https://vvtesh.sarahah.com/

Information Retrieval

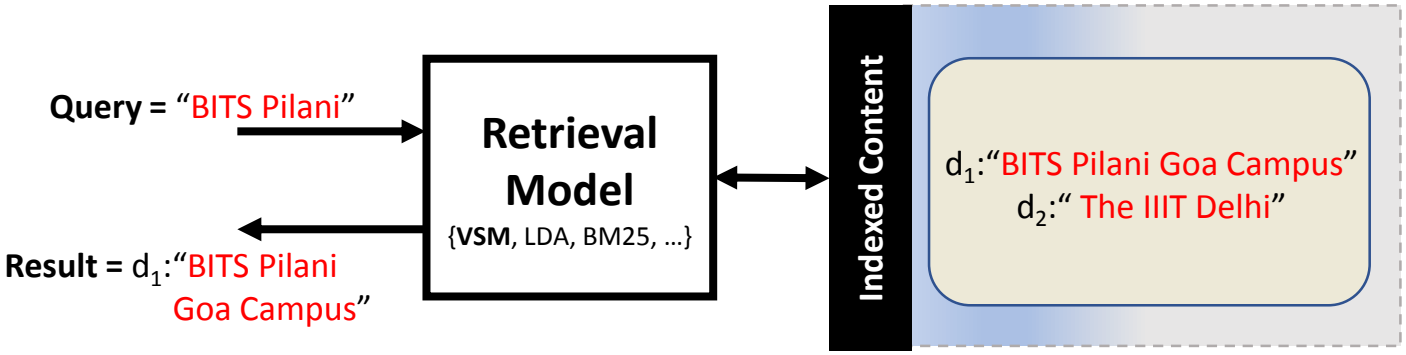
Venkatesh Vinayakarao

Term: Aug – Dec, 2018
Indian Institute of Information Technology, Sri City

The world is one big data problem.
– Andrew McAffee.

Term Weights

Not Every Word is Important!



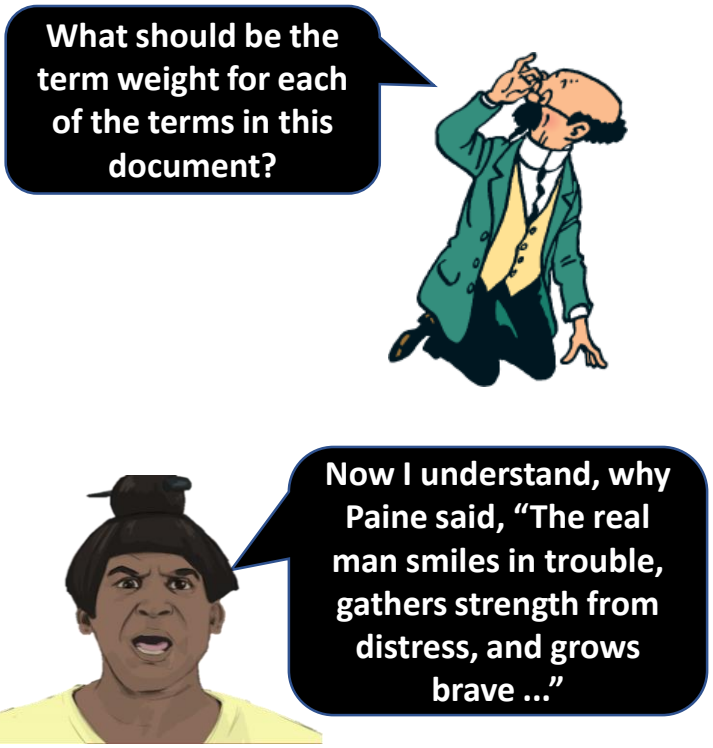
Let us add Term Weights

	BITS	the (* 0)	Pilani	Goa	Campus	IIIT	Delhi
q	1	1 * 0 = 0	1	0	0	0	0
d_1	1	0 * 0 = 0	1	1	1	0	0
d_2	0	1 * 0 = 0	0	0	0	1	1

$sim(q,d_1) = 0.71$

$sim(q,d_2) = 0$

Term Weighting with Inverse Document Frequency



Indexed Content

d_1: *"An inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely."*

d_2: *"A high weight in tf–idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms "*

Inverse Document Frequency

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

where N = |D| = Total no. of documents.

$$idf(\text{"the"}, \{\mathbf{d_1}, \mathbf{d_2}\}) = \log \frac{2}{2} = 0$$

$$idf(\text{"batsmen"}, \{\mathbf{d_1}, \mathbf{d_2}\}) = \log \frac{2}{1} = 0.3$$

But, do you see any problem? Clue… divide by zero.

Indexed Content

d_1: *"An inverse document frequency factor is incorporated which diminishes **the** weight of terms that occur very frequently in **the** document set and increases **the** weight of terms that occur rarely."*

d_2: *"Sachin Ramesh Tendulkar is a former Indian international cricketer and a former captain of **the** Indian national team, regarded as one of **the** greatest batsmen of all time."*

Quiz

- Consider a document containing 100 words wherein the word *cat* appears 3 times.
  - The term frequency (i.e., tf) for *cat* is _____

Quiz

- Consider a document containing 100 words wherein the word *cat* appears 3 times.
  - The term frequency (i.e., tf) for *cat* is 3

*Assumption: No Length Normalization.

