- **Idea**: The most important terms in a document are the terms that occur most often.

- So, let's count the number of occurrences of each term in each document.
  In other words, let's count the **frequency** of each term in each document.

- For example, consider the following three documents:

big big big big data class
data big data science
science big data

- Let's construct a matrix, where:
  - there is one row per **document**,
  - one column per unique **term**, and
  - the value in row d and column t is the **number of occurrences of term t in document d**.

	big	data	class	science
big big big big data class	4	1	1	0
data big data science	1	2	0	1
science big data	1	1	0	1

Bag of words

- The **bag of words** model represents documents as **vectors of word counts**, i.e. **term frequencies**.
  The matrix below was created using the bag of words model.

- Each **row** in the bag of words matrix is a **vector representation** of a document.

	big	data	class	science
big big big big data class	4	1	1	0
data big data science	1	2	0	1
science big data	1	1	0	1

- For example, we can represent the document 2, **data big data science**, with the vector $\vec{d_2}$:

document as vector.

$$\vec{d_2} = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \end{bmatrix}$$

Vectors

$$\vec{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

n components: $\vec{v} \in \mathbb{R}^n$

Example:

$$\vec{a} = \begin{bmatrix} 2 \\ 1 \\ -4 \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 0 \\ 5 \\ 1 \end{bmatrix} \qquad \vec{a}, \vec{b} \in \mathbb{R}^3$$

① Add vectors elementwise

$$\vec{a} + \vec{b} = \begin{bmatrix} 2 \\ 6 \\ -3 \end{bmatrix}$$

② Scalar multiplication

$$-3\vec{a} = \begin{bmatrix} -6 \\ -3 \\ 12 \end{bmatrix}$$

The dot product

Given that $\vec{a}, \vec{b} \in \mathbb{R}^n$, the dot product is:

$$\vec{a} \cdot \vec{b} = \overbrace{a_1 b_1 + a_2 b_2 + \dots + a_n b_n}^{\text{a scalar!}}$$

e.g. $\vec{a} = \begin{bmatrix} 2 \\ 1 \\ -4 \end{bmatrix}$ $\vec{b} = \begin{bmatrix} 0 \\ 5 \\ 1 \end{bmatrix}$

$$\vec{a} \cdot \vec{b} = (2)(0) + (1)(5) + (-4)(1)$$

$$= 0 + 5 - 4$$

$$= \boxed{1}$$

$\boxed{\text{scalar!}}$ not a vector

Geometric definition

Assume $\vec{a}, \vec{b} \in \mathbb{R}^n$.

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \, \|\vec{b}\| \cos\theta$$

magnitude

the angle between them

$$\vec{a} = \begin{bmatrix} 2 \\ 1 \\ -4 \end{bmatrix} \qquad \vec{b} = \begin{bmatrix} 0 \\ 5 \\ 1 \end{bmatrix}$$

previous slide: $\vec{a} \cdot \vec{b} = 1$

equal!

$$\vec{a} \cdot \vec{b} = \boxed{(\sqrt{21})(\sqrt{26}) \cos\theta}$$

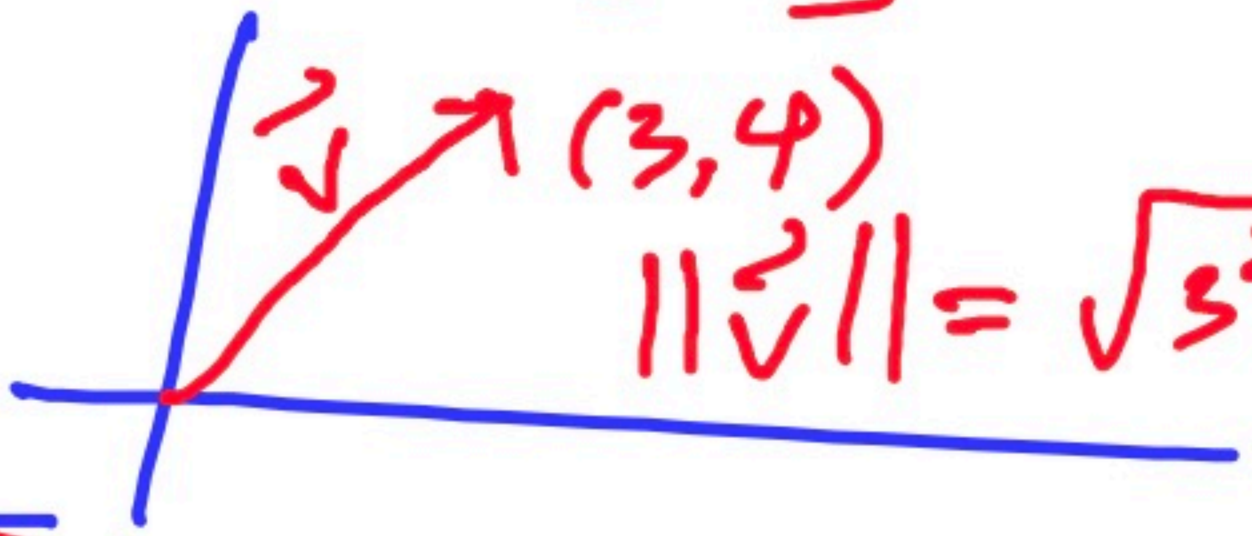
$$\|\vec{a}\| = \sqrt{2^2 + 1^2 + (-4)^2} = \sqrt{21} \qquad \|\vec{b}\| = \sqrt{0^2 + 5^2 + 1^2} = \sqrt{26}$$

Aside:

