

1906003132015

Doğal Dil İşleme

BAİBÜ Bilgisayar Müh.

Dr. Araş. Gör. İsmail Hakkı Parlak

ismail.parlak@ibu.edu.tr

Oda: 329

Morfoloji

- **Morfoloji:** Biçimbilim, yapıbilim. Sözcüklerin içyapısını inceleyen dilbilimi alt dalıdır. Kök, gövde, ek gibi bileşenleri inceler.
- **Hece:** Seslem. Ağzın tek hareketinde ifade edilebilen, 1 veya daha çok kelimeden oluşan birimler. Anlamalı veya anlamsız olabilir.
- **Morfem:** Dilin en küçük anlamalı birimi. Anlam ayırıcılar. Kelime kök ve eklerine ayrıştırıldığında morfemlerine ayrıştırılmış olur. Yapım ekleri köke dahil edilir. Çekim ekleri köke dahil edilmez

Tokenization

- "İki, iki daha 5 eder. Her şey bu kadar basit." → ["İki, iki daha 5 eder", "Her şey bu kadar basit"]
- "Naber? İyi misin?"
 - ["Naber?", "İyi misin?"]
 - ["Naber", "İyi misin"]
 - ["Naber", "İyi", "misin"]
 - ["naber?", "iyi misin?"]
 - ["naber", "iyi"]
 - ...

Stop Words (Dolgu Sözcükleri)

- The, to, it, as, ...
- Ben, ama, veya, ile, şey, ...
- Veri setinde çok sayıda bulunup da model için faydalı olmayabilecek kelimeler.
- Sayıları çok olduğu için dağılımda dengesizlik yaratabilirler.
- Doküman benzerlikleri araştırılırken modellerin başarısını düşürebilirler.