Quiz

- Consider a document containing 100 words wherein the word *cat* appears 3 times.
- Now, assume we have 10 million documents and the word *cat* appears in one thousand of these.
- Then, the inverse document frequency (i.e., idf) is ____

Quiz

- Consider a document containing 100 words wherein the word *cat* appears 3 times.
- Now, assume we have 10 million documents and the word *cat* appears in one thousand of these.
- Then, the inverse document frequency (i.e., idf) is log(10,000,000 / 1,000) = 4

Quiz

- Say, D is the set of all documents in our collection.
- |D| = 1,000,000. Find the idf values for each term.

Term	df	idf
Calpurnia	1	
animal	100	
Sunday	1000	
fly	10,000	
under	100,000	
the	1,000,000	

Quiz

- Say, D is the set of all documents in our collection.
- |D| = 1,000,000. Find the idf values for each term.

Term	df	idf
Calpurnia	1	6
animal	100	4
Sunday	1000	3
fly	10,000	2
under	100,000	1
the	1,000,000	0

Symmetric Relation

- Examples of symmetric relationship
  - a is married to b
  - a = b

A relation R is symmetric if $(a,b) \in R$ implies that $(b,a) \in R$.

- Examples of anti-symmetric relationship
  - a < b does not imply b < a. So, <(a,b) is anti-symmetric.

A relation is anti-symmetric if (a,b)∈R and (b,a)∈R only when a=b.

Neither Symmetric Nor Anti-S…

Can a relation be neither symmetric nor anti-symmetric?

a <= b

Quiz: Compute IDF Values

- Given a document with the terms A, B and C with the following frequencies A: 3, B: 2, C: 1. The document belongs to a collection of 10,000 docs. The document frequencies are: A: 50, B:1300, C:250. Compute the IDF values.

Quiz: Compute IDF Values

- Given a document with the terms A, B and C with the following frequencies A: 3, B: 2, C: 1. The document belongs to a collection of 10,000 docs. The document frequencies are: A: 50, B:1300, C:250. Compute the IDF values.

A idf = log(10000/50) = 5.3;
B idf = log(10000/1300) = 2.0;
C idf = log(10000/250) = 3.7

Source: http://disi.unitn.it/~bernardi/Courses/DL/Slides_11_12/measures.pdf

Revision: L^infty-Norm

- A general vector norm |x| is
  - |x| > 0 when x!=0 and |x|=0 iff x=0.
  - |kx|=|k||x| for any scalar k.
  - |x+y|<=|x|+|y|.
- L^infty-Norm

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad |\mathbf{x}|_\infty = \max_i |x_i|.$$

Read http://mathworld.wolfram.com/VectorNorm.html

Revision: Norms

Compute for the vector x = (1,2,3)

Name	Symbol	Value
L^1-norm	$\|x\|_1$	
L^2-norm	$\|x\|_2$	
L^3-norm	$\|x\|_3$	
L^4-norm	$\|x\|_4$	
L^Infty-norm	$\|x\|_\infty$	

Revision: Norms

Compute for the vector x = (1,2,3)

Name	Symbol	Value
L^1-norm	$\|x\|_1$	6
L^2-norm	$\|x\|_2$	$\sqrt{14}$
L^3-norm	$\|x\|_3$	$6^{2/3}$
L^4-norm	$\|x\|_4$	$2^{1/4}\sqrt{7}$
L^Infty-norm	$\|x\|_\infty$	3

Quiz

- Given the following table, what information is missing to compute IDF?

Term Frequencies

	Doc1	Doc2	Doc3
Car	27	4	24
Auto	3	33	0
Insurance	3	33	0
Best	14	0	17

Quiz

- Reuters collection has 806,791 documents. Can you now compute the IDF?

	df	idf
Car	18165	
Auto	6723	
Insurance	19241	
Best	25235	

Quiz

- Reuters collection has 806,791 documents. Can you now compute the IDF?

	df	idf
Car	18165	1.65
Auto	6723	2.08
Insurance	19241	1.62
Best	25235	1.5

Quiz

- Why are we taking the log while computing IDF?

Bonus Task – 4% Marks.

- Give the story of log function in a way that the entire class can understand.
- Cover logarithm function in as much detail as possible including but not limited to:
  - What is “log” function? What is the difference between “ln”, “$\log_2$” and “$\log_{10}$”? When to go for what? What are the common applications of log? How does the log curve look like? What do we mean when we say “linear”, “sub-linear” and “exponential”? Which of these is related to “log” and how? Work out some examples wherever possible.
- Make a 10 to 15 minute presentation to the class. Only the top presentations (max 6) will get the marks. This task is optional.

Collection Vs. Document Frequency

- Collection Frequency = # Occurrences of the word in the entire collection.
- Document Frequency = # Documents containing the word.

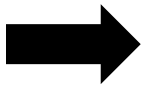
Word	Collection Frequency	Document Frequency
insurance	10440	3997
try	10422	8760

Two Ideas

- Document containing more occurrences of query term is more relevant to the query.
- Terms that occur in fewer documents are more important in the query (for relevance computation).

$$\textbf{Relevance} \propto \textbf{tf}$$

$$\textbf{Relevance} \propto \frac{1}{df}$$



Relevance ∝ TF * IDF

tf-idf weighting

Term weight is computed as follows:

$$w_{t,d} = \log(1 + \mathrm{tf}_{t,d}) \times \log_{10}(N / \mathrm{df}_t)$$

Document is scored as follows:

$$\mathrm{Score}(q,d) = \sum_{t \in q \cap d} \mathrm{tf.idf}_{t,d}$$

Thank You.