

# DOĞAL DİL İŞLEMESİNE GİRİŞ

BAHAR DÖNEMİ - 2023-2024

BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

BURSA TEKNİK ÜNİVERSİTESİ 2010

DR. ÖĞR. ÜYESİ HAYRİ VOLKAN AGUN

# Özet

---

- ❑ Kelime Torbası Yaklaşımı - Bag of words ?
- ❑ Sınıflandırma
- ❑ Ağırlıklandırma
- ❑ Kaynaklar



# Kelime Torbası

Doküman / İçerik	Kelime Çıkarımı	Sözlük
Edirne'de Meriç, Tunca ve Arda nehirleri etrafında 2315 parça bakımlı sebze, meyve ve dut bahçeleri vardır. Kayısı, erik, ayva, dut, muşmula ve diğer meyveleri boldur.	{ Edirne, Meriç, Tunca, Arda, nehirleri, etrafında 2315, parça, bakımlı, sebze, meyve, ve, dut, bahçeleri, vardır, ., Kayısı, erik, ayva, dut, muşmula, ve, diğer, meyveleri, boldur, . }	edirne, meriç, tunca, arda, nehirleri, etrafında, parça, bakımlı, sebze, meyve, dut, muşmula, meyveleri, lozan, yunanistan, savaş, tazminatı, kayısı, erik, ayva, antlaşması, karaağaç, türkiye, anısına, anıtı, anlaşma, ilçenin, inşa, eylül
Lozan Antlaşması ile Yunanistan'dan savaş tazminatı olarak geri alınan Karaağaç'ın 15 Eylül 1923'te Türkiye'ye katılmasıyla ilin sınırı bugünkü halini aldı. Antlaşma anısına inşa edilen Lozan Anıtı ilçenin Karaağaç mahallesindedir.	{ Lozan, Antlaşması, ile Yunanistan, dan, savaş, tazminatı, olarak, geri, alınan, Karaağaç,ın, 15, Eylül, 1923, te, Türkiye, ye, ktılmasıyla, ilin, sınırı, bugünkü, halini, aldı, ., Antlaşma, anısına, inşa, edilen, Lozan, Anıtı, ilçenin, Karaağaç, mahallesindedir, . }	ve, dan, ile, diğer, olarak,  edilen, aldı, mahallesindedir, vardır, boldur, alınan,  2315, 15, 1923

# Kelime Torbası

Doküman / İçerik	Kelime Çıkarımı	Sözlük
Edirne'de Meriç, Tunca ve Arda nehirleri etrafında 2315 parça bakımlı sebze, meyve ve dut bahçeleri vardır. Kayısı, erik, ayva, dut, muşmula ve diğer meyveleri boldur.	{ Edirne, Meriç, Tunca, Arda, nehirleri, etrafında 2315, parça, bakımlı, sebze, meyve, ve, dut, bahçeleri, vardır, ., Kayısı, erik, ayva, dut, muşmula, ve, diğer, meyveleri, boldur, . }	edirne, meriç, tunca, arda, nehirleri, etrafında, parça, bakım, sebze, meyve, dut, muşmula, meyveleri, lozan, yunanistan, savaş, tazminat, kayısı, erik, ayva, antlaşması, karaağaç, türkiye, anısına, anıt, antlaşma, ilçenin, inşa, eylül
Lozan Antlaşması ile Yunanistan'dan savaş tazminatı olarak geri alınan Karaağaç'ın 15 Eylül 1923'te Türkiye'ye katılmasıyla ilin sınırı bugünkü halini aldı. Antlaşma anısına inşa edilen Lozan Anıtı ilçenin Karaağaç mahallesindedir.	{ Lozan, Antlaşması, ile Yunanistan, dan, savaş, tazminatı, olarak, geri, alınan, Karaağaç,ın, 15, Eylül, 1923, te, Türkiye, ye, kılmasıyla, ilin, sınırı, bugünkü, halini, aldı, ., Antlaşma, anısına, inşa, edilen, Lozan, Anıtı, ilçenin, Karaağaç, mahallesindedir, . }	ve, dan, ile, diğer, olarak, edilen, aldı, mahallesindedir, vardır, boldur, alınan, katılmasıyla  2315, 15, 1923

# Kelime Torbası

Doküman / İçerik	Kelime Çıkarımı	Sözlük
Edirne'de Meriç, Tunca ve Arda nehirleri etrafında 2315 parça bakımlı sebze, meyve ve dut bahçeleri vardır. Kayısı, erik, ayva, dut, muşmula ve diğer meyveleri boldur.	{ Edirne, Meriç, Tunca, Arda, nehirleri, etrafında 2315, parça, bakımlı, sebze, meyve, ve, dut, bahçeleri, vardır, ., kayısı, erik, ayva, dut, muşmula, ve, diğer, meyveleri, boldur, . }	edirne, meriç, tunca, arda, nehir, etraf, parça, bakım, sebze, meyve, dut, kayısı, erik, ayva, muşmula, meyve, yunanistan, <b>savaş tazminatı</b> , <b>lozan antlaşması</b> , karaağaç, türkiye, <b>lozan anıtı</b> , ilçe, inşa
Lozan Antlaşması ile Yunanistan'dan savaş tazminatı olarak geri alınan Karaağaç'ın 15 Eylül 1923'te Türkiye'ye katılmasıyla ilin sınırı bugünkü halini aldı. Antlaşma anısına inşa edilen Lozan Anıtı ilçenin Karaağaç mahallesindedir.	{ Lozan, Antlaşması, ile Yunanistan, dan, savaş, tazminatı, olarak, geri, alınan, Karaağaç, ın, 15, Eylül, 1923, te, Türkiye, ye, katılmasıyla, ilin, sınırı, bugünkü, halini, aldı, ., Antlaşma, anısına, inşa, edilen, Lozan, Anıtı, ilçenin, Karaağaç, mahallesindedir, . }	ve, dan, ile, diğer, ol,  <b>edilen, aldı, mahale, var, bol, al, katıl</b>  15 Eylül 1923 2315