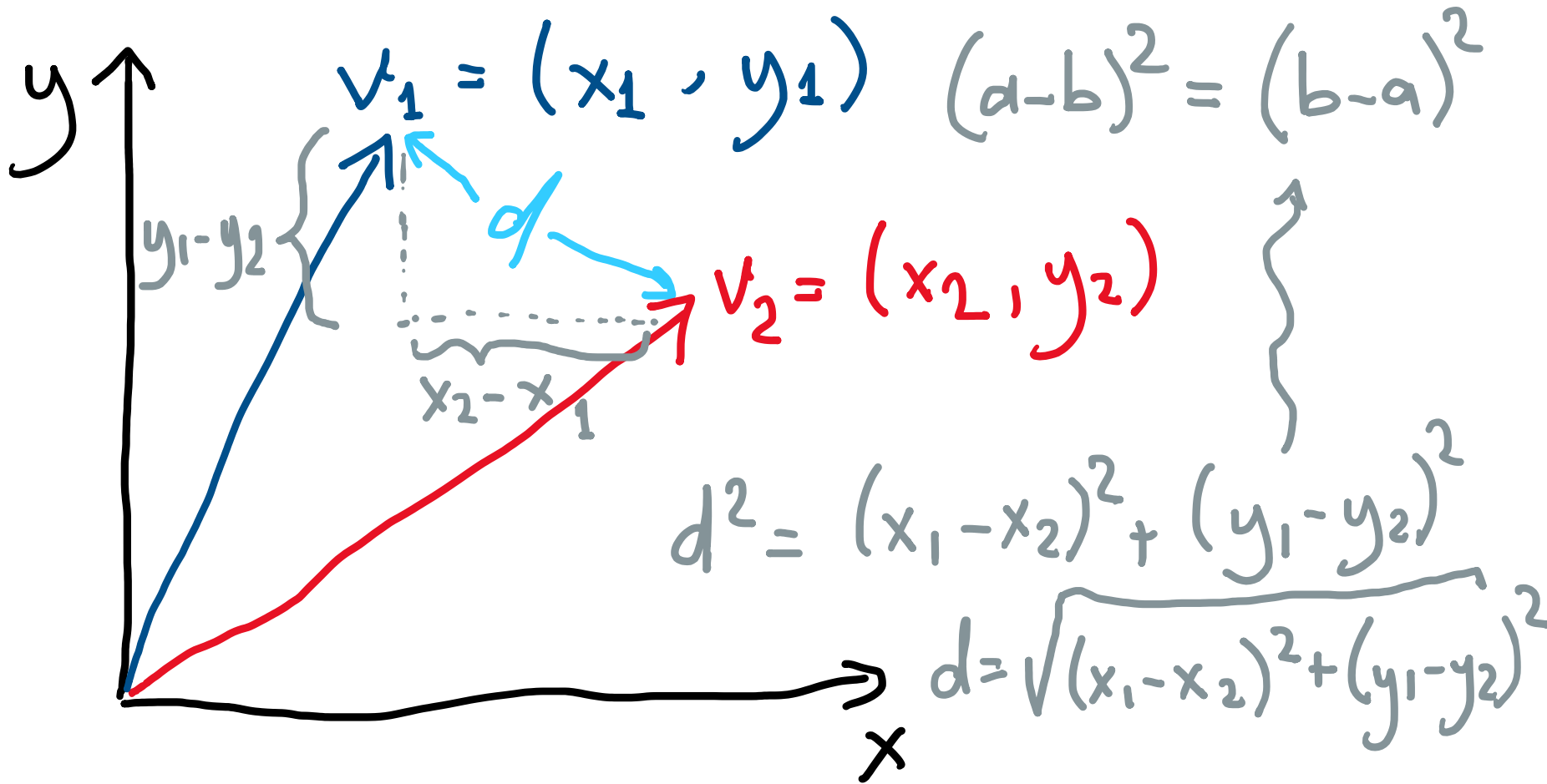
Lemmatization & Stemming

- eat, eaten, ate, eating, ... Aynı eylemi ifade eden farklı yazılışlar. Hepsı ayrı mı değerlendirilmeli?
- **Stemming:** Kelimenin sonunu kesip atma.
 - apples -> apple
 - ponies -> poni
 - better -> better
 - was -> wa
- **Lemmatization:** Kelimenin kökünü bulma.
 - apples -> apple
 - ponies -> pony
 - better -> good
 - was -> be

Vektör Benzerlikleri

- d1: ["a", "b", "a", "b", "a", "a", "b"] -> a:4, b:3
 - d2: ["b", "b", "a", "a", "b", "b", "b"] -> a:2, b:5
 - d3: ["b", "a", "a", "b", "b", "a", "a", "a", "b", "b", "a", "a", "b", "a",] -> a:8, b:6
-
- d_i : Dokümanlar
 - Hangi dokümanlar daha benzer?

Euclidean Distance (Öklit Uzaklığı)

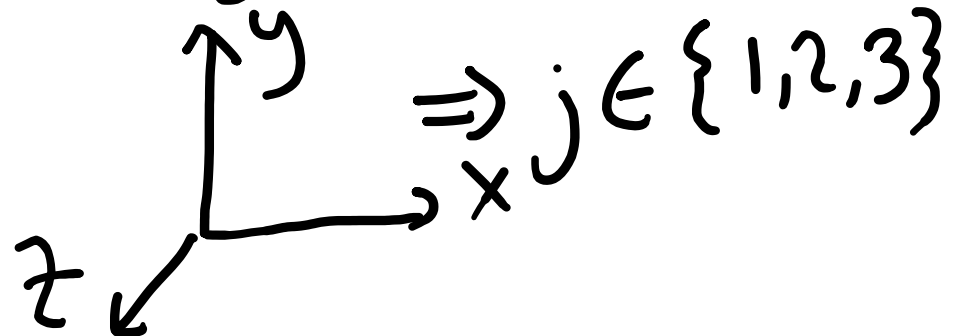
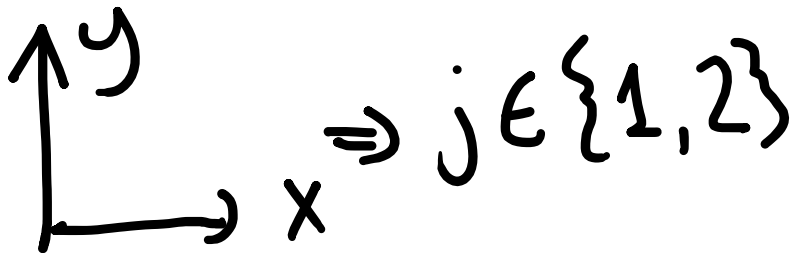


