VEKTORROMMODELLER

Vektorrommodeller

Representerer språklige data som trekkvektorer

- hvert trekk representerer én dimensjon
- Hvert objekt er ett punkt i vektorrommet
- Vektorrommet kan representeres som en matrise

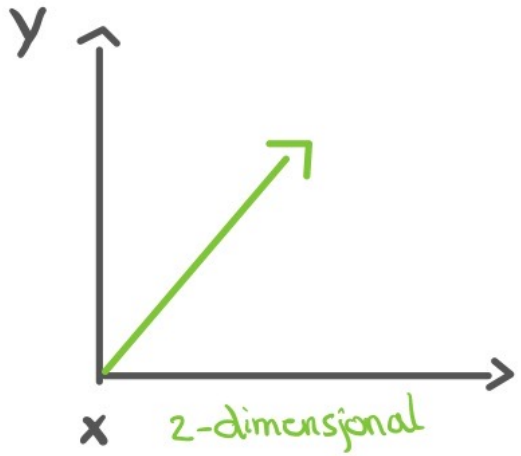
Måle likhet

Vi kan måle likhet mellom ord eller dokumenter ved å måle avstanden mellom vektorer

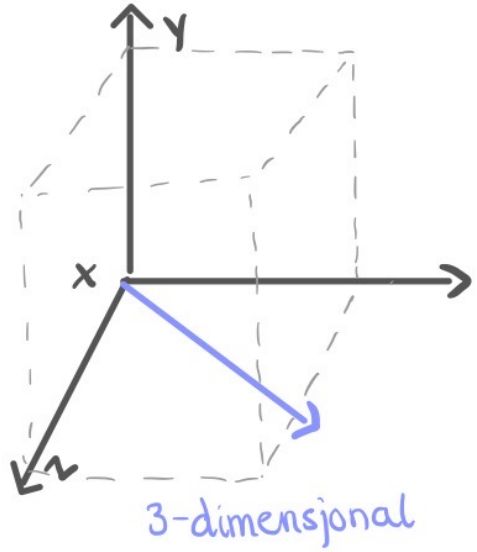
HVA ER EN VEKTOR?

- En *n*-dimensjonal vektor er en liste av reelle tall med en størrelse og en retning

$[x_0]$

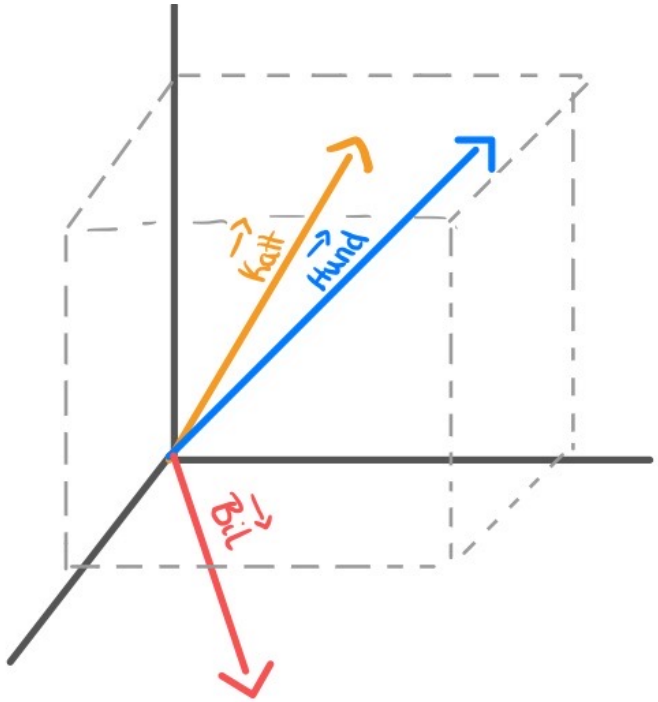


$[x_0, x_1]$



$[x_0, x_1, x_2]$

TREKKVEKTORER



3-dimensjonal vektor med tre dokumenter om katt, hund og bil

- Hvert punkt er en trekkvektor som representerer dokumentet
- Hvert trekk kan f.eks. være antall forekomster av ord

Katt = [katt:10, hund:3, bil:0]

Hund = [katt:4, hund:15, bil:1]

Bil = [katt:0, hund:1, bil:15]

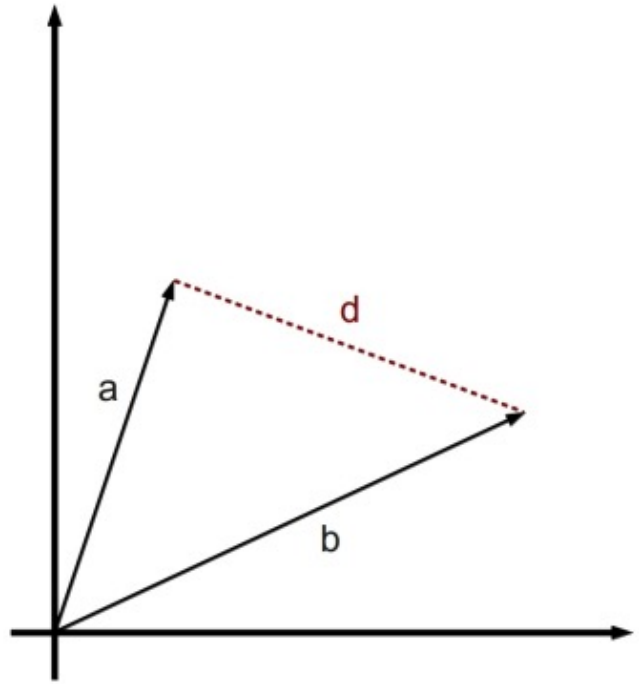
- Et vektorrom kan ha flere hundre tusen av dimensjoner!

Euklidsk avstand

Måles mellom ytterpunktene til to vektorer

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{a}_i - \boldsymbol{b}_i)^2}$$

Men: Er ikke alltid optimal for dokumenter av ulik lengde



Cosinus-likhet

Viser om to vektorer peker i (ca.) samme retning

Måles mellom to vektorer

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\sum_i \boldsymbol{a}_i \boldsymbol{b}_i}{\sqrt{\sum_i \boldsymbol{a}_i^2} \sqrt{\sum_i \boldsymbol{b}_i^2}} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}$$