

# Vector Space Model



**SDSC** SAN DIEGO  
SUPERCOMPUTER CENTER

# After this video you will be able to

- Explain the term “vector model”
- Describe the concepts of similarity function and similarity search
- Recognize that many document and image search engines use vector models and similarity search

# Document Vector

3 documents



d1: "new york times"  
d2: "new york post"  
d3: "los angeles times"

- Let's create the term frequency matrix

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

# Document Vector

3 documents

d1: "new york times"  
d2: "new york post"  
d3: "los angeles times"

- Inverse document frequency

<u>TERM</u>	<u>DOC-FREQUENCY</u>	<u>IDF</u>
angeles	1	$\log_2(3/1) = 1.584$
los	1	$\log_2(3/1) = 1.584$
new	2	$\log_2(3/2) = 0.584$
post	1	$\log_2(3/1) = 1.584$
times	2	$\log_2(3/2) = 0.584$
york	2	$\log_2(3/2) = 0.584$

# The tf-idf matrix

<u>TERM</u>	<u>DOC-FREQUENCY</u>	<u>IDF</u>
<i>angeles</i>	1	$\log_2(3/1) = 1.584$
<i>los</i>	1	$\log_2(3/1) = 1.584$
<i>new</i>	2	$\log_2(3/2) = 0.584$
<i>post</i>	1	$\log_2(3/1) = 1.584$
<i>times</i>	2	$\log_2(3/2) = 0.584$
<i>york</i>	2	$\log_2(3/2) = 0.584$

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

# The tf-idf matrix

<u>TERM</u>	<u>DOC-FREQUENCY</u>	<u>IDF</u>
angeles	1	$\log_2(3/1) = 1.584$
los	1	$\log_2(3/1) = 1.584$
new	2	$\log_2(3/2) = 0.584$
post	1	$\log_2(3/1) = 1.584$
times	2	$\log_2(3/2) = 0.584$
york	2	$\log_2(3/2) = 0.584$

	angeles	los	new	post	times	york	Length
d1	0	0	0.584	0	0.584	0.584	1.011
d2	0	0	0.584	1.584	0	0.584	1.786
d3	1.584	1.584	0	0	0.584	0	2.316

$\text{Length of } d1 = \sqrt{0.584^2 + 0.584^2 + 0.584^2} = 1.011$

# Searching in Vector Space

query



q: new new york

- Max frequency of a term ("new") = 2
- Create the query vector

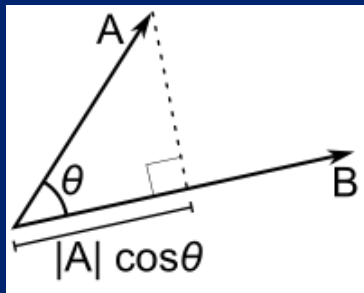
$Q [0 \ 0 \ (2/2)*0.584=0.584 \ 0 \ (1/2)*0.584=0.292 \ 0]$   $length(q)=0.652$

- A similarity function between two vectors is a measure of how far they are apart

# Similarity Function

- Many possible functions
- Cosine distance

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



$$\text{cosSim}(d1,q) = (0.584 \cdot 0.584 + 0.584 \cdot 0.292) / (1.011 \cdot 0.652) = 0.776$$

$$\text{cosSim}(d2,q) = (0.584 \cdot 0.584) / (1.786 \cdot 0.652) = 0.292$$

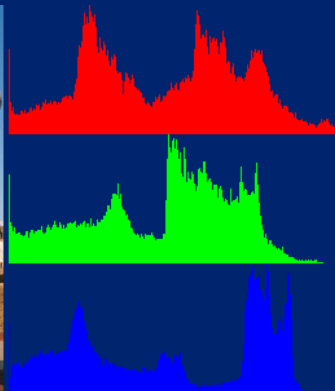
$$\text{cosSim}(d3,q) = (0.584 \cdot 0.292) / (2.316 \cdot 0.652) = 0.112$$



# Query Term Weighting

- Every query term may optionally be associated with a weighting term
  - $Q = \text{York}^1 \text{ times}^2 \text{ post}^5$ 
    - $\text{wt}(\text{York}) = 1/(1+2+5) = 1/8 = 0.125$
    - $\text{wt}(\text{times}) = 2/8 = 0.25$
    - $\text{wt}(\text{post}) = 5/8 = 0.625$
  - Multiply the query vector with these weights
  - “new york post” ranks first

# Image Search



	0-31	32-63	64-95	96-127	128-159	159-191	192-223	223-255
Red	0.04	0.12	0.23	0.06	0.24	0.12	0.13	0.06
Green	0.05	0.07	0.11	0.07	0.26	0.24	0.17	0.03
Blue	0.08	0.13	0.16	0.08	0.03	0.12	0.19	0.21