# Kelime Torbası/Çantası

---

- ☐ Her bir doküman için:
- ☐ Kelime sınırlarının bulunması ile kelimeler çıkarılır.
- ☐ Kelimeler sıralı bir şekilde bir dizide toplanır. Buna kelime dizisi denir.
- ☐ Kelime dizindeki kelimelerde bulunan ekler kaldırılır ve kelimeler birlikte geçme sıklıklarına göre gruplandırılır.
- ☐ Kelime dizisinde bulunan tüm kelimeler sözlüğe eklenir.
- ☐ Doküman torbası her bir kelimenin tek bir kez geçtiği bir küme şeklinde ifade edilir.
- ☐ Her bir doküman sözlükte geçen kelimeleri barındırıp barındırmadığına göre bir vektör ile ifade edilir.
- ☐ Bu vektörde dokümanda geçen kelimeler
  - Geçip geçmeme durumuna göre 1 yada 0 ile
  - Geçme sayısına göre rakamla.
  - Geçme ağırlığına göre kayar nokta sayısı ile (tf\*idf)

# Doküman Vektörü

❑ Sözlükteki kelimelerin her bir kök olarak dokümanda kaç kere geçiyor.

Doküman / İçerik	Sözlük	Doküman Vektörü
Edirne'de Meriç, Tunca ve Arda nehirleri etrafında 2315 parça bakımlı sebze, meyve ve dut bahçeleri vardır. Kayısı, erik, ayva, dut, muşmula ve diğer meyveleri boldur.	edirne, meriç, tunca, arda, nehir, etraf, parça, bakım, sebze, <b>meyve</b> , dut, muşmula, kayısı, erik, ayva, <b>yunanistan</b> , <b>savaş tazminatı</b> , <b>lozan antlaşması</b> , <b>karaağaç</b> , <b>türkiye</b> , <b>lozan anıtı</b> , ilçe, inşa, <b>ve</b> , dan, ile, diğer, ol, edilen, aldı, mahale, var, bol, al, katıl, 15 Eylül 1923, 2315	<b>Geçip/geçmeme durumu:</b> [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ..., 1, 0, 0, 0, 1]  <b>Geçme sayısı:</b> [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, <b>2</b> , 0, 0, 0, 0, <b>2</b> , ..., 1, 0, 0, 0, 1]  ve, meyve 2 kez geçiyor.



# Doküman Vektörü

- ❑ Bazı dokümanlar çok kısa ve bazı dokümanlar çok uzun olabilir.
- ❑ Dokümanların çok uzun olması içerisinde çok fazla kelime barındırması neyi etkiler? Neden önemlidir?
- ❑ Bir dokümanın uzunluğu içerisinde barındırdığı toplam sözcük sayısı ve farklı sözcük sayısı ile ifade edilir.
- ❑ Doküman vektörü bulunduğundan sonra doküman uzunluğunun hesaplanmasında norm kullanılır.
- ❑  $\sqrt{x_1 * x_1 + x_2 * x_2 + \dots + x_n}$

Doküman / İçerik	Doküman Vektörü
Edirne'de Meriç, Tunca ve Arda nehirleri etrafında 2315 parça bakımlı sebze, meyve ve dut bahçeleri vardır. Kayısı, erik, ayva, dut, muşmula ve diğer meyveleri boldur. Arda veya Arda Nehri güney Bulgaristan'da Rodop dağlarından doğar ve Yunanistan - Türkiye sınırında, Yunanistan topraklarında Meriç nehrine karışır. Bugün Kırcaali'den geçen Arda nehri üzerindeki en büyük köprü eğlence merkezine dönüştü. Arda nehri nereden doğar nereye dökülür? Tunca hangi akarsuyun kolu? Aras nehri nereden geçiyor? Meriç ve Tunca Nehri nerede?	$[1, 1, 1, 1, 5, 1, 1, 1, 1, 1, 1, 0, 0, 0, 2, 2, 1, 1 \dots, 1, 0, 3, 5, 1]$  Doküman uzunluğu (norm): $  x  _2 = \sqrt{14 + 58 + 9} = 9$



# Sınıflandırma

---

- ❑ Dokümanlar birçok farklı şekilde sınıflandırılabilir, örneğin: konulara göre
  - duygu duruma göre
  - gereksiz yada gerekli olma durumuna göre
  - içerisindeki dile göre türkçe, japonca gibi
  - nefret söylemi barındırıp barındırmamasına göre
  - gerçek bir sosyal medya kullanıcısı olup olmama durumuna göre

# Sınıflandırma

---

- ❑ Yapılan sınıflandırmada bir doküman birden fazla farklı sınıfa ait olabilir. Buna **çok etiketli çoklu sınıflı** sınıflandırma denir.
- ❑ Eğer bir doküman sadece birden fazla sınıf içerisinde sadece bir tanesine ait ise o zaman buna **çok sınıflı** sınıflandırma denir.
- ❑ Eğer bir doküman 2 sınıftan sadece birine ait ise o zaman buna **ikili sınıflandırma** denir.

2010

# K - en yakın sınıflandırma

---

- ❑ Bir dokümanın kendisine ait en yakın dokümana atanarak sınıflandırılmasına k-nn sınıflandırma denir.
- ❑ K değeri dokümana en yakın olan k adet komşuyu temsil eder. Bu komşular içerisinde dokümana en yakın olan sınıf yoğunluğu ile dokümanın sınıfı bulunabilir.
- ❑ Bazı durumlarda k-nn sınıflandırmada sınıfı temsil eden tüm dokümanların ortalaması kullanılır. Buna en yakın merkez yaklaşımı denir. Bir sınıfın merkez vektörüne en yakın olan doküman o sınıfa aittir.

2010

# Yakınlık hesabı

---

İki dokümanın birbirine olan uzaklığı vektörler üzerinden hesaplanır. Temelde iki farklı yakınlık hesabı vardır.

- ❑ **Orantısal yakınlık hesabı:** Dokümanın büyüklüğü göz önünde bulundurulmadan içerisinde geçen kelimelerin ve bu kelimelerin geçme sayılarının orantısal olarak birbirine benzer olup olmadığını kıyaslar.
- ❑ **Mesafe hesabı:** Euclid yada benzeri bir uzaklık hesabı kullanarak mesafesel uzaklığını hesaplar.