$\vec{v} \in \mathbb{R}^n$,

$$\|\vec{v}\|^2 = \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}$$

e.g. $\vec{v} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$



$$\|\vec{v}\| = \sqrt{3^2 + 4^2} = 5$$

$$\vec{a} \cdot \vec{b} = 1$$

also

$$\vec{a} \cdot \vec{b} = \sqrt{21}\,\sqrt{26}\,\cos\theta$$

$$\Longrightarrow 1 = \sqrt{21}\,\sqrt{26}\,\cos\theta$$

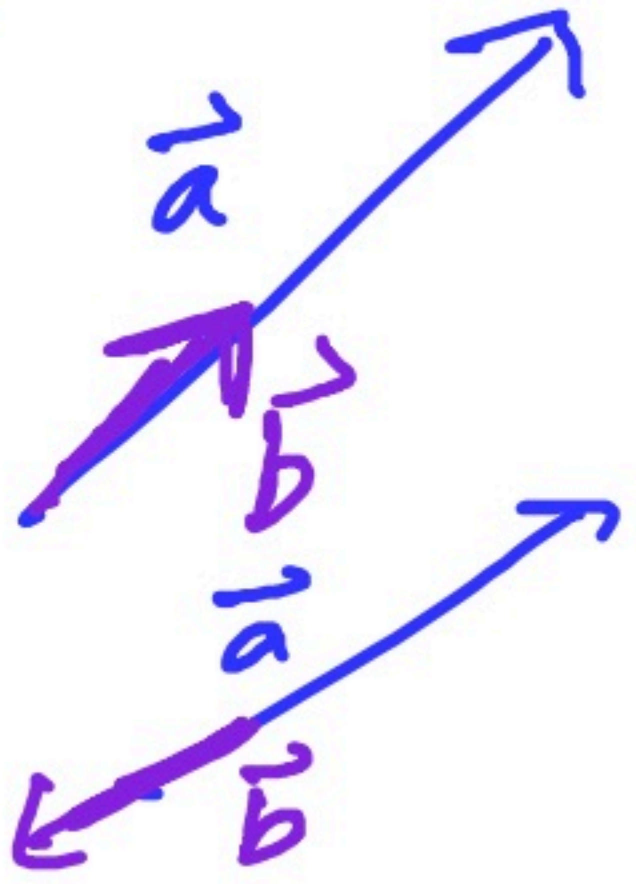
$$\Rightarrow \boxed{\cos\theta = \frac{1}{\sqrt{21}\,\sqrt{26}}}$$

the more similar two vectors are, the larger $\cos\theta$ is!

$$\boxed{-1 \le \cos\theta \le 1}$$

$\cos\theta = 1$ when $\theta = 0^\circ$

$\cos\theta = -1$ when $\theta = 180^\circ$



Cosine similarity

- To measure the similarity between two documents, we can compute the **cosine similarity** of their vector representations:

$$\text{cosine similarity}(\vec{u}, \vec{v}) = \cos\theta = \boxed{\frac{\vec{u}\cdot\vec{v}}{|\vec{u}||\vec{v}|}}$$

$$\vec{u}\cdot\vec{v} = ||\vec{u}||\,||\vec{v}||\cos\theta$$

$$\Rightarrow \cos\theta =$$

Normalizing

- Why can't we just use the dot product – that is, why must we divide by $|\vec{u}||\vec{v}|$ when computing cosine similarity?

$$\text{cosine similarity}(\vec{u}, \vec{v}) = \cos\theta = \boxed{\frac{\vec{u} \cdot \vec{v}}{|\vec{u}||\vec{v}|}}$$

normalizing

the dot product.

	big	data	class	science
big big big big data class	4	1	1	0
data big data science	1	2	0	1
science big data	1	1	0	1

- Consider the following two *pairs* of documents:

Pair	Dot Product	Cosine Similarity
big big big big data class and **data big data science**	6	0.577
science big data and **data big data science**	4	0.943

- **"big big big big data class"** has a large dot product with **"data big data science"** just because the former has the

- The **inverse document frequen**… , . . . is:

$$\text{idf}(t) = \log\left(\frac{\text{total \# of documents}}{\text{\# of documents in which } t \text{ appears}}\right)$$

- **Example**: What is the inverse document frequency of **"billy"** in the following three documents?
  - "my brother has a friend named **billy** who has an uncle named **billy**"
  - "my favorite artist is named jilly boel"
  - "why does he talk about someone named **billy** so often"

if term in only one doc: $\log\left(\frac{n}{1}\right)$ ✓.

- **Answer**: $\log\left(\frac{3}{2}\right) \approx 0.4055$.

Here, we used the natural logarithm. It doesn't matter which log base we use, as long as we keep it consistent throughout all of our calculations.

if term in every doc: $\log\left(\frac{n}{n}\right) = 0$

- Intuition: If a word appears in every document (like **"the"** or **"has"**), it is probably not a good summary of any one document.

- Think of $\text{idf}(t)$ as the "rarity factor" of t across documents – the larger $\text{idf}(t)$ is, the more rare t is.

$$\text{idf}(t) \text{ large} \implies t \text{ rare across all documents}$$
$$\text{idf}(t) \text{ small} \implies t \text{ common across all documents}$$

Term frequency-inverse document frequency

- The **term frequency-inverse document frequency (TF-IDF)** of term t in document d is the product:

$$\begin{aligned}
\text{tfidf}(t, d) &= \text{tf}(t, d) \cdot \text{idf}(t) \\
&= \underbrace{\frac{\#\text{ of occurrences of } t \text{ in } d}{\text{total }\#\text{ of terms in } d}}_{\text{how common is } t \text{ in } d\text{?}} \cdot \underbrace{\log\left(\frac{\text{total }\#\text{ of documents}}{\#\text{ of documents in which } t \text{ appears}}\right)}_{\text{how rare is } t \text{ overall.}}
\end{aligned}$$

?