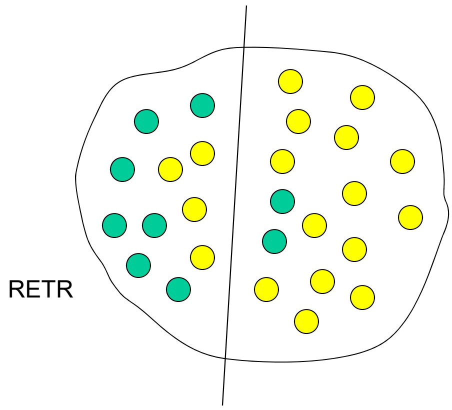
- How do we know which documents are relevant if they are not retrieved?
- Manually verified test collections
  - It is decided by humans whether or not a document is relevant to a query



When are Precision and Recall not Defined?

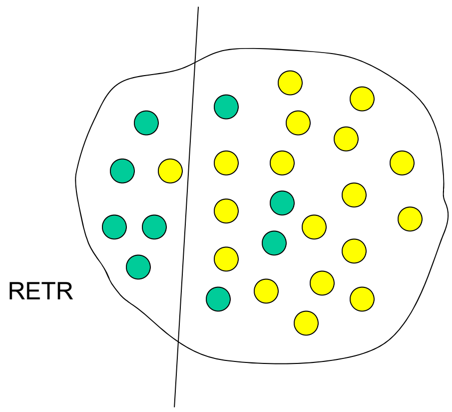
- If recall is 0, then precision is not defined
- Precision and recall are considered only if the query has relevant document in the collection. Otherwise recall is not defined, and with no recall precision is not defined

Example



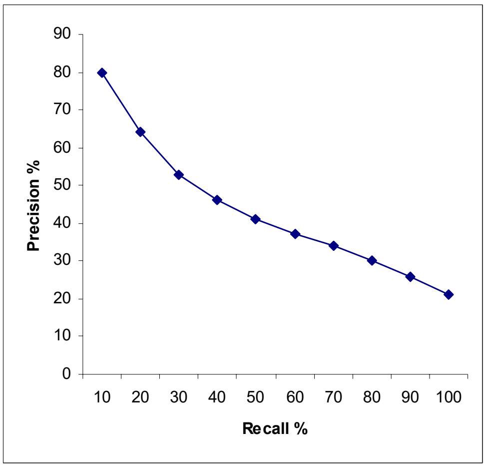
$P = 7 / 11 \approx 64\%$
$R = 7 / 9 \approx 78\%$

Moving the Cut-off Line

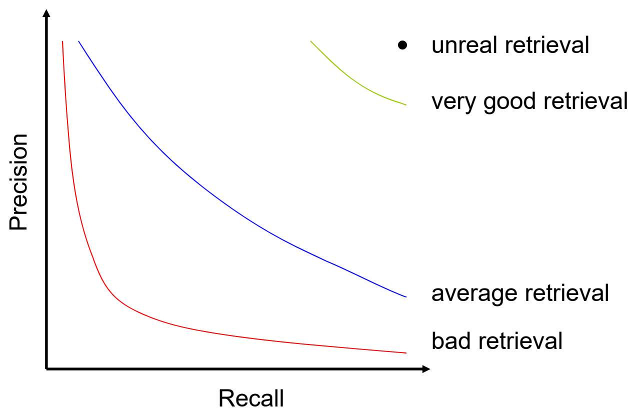


$P = 5 / 6 \approx 83\%$
$R = 5 / 9 \approx 56\%$

Precision-Recall Curve: Trade-off between *P* and *R* Values by Moving the Cut-off Line

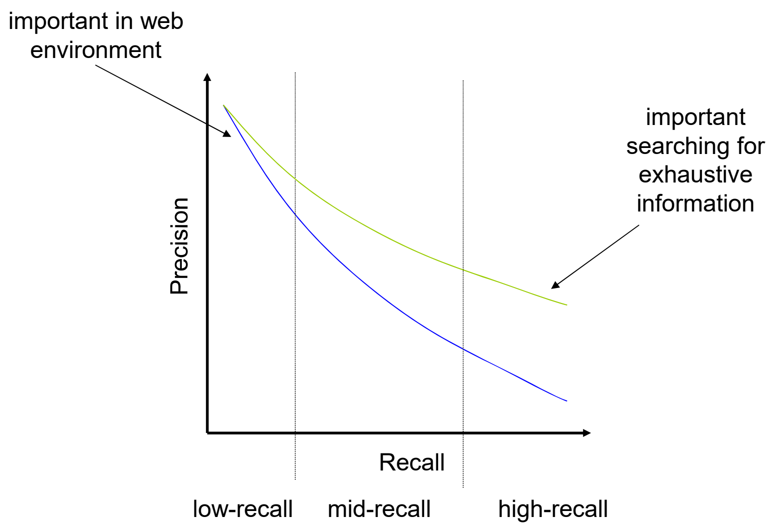


Quality of Retrieval and Precision-Recall Curves



A better retrieval algorithm has better trade-off between precision and recall

Precision-Recall Regions



Today

+ Vector-space retrieval model
  + Similarity between documents
  + Measuring importance of a word
+ Brief mention of the probabilistic retrieval model
+ Evaluation of retrieval
  + Retrieval performance measures

cosine similarity

tf.idf-score

BM25

precision and recall