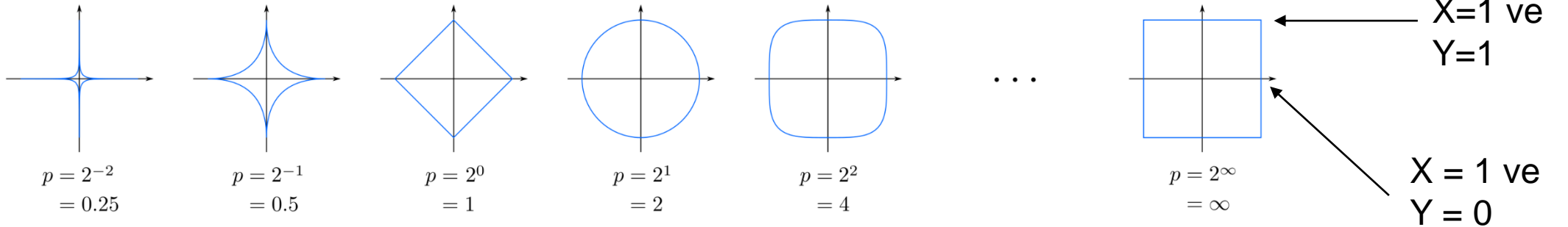
2010

# Minkowski Uzaklığı

- ❑ Genel olarak Euclid uzaklığı ve Manhattan uzaklığının bir türevidir.
- ❑ X ve Y dokümanları için verilen X ve Y vektörleri yukarıdaki formülasyonda p değeri 1 için Manhattan ve p değeri 2 için Euclid uzaklığını göstermektedir. Tüm p değerleri için Minkowski uzaklığı aşağıdaki geometrik şekil çevresi üzerindeki tüm farklı değerleri alabilir.

$$D(X, Y) = \left( \sum_{i=1}^n |x_i - y_i|^p \right)^{\frac{1}{p}}.$$

$$X = (x_1, x_2, \dots, x_n) \text{ and } Y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$$



# Manhattan Uzaklığı

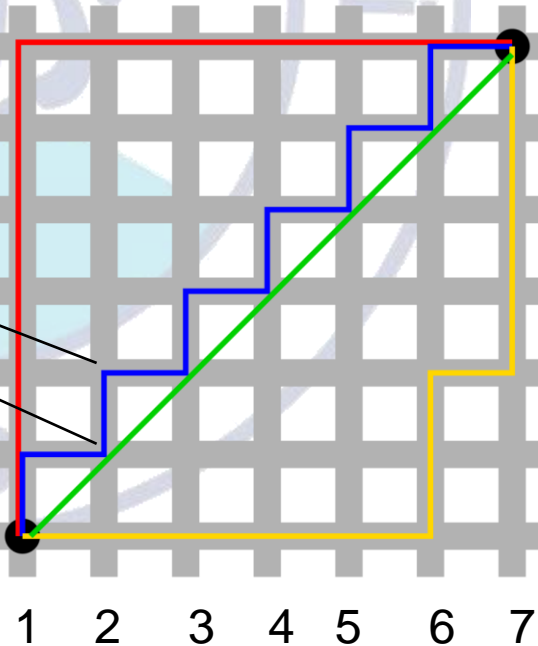
- ❑ New York'daki Manhattan bölgesinde bulunan iki nokta üzerinde yollar üzerinden taksi ile gidilerek ölçülen mesafedir. Bu bir düzlem üzerinde aşağıdaki şekilde ifade edilebilir.

$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|,$$

i=2 için  
p<sub>i</sub>=2 ve q<sub>i</sub>=3

P noktası

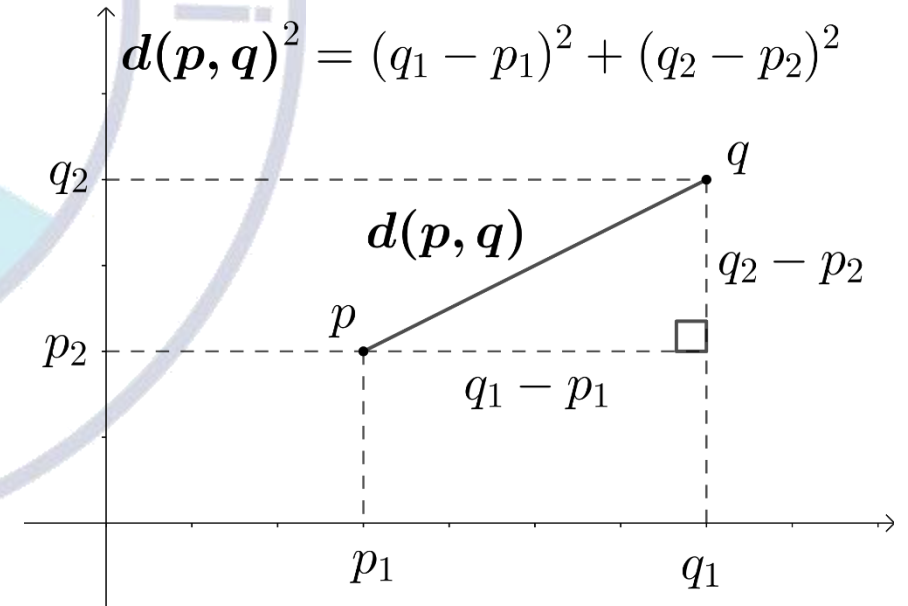
Q noktası



# Euclid Uzaklığı

- ❑ Pisagor teoreminde hesaplanan hiptenüs mesafesidir. Ancak bu hesaplama çok boyutlu vektörler için aşağıdaki gibidir.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_i - q_i)^2 + \dots + (p_n - q_n)^2}.$$





# Mahalanobis uzaklığı

- ❑ Mahalanobis mesafesi bir doküman vektörü ile bir sınıfa ait merkez vektörü arasında aşağıdaki gibi hesaplanır.
- ❑ S vektörlerin her biri için kovaryans hesabını ifade eder.
- ❑ Kovaryans nedir? Nasıl hesaplanır? Neyi ifade eder?

