---

1. Document Representation

Suppose we have a small vocabulary:

$$\text{Vocab}=\{\text{cat, dog}\}$$

We represent documents as **term-frequency vectors**.

- **Doc1**: `"cat cat dog"` → vector = (2, 1)
- **Doc2**: `"cat dog"` → vector = (1, 1)
- **Doc3**: `"dog dog dog dog"` → vector = (0, 4)

---

2. Euclidean Distance

$$d_E(A,B)=\sqrt{\sum(A_i-B_i)^2}$$

- Distance(Doc1, Doc2)

$$¿\sqrt{(2-1)^2+(1-1)^2}=\sqrt{1}=1$$

- Distance(Doc1, Doc3)

$$¿\sqrt{(2-0)^2+(1-4)^2}=\sqrt{4+9}=\sqrt{13}\approx 3.6$$

□ Euclidean says **Doc1 is closer to Doc2** (which is fine).
But notice something important:

- If we just **repeat Doc2 multiple times**, its vector length increases, and Euclidean distance also increases—even though the content is the same!

Example: Doc2 = (1,1), Doc2_long = (10,10)

$$d_E((1,1),(10,10))=\sqrt{(9^2+9^2)}=\sqrt{162}\approx 12.7$$

□ This means the **same document but longer looks "far"** under Euclidean distance.

---

3. Cosine Similarity

Cosine measures **angle**, not length:

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||}$$

- Similarity(Doc1, Doc2):

$$\text{¿} \frac{(2)(1)+(1)(1)}{\sqrt{2^2+1^2} \cdot \sqrt{1^2+1^2}} = \frac{2+1}{\sqrt{5} \cdot \sqrt{2}} = \frac{3}{\sqrt{10}} \approx 0.95$$

- Similarity(Doc1, Doc3):

$$\text{¿} \frac{(2)(0)+(1)(4)}{\sqrt{5} \cdot \sqrt{16}} = \frac{4}{\sqrt{80}} = \frac{4}{8.94} \approx 0.45$$

- Similarity(Doc2, Doc2_long):

$$\text{¿} \frac{(1)(10)+(1)(10)}{\sqrt{2} \cdot \sqrt{200}} = \frac{20}{1.41 \cdot 14.14} = \frac{20}{20} = 1$$

□ Cosine gives **1.0 (perfect match)** for the same doc regardless of length.
□ It correctly shows Doc1 is more similar to Doc2 (0.95) than to Doc3 (0.45).

---

4. Conclusion

- **Euclidean Distance** penalizes document length, so longer documents always look "farther," even if they have the same proportions of words.

- **Cosine Similarity** removes the effect of document length (normalizes vectors) and focuses only on word distribution.

That’s why in **text mining, IR, and NLP**, we prefer **Cosine similarity/distance** over Euclidean.