Euclidean Distance

$$\|v_1 - v_2\|_2 = \sqrt{(v_{11} - v_{21})^2 + (v_{12} - v_{22})^2 + \dots + (v_{1D} - v_{2D})^2}$$

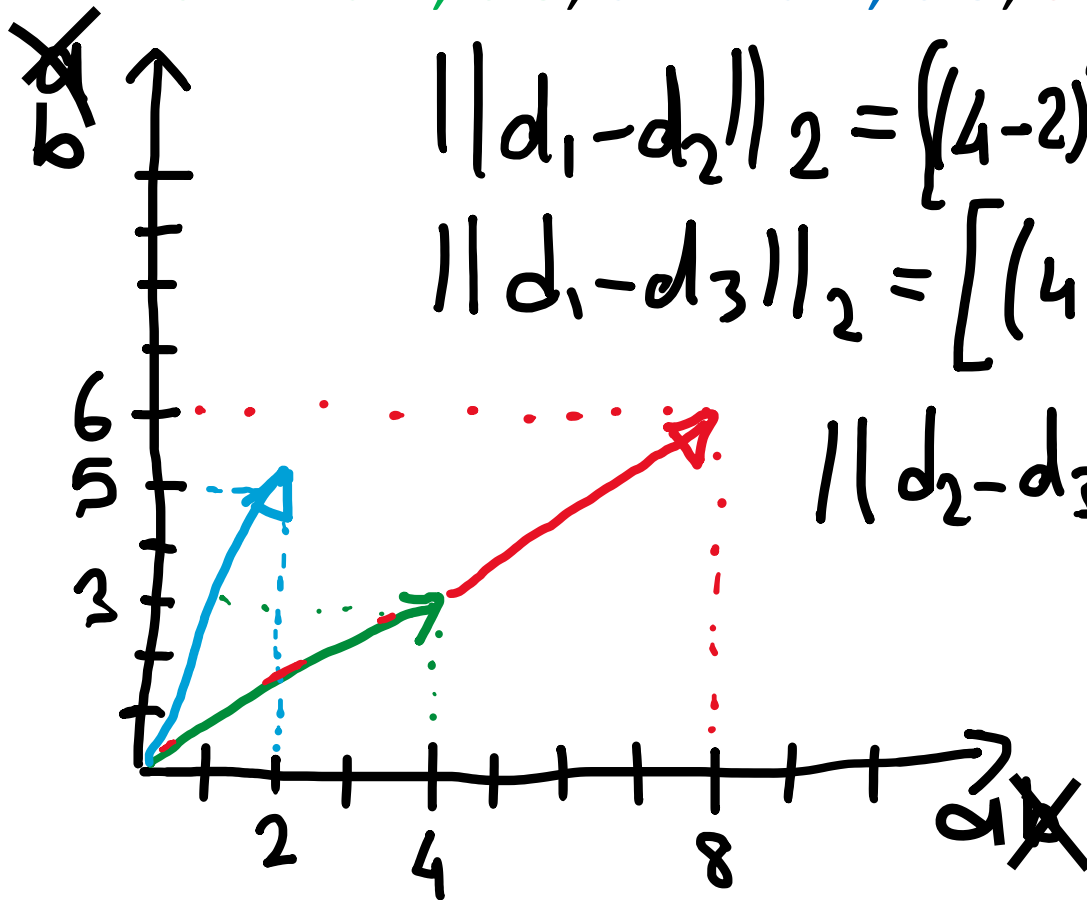
v_{ij} : $i \Rightarrow$ hangi vektör

$j \Rightarrow$ vektörün j 'inci bileşeni



Benzerlik

- d1 -> a:4, b:3; d2 -> a:2, b:5; d3 -> a:8, b:6



$$\|d_1 - d_2\|_2 = \{(4-2)^2 + (3-5)^2\}^{1/2} = 8^{1/2}$$

$$\|d_1 - d_3\|_2 = \{(4-8)^2 + (3-6)^2\}^{1/2} = 25^{1/2}$$