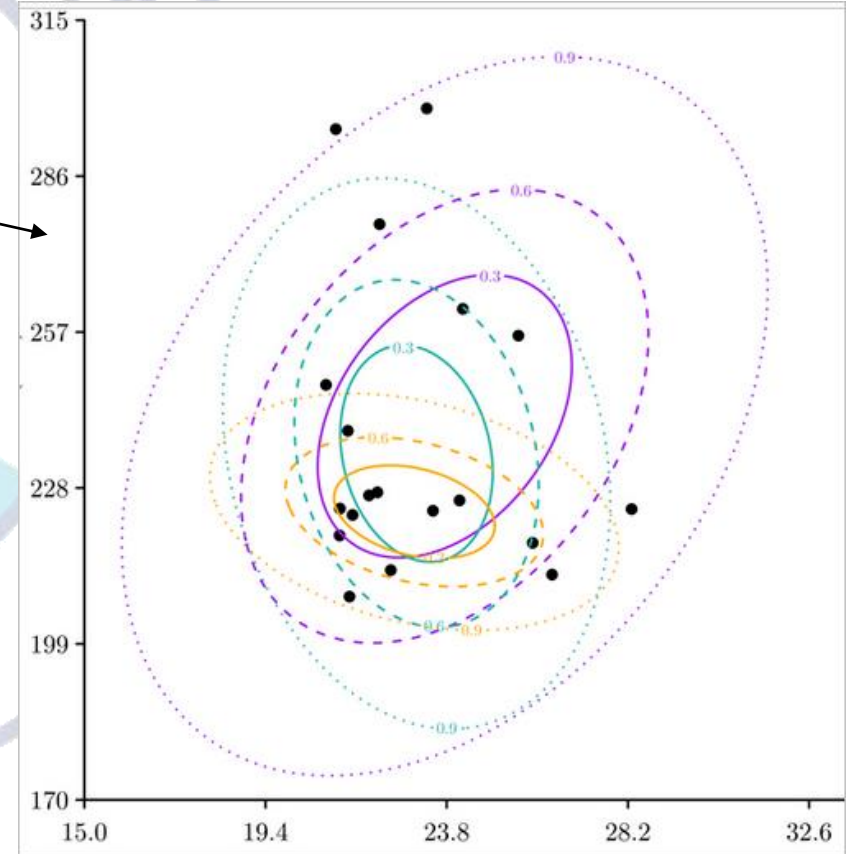
$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \mathbf{S}^{-1} (\vec{x} - \vec{\mu})}.$$

# Mahalanobis uzaklığı

x1 ve x2 değerleri  
için kovaryans

$$\text{cov}(X, Y) = \sum_{i=1}^n p_i (x_i - E(X))(y_i - E(Y)).$$

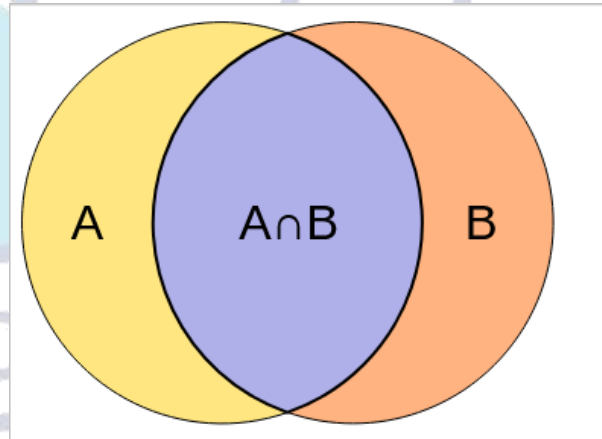
Merkez vektörü



# Jaccard index

- ❑ Jaccard index ile iki vektörün sahip olduğu ortak eleman sayısı hesaplanır.
- ❑ Örneğin:  $X=[1, 0, 1, 1, 0]$  ile  $Y=[1, 0, 1, 0, 1]$  vektörlerinde ortak eleman indisleri: 1. ve 3. indislerdir. Bu durumda  $|A \cap B| = 2$  ve  $|A \cup B| = 4$  olacaktır.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



# Cosinüs uzaklığı

- ❑ A ve B vektörleri için cosinüs benzerliği yanda verilmiştir.
- ❑ Cosinüs ile orantısal uzaklık ölçülür. Örneğin türkiye, anıtkabir, elma dağ kelimelerinin geçtiği iki doküman vektörü aşağıda verilmiştir.
- ❑  $X = [1, 1, 1]$  ve  $Y = [1, 5, 5]$  ve  $Z = [3, 15, 15]$  dokümanları arasındaki
- ❑ cosinüs benzerliği X ve Y için 0.88 ve Y ve Z için 1.0 çıkmaktadır.
- ❑ euclid uzaklığı X ve Y için  $\sqrt{32}$ , ~5.65 ve Y ve Z için 14.28
- ❑ Bu cosinüs farkını uzaklığa dönüştürmek için  $(1 - \cos) = 0.22$  ve 0.0 değerleri elde edilir.
- ❑ Bu durumda cosinüs içim Y ve Z bir birine çok yakın (0.0) iken euclid için uzaklık X ve Y nin kinden (14.28) çok daha fazladır.

$$\frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

# K-nn sınıflandırıcı

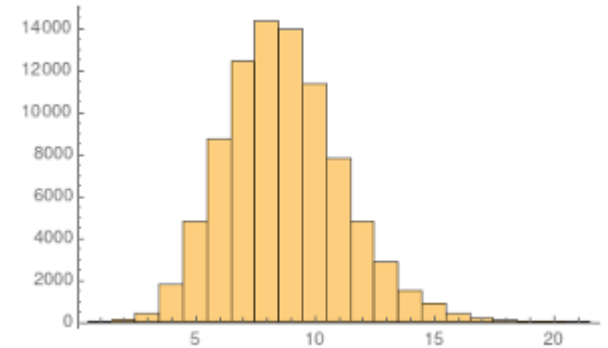
---

- ☐ Cosinus yakınlık hesabı kullanarak aşağıdaki T vektörünü doğru sınıfa atayınız.
- ☐ Her bir vektörün sınıfı: [X,Spor], [Y, Politika], [Z, Güncel], [A,Spor], [B, Politika], [C, Güncel]
- ☐  $\text{cosinus}(T,X) = 0.5$ ,  $\text{cosinus}(T, Y) = 0.6$ ,  $\text{cosinus}(T, Z)=0.3$ ,  $\text{cosinus}(T, A) = 0.6$ ,  $\text{cosinus}(T, B) = 0.4$ ,  $\text{cosinus}(T, C) = 0.4$
- ☐ K=1 için T hangi sınıftadır, K=4 için T hangi sınıftadır.
- ☐ Merkez yakınlık hesabı kullanılmış olsaydı ne olurdu?

# Naive Bayes sınıflandırıcı

- ❑ Naïve Bayes en sık kullanılan sınıflandırıcıdır. Bu sınıflandırıcıda her bir kelimenin belirli bir sınıfa ait olasılığı kullanarak her bir sınıfı ait olma olasılıklarını özellik dağılımından üretir. Dolayısıyla naive Bayes üretici sınıflandırıcı türünde bir makine öğrenmesi yöntemidir.
- ❑ Eğer dokümandaki tüm kelimelerin bir sınıfa ait olma olasılıkları çarpımı yüksek ise o zaman bu doküman o sınıfa aittir denir.
- ❑ Bu prensibe maksimum sonsal olasılık (maximum a posterior probability) – MAP denir. Buna göre tüm sınıfa ait olma olasılığı en yüksek olan sınıf doğru sınıf olarak kabul edilir.

$$p(c|x) = P(c) * \prod_{i=1} \frac{p(x_i|c)}{p(x_i)}$$
$$freq_c / total * \prod_{i=1} freq(x_i, c) / freq(x_i)$$





# Naïve Bayes sınıflandırıcı

	class	SAVAŞ	ZAFER	GEL	KAR	ZAM	DOLAR	ORAN	DAĞ	SU	ÇAM	TOPLAM
D1	A	1	1	1	0	0	0	0	0	1	0	5
D2	B	0	1	0	1	1	1	1	0	0	0	4
D3	C	0	0	0	0	0	0	1	1	1	1	2
D4	B	0	0	0	1	0	1	1	0	1	0	4
D5	B	0	1	0	0	1	0	1	0	0	0	4
D6	A	0	0	1	0	0	0	1	1	1	0	5
D7	A	1	0	1	0	0	0	0	0	1	0	5
D8	A	0	0	0	0	0	0	0	0	0	0	5
D9	A	1	0	1	0	0	0	0	1	0	0	5
D10	B	0	1	1	1	1	1	1	0	1	0	4
D11	C	0	0	0	0	0	1	1	1	1	1	2
Toplam		3	4	5	3	3	4	7	4	7	2	



# Naive Bayes sınıflandırıcı

Naive bayes ile;

$$\square p(\text{SAVAŞ} | A) = \#(\text{SAVAŞ} \& A) / \#(A) = 3.0 / 5.0$$

$$\square p(A) = \#A / \text{Total} = 5.0 / 11.0$$

$$\square p(\text{SAVAŞ}) = \#(\text{SAVAŞ}) / \text{Total} = 3.0 / 11$$

$$\square p(A | \text{SAVAŞ}) = P(A) * P(\text{SAVAŞ} | A) / P(\text{SAVAŞ}) = 5.0/11.0 * 3.0/5.0 * 11.0/3.0$$

$p(A | D)$  = Doküman içerisindeki tüm kelimelerin A kategorisine ait olma olasılıkları çarpımı  
=  $P(A|\text{SAVAŞ}) * P(A| \text{ZAFER}) * P(A| \text{GEL}) \dots P(A| \text{ÇAM})$

Peki bazı kelimeler D dokümanı içinde geçiyor fakat A kategorisinde hiç geçmiyorsa?

# Naive Bayes sınıflandırıcı

---

- ❑ Bu durumda Naive Bayes sınıflandırıcı yumuşatma yada düzleme yöntemi (smoothing) kullanılır.
- ❑ Yumuşatma ile 0 yada sonsuz olan olasılık değerleri küçük sayılara yuvarlanır.
- ❑ Örneğin:  $\#(\text{SAVAŞ} \ \&\& \ B) + 1.0 / \#(B) + 1.0$
- ❑ Peki olasılık çarpımlarındaki en önemli problem nedir?
- ❑ Olasılık 1.0 değerinden küçük olduğu için çarpım uzunluğu arttıkça değer küçülür ve bunu engellemek için negative logaritma kullanılır ve çarpım toplama dönüşür.
- ❑  $-\log(a * b * \dots) = -\log(a) - \log(b) \dots$

# Kelime Çantası

- ❑ Tüm gördüğümüz ağırlıklandırma, sınıflandırma ve uzaklık yöntemlerinde kelimeler arasındaki ilişki kullanılmadı.
- ❑ Bu yöntemlere genel olarak özellikler (kelime/terimler) bağımsız özdeş dağılıma sahiptir. Kısaca kelime sınıf ve doküman dağılımları birbirinden bağımsızdır veya birbirini etkilemez.
- ❑ Bağımsız olma durumunda : örneğin; kale kelimesi ve futbol kelimesin savaş kategorisinde olan dokümanlarda geçme durumları birbirinden bağımsız ise o zaman geçme olasılıkları için aşağıdaki durum oluşur.

$$P(futbol \cap kale | c = savaş) = P(futbol | c = savaş) * p(kale | c = savaş)$$

- ❑ İki olayın birbirinden bağımsız değilse bu durumda bağımlı durum oluşur. Örneğin yukarıdaki örnek için kategori spor olsun. Bu durumda futbol ve kale kelimelerinin kategori içerisinde birlikte geçme olasılıkları aşağıdaki gibi hesaplanabilir.

$$P(futbol \cap kale | c = spor) = P(futbol | c = spor, w = kale) * p(kale | c = spor)$$