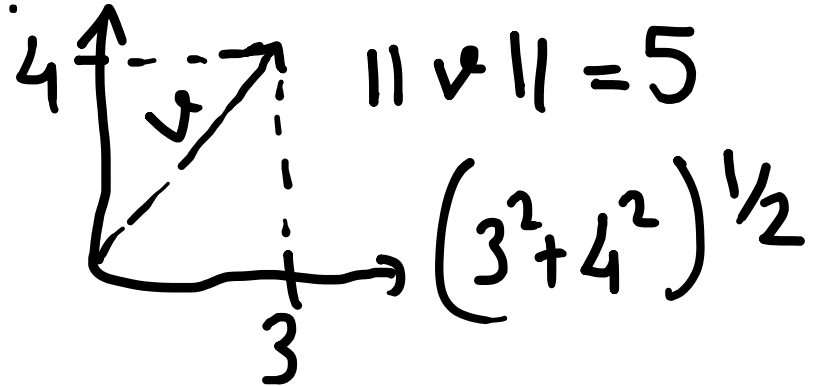
$$\|d_2 - d_3\|_2 = \{(2-8)^2 + (5-6)^2\}^{1/2} = 37^{1/2}$$

$$\begin{aligned} d_1 - d_2 &: 8 \checkmark \\ d_1 - d_3 &: 25 \\ d_2 - d_3 &: 37 \end{aligned}$$

Çözüm 1: Normalizasyon

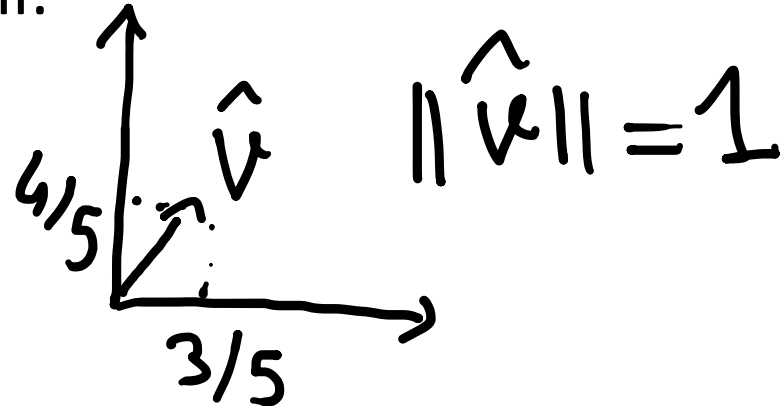
- Vektörlerin boyunu 1'e indirmek.
- D boyutlu bir vektörün boyu:

$$\|v\| = \left(\sum_{j=1}^D v_j^2 \right)^{1/2}$$

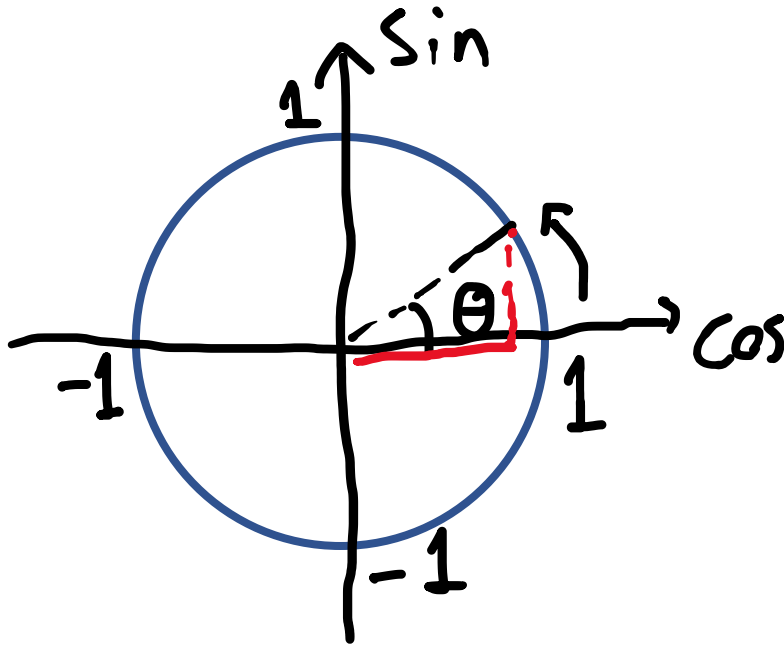


- Normalize etme işlemi:

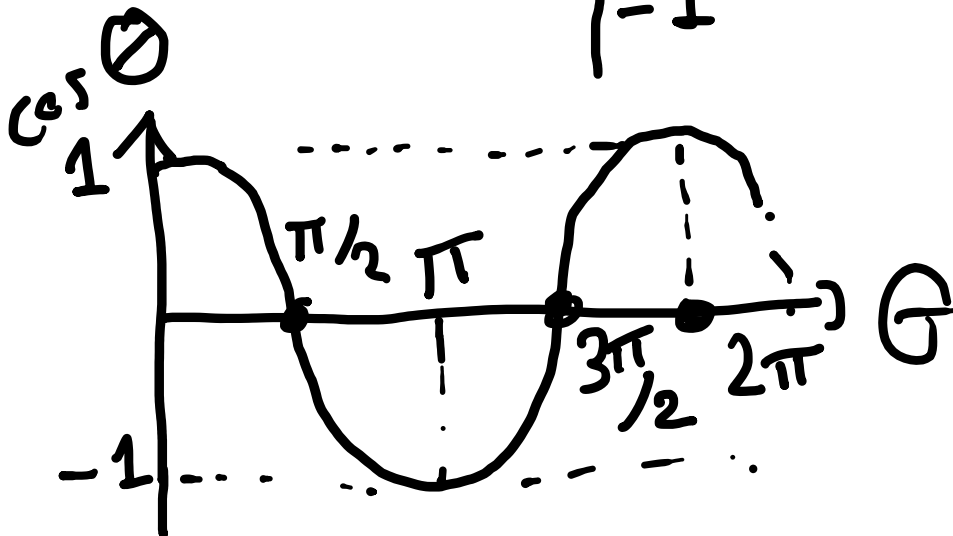
$$\hat{v} = \frac{v}{\|v\|}$$



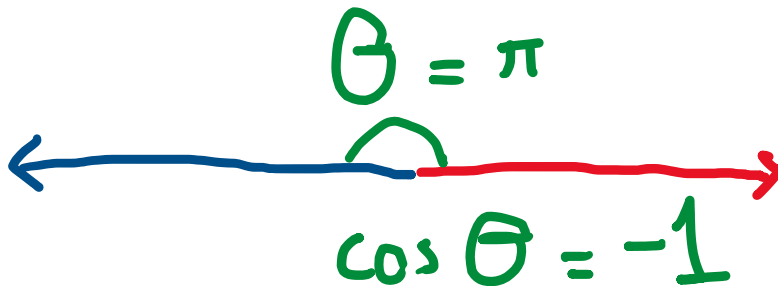
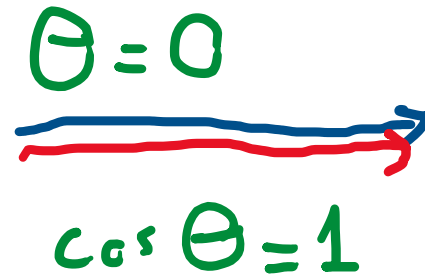
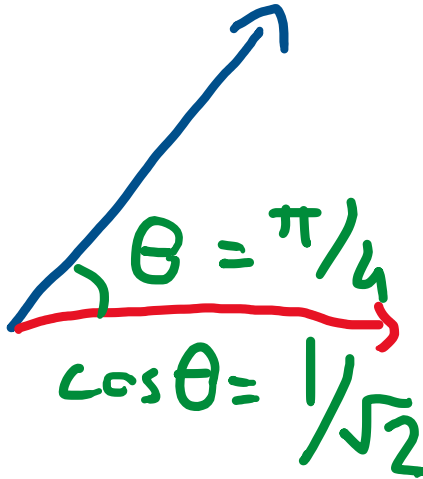
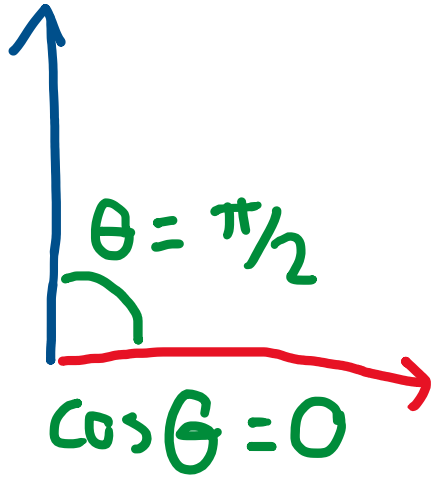
Çözüm 2: Cosine Similarity



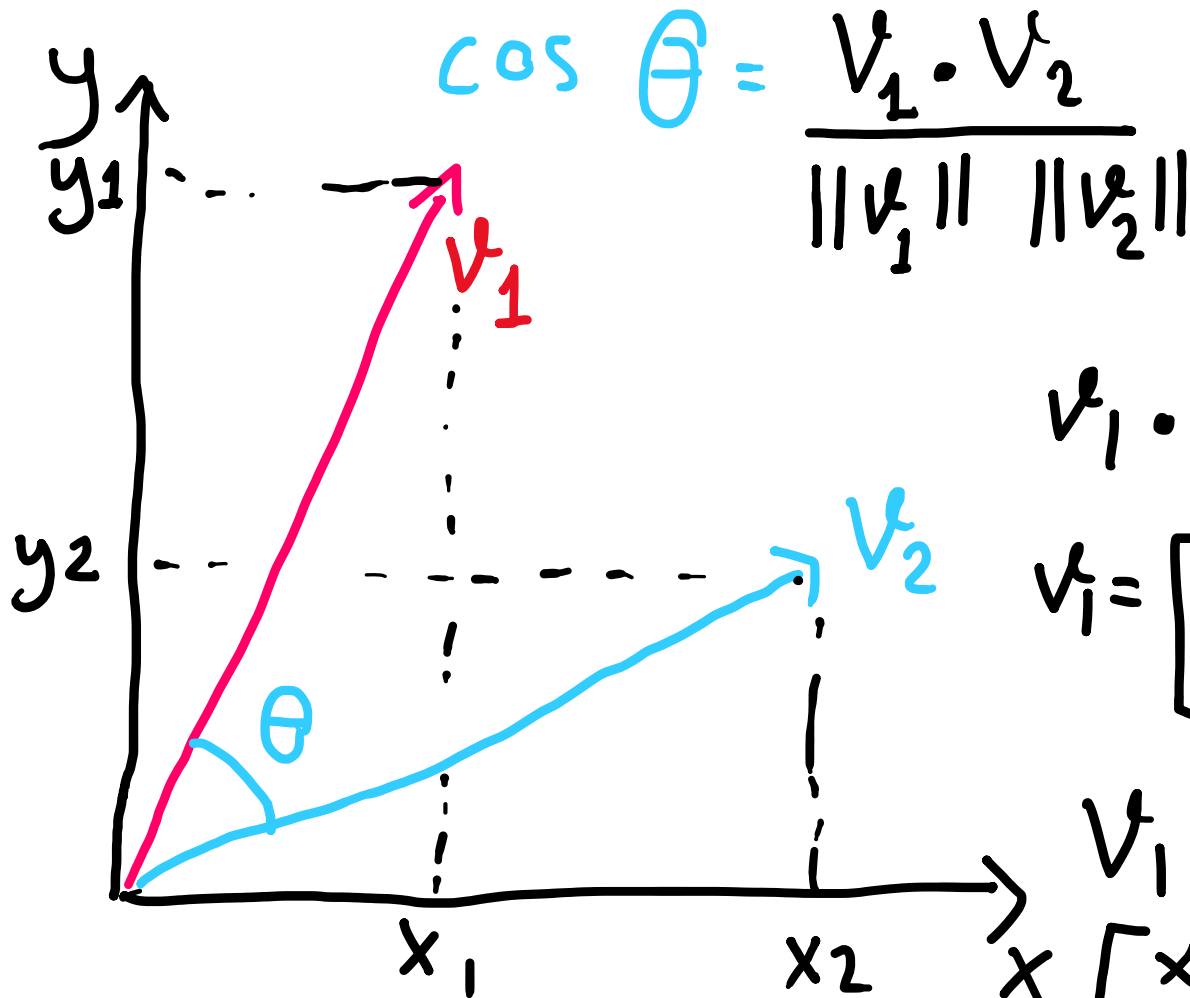
θ	$\cos \theta$
0	1
$\pi/2$	0
π	-1
$3/2\pi$	0
2π	1



Çözüm 2: Cosine Similarity



Çözüm 2: Cosine Similarity



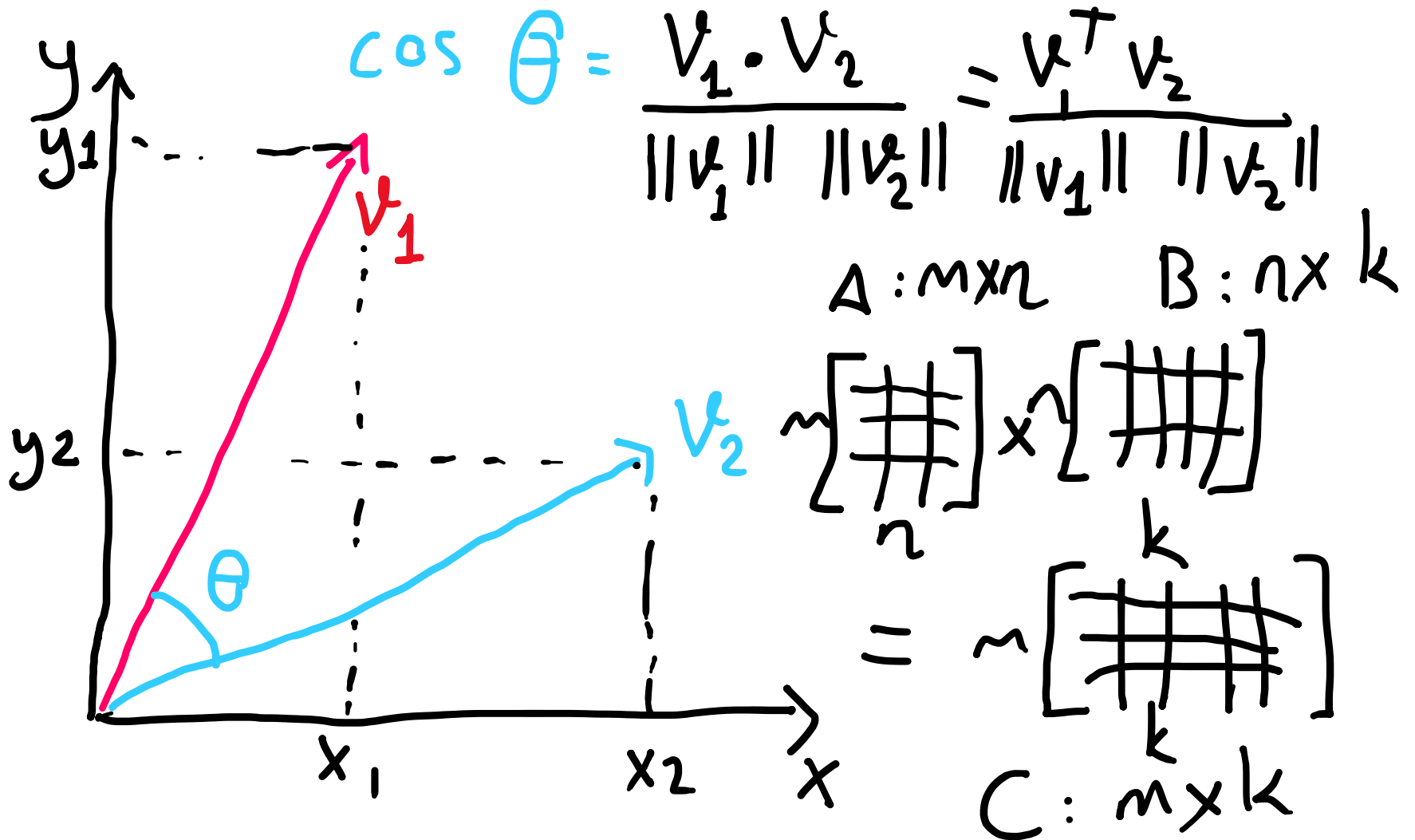
$$\cos \theta = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$$

$$v_1 \cdot v_2 = x_1 x_2 + y_1 y_2$$

$$v_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

$$v_1 \cdot v_2 = v_1^T v_2$$
$$\begin{bmatrix} x_1 & y_1 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

Çözüm 2: Cosine Similarity



Python'a Giriş

- Google colab.