# Ağırlıklandırma

- ❑ Terimlerin doküman içerisindeki geçme sıklığı terim ve doküman arasındaki ilişkiyi belirler.
- ❑ Bazen terimlerin dokümaları temsil etmesi mümkün olmaz. Örneğin fonksiyon kelimeleri (stop words) olan ve, veya, nasıl gibi sık geçen kelimeler dokümanları temsil etmeyebilir. Bu durumda terimin doküman içerisindeki ağırlığının bulunması sınıflandırma sonucunda çok etkilidir.
- ❑ Standard yöntemde kelimeler geçip geçmeme ile ifade edilir.
- ❑ İkili ağırlık (var/yok) = [0, 1, 1, 1, 0, 0, 0,..., 1]
- ❑ Frekans ağırlık (kaç defa) = [0, 2, 1, 1, 0, 0, 0, ..., 7]
- ❑ Frekans oran (norm) = frekans/doküman uzunluğu = [0, 2.0/11, 1.0/11, 1.0/11.0,..., 7.0/11.0]
- ❑ Frekans log =  $\log(\text{frekans}+1.0)$
- ❑ Ters doküman frekansı (IDF): (toplam doküman sayısı / terimin kaç dokümanda geçtiği) =  $N / df$
- ❑  $TF*IDF$  = Frekans ağırlık \* ters doküman frekansı

2010

# Örnek

	class	KAVŞAK	DENİZ	DAĞ
D1	A	1	3	5
D2	B	0	2	0
D3	C	3	0	4
D4	B	0	0	0
D5	B	0	1	0
D6	A	1	0	4
Toplam		3	4	5

☐ Ters doküman ağırlığı (IDF) kullanarak aşağıdaki dokümanları içerisinde geçen DAĞ terimini her bir doküman için ağırlıklandırınız.

☐ Dokümanlarda sadece 3 terim olduğunu kabul ediniz.

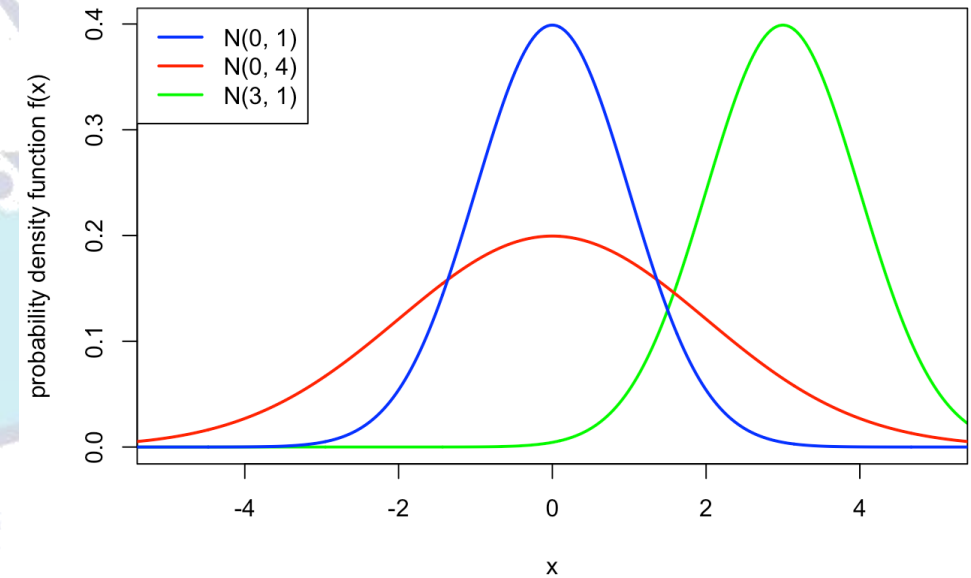
# Tek deęişkenli normal daęılım

Normal daęılım çoęu zaman karřımıza çıkan bir daęılım fonksiyonudur.

Genelde an eęrisi daęılımı olarak da bilinmektedir.

Her bir kelime aęırlıklandırıldıktan sonra üç farklı kategori için yandaki gibi üç farklı normal daęılım ile modellenenebilir.

Burada  $x$  kelimesinin geme ortalaması yeřil kategori ile belirtilen durum için 3 tür. Buna göre yeřil kategoriden bir doküman seçtięimde kelime aęırlığı en yüksek ihtiamalle üç olacaktır.





# Tek deęişkenli normal dağılım

Yandaki formül belirli bir noktada normal dağılım üzerinde x girdisi için dağılımın olasılık deęerini vermektedir. Burada  $\mu$  ortalama,  $\sigma$  standard sapma ve  $\pi$  (pi) 3.14 sabit olarak alınmalıdır. Örneęin spor dokümanları için **futbol** kelimesi aşığıdaki gibi  $f(x)$  idf deęerler alsın;

Doküman 1: 5.0  
Doküman 2: 3.75  
Doküman 5: 1.25  
Doküman 8: 2.0

Bu durumda spor dokümanları için ortalama 3.0 olacaktır. Bu durumda yandaki formülde ortalama  $\mu = 3$  tür. Ve standard sapmada yandaki formüle göre ve  $x_i$  dokümanda geęme miktarını belirtirse;

$$\sqrt{((5 - 3)^2 + (3.75 - 3)^2 + (1.25 - 3)^2 + (2 - 3)^2) / 4}$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$



# Tek değişkenli normal dağılım

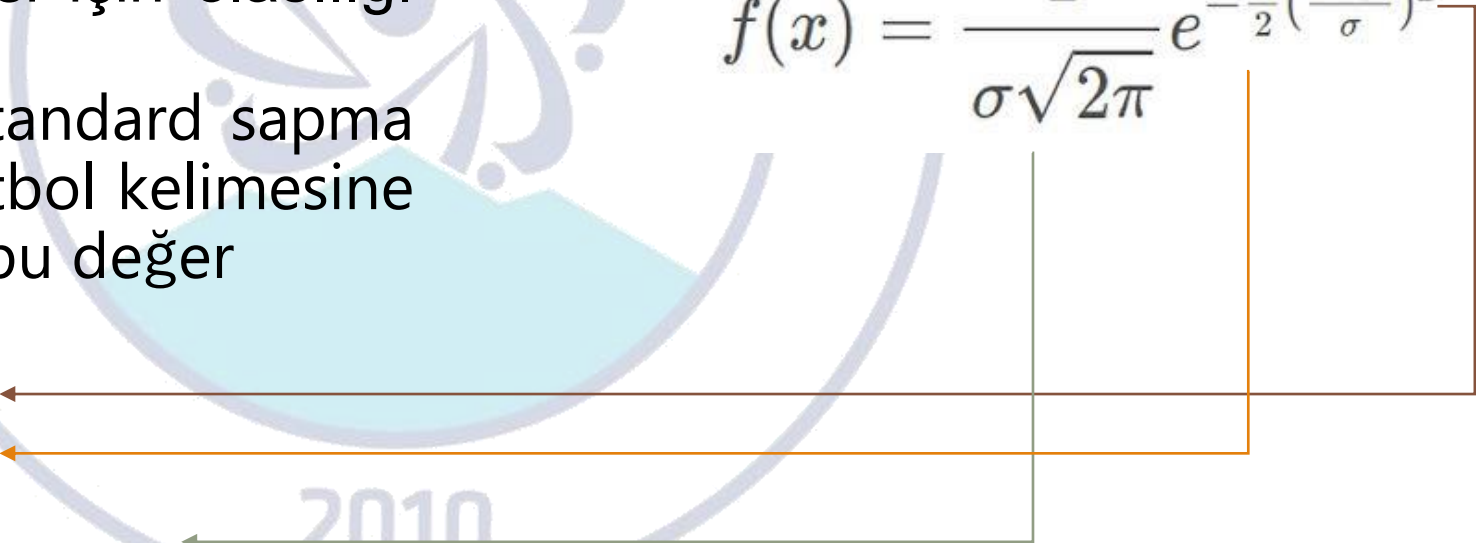
Bu durumda yandaki formülü kullanarak spor kategorisi için olasılığı hesaplayalım.

Eğer ortalama  $\mu=3$  ve standard sapma  $\sigma=2$  ise dokümandaki futbol kelimesine ait tf x idf 3 ise o zaman bu değer

Adım 1:  $((3 - 3)/2)^2 = 0$

Adım 2:  $e^{-1/2 \times 0} = e^0 = 1$

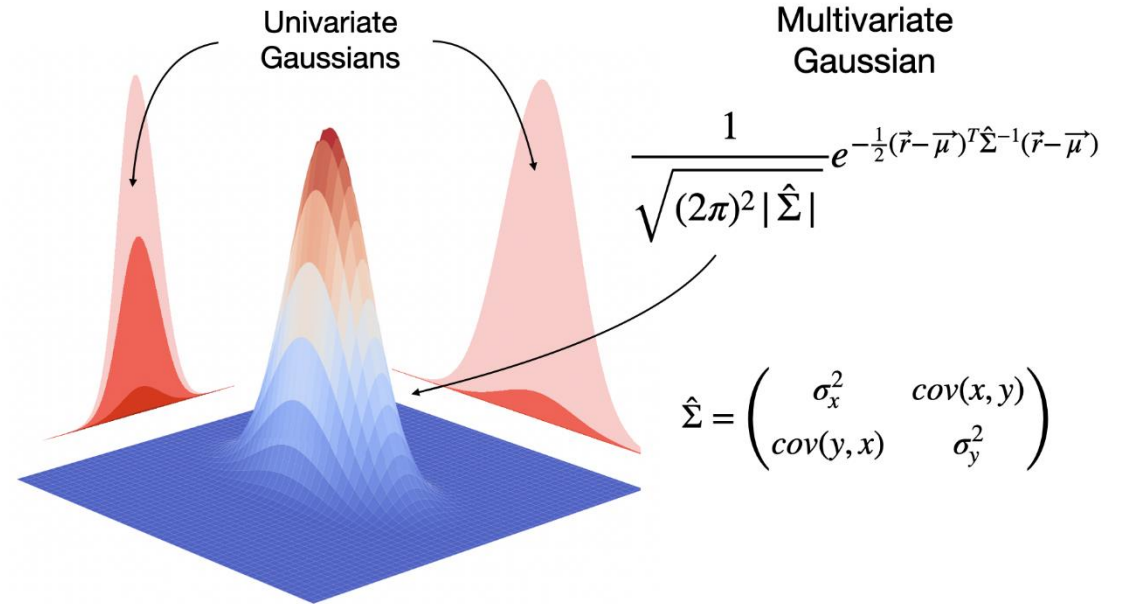
Adım 3:  $f(x) = 1 / 5.011 = 0.19$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$


# Çok değişkenli normal dağılım

Çoğu zaman kelimeler arasında belirli bir bağımlılık vardır ve bir kelimelerin geçmesi diğer bir kelimenin geçme olasılığını yada kaç kere geçeceğini etkiler.

Böyle bir durumda çok değişkenli bir olasılık dağılım fonksiyonu olan çok değişkenli normal dağılım (multivariate gaussian) fonksiyonunu kullanmamız daha doğru sonuçlar üretecektir.



# Çok değişkenli normal dağılım

Çok değişkenli normal dağılımda kovaryans matrisi hesabı yapmamız gerekir.

Kovaryans matrisi değişkenlerin yada kelimelerin diğer gelişmelerle geçme frekanslarının ortalamadan ne kadar sapıp saptığını belirten bir matristir.

Kovaryans matrisi her değişken için diğer değişkenlerle arasında hesaplanır.

Kovaryans hesabında  $x_i$  doküman vektörünü ve  $\bar{X}$  ise belirli bir kategoriye ait olan dokümanların vektör ortalamasını göstermektedir. Bu kategoride toplam N adet doküman yer almaktadır.

$$\text{Cov}(A) = \begin{bmatrix} \frac{\sum (x_i - \bar{X})(x_i - \bar{X})}{N} & \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N} \\ \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{N} & \frac{\sum (y_i - \bar{Y})(y_i - \bar{Y})}{N} \end{bmatrix}$$

$$= \begin{bmatrix} \text{Cov}(X, X) & \text{Cov}(Y, X) \\ \text{Cov}(X, Y) & \text{Cov}(Y, Y) \end{bmatrix}$$

# Örnek Kovaryans hesabı

- Sözlüğümüzün 3 adet kelime içerdiğini düşünelim.
- Bu durumda üç döküman için aşağıdaki gibi 3 vektör verebiliriz.
- Birinci kelimenin geçme frekansları ortalaması  $(2+3+4)/3 = 3$  olacaktır. Benzer şekilde diğer kelimelerde aşağıdaki gibi hesaplanabilir.

Belge 1: [2, 3, 4]

Belge 2: [3, 4, 5]

Belge 3: [4, 5, 6]

$$\bar{X} = \frac{2+3+4}{3} = 3$$

$$\bar{Y} = \frac{3+4+5}{3} = 4$$

$$\bar{Z} = \frac{4+5+6}{3} = 5$$

- Burada 1. indeks 1. kelmeyi ifade etmektedir.

# Örnek Kovaryans hesabı

- Sözlüğümüzün 3 adet kelime içerdiğini düşünelim.
- Bu durumda üç doküman için aşağıdaki gibi 3 vektör verebiliriz.
- Ortalamalar bulunduktan sonra kovaryans matrisinde 3 kelime için birbiri arasında ortalamadan nasıl değiştiğini hesaplarız.
- Tüm dokümanlar için 1. kelimenin 2. kelime ile birlikte değişimi aşağıdaki gibi toplamın ortalaması şeklinde hesaplanır.

Belge 1: [2, 3, 4]

Belge 2: [3, 4, 5]

Belge 3: [4, 5, 6]

- Burada 1. indeks 1. kelmeyi ifade etmektedir.

$$\text{Cov}(X, Y) = \frac{(2-3)(3-4) + (3-3)(4-4) + (4-3)(5-4)}{3-1}$$

$$\text{Cov}(X, Y) = \frac{(-1)(-1) + (0)(0) + (1)(1)}{2}$$

$$\text{Cov}(X, Y) = \frac{1+0+1}{2}$$

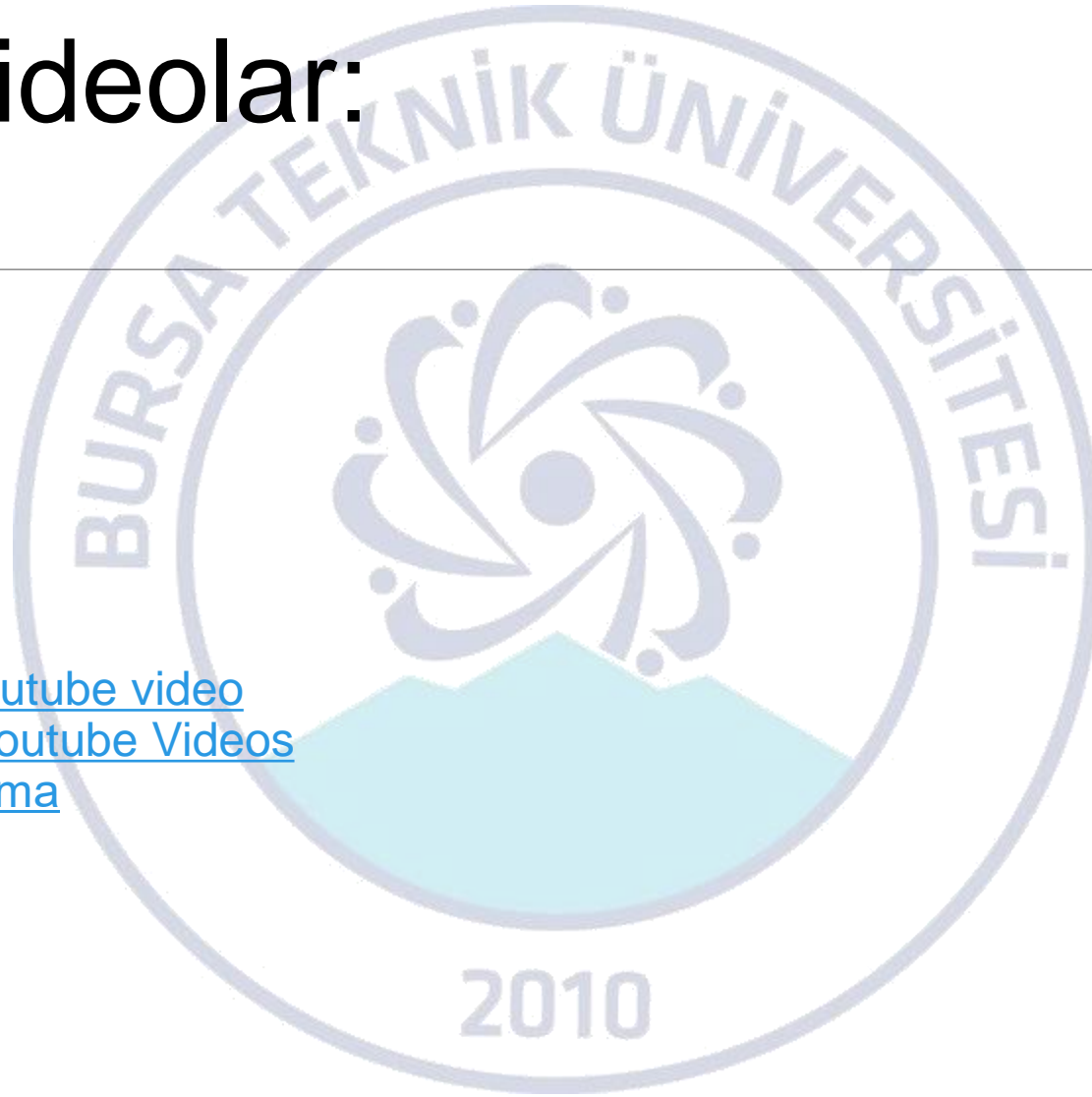
$$\text{Cov}(X, Y) = \frac{2}{2}$$

$$\text{Cov}(X, Y) = 1$$

# Örnek videolar:

---

- ☐ [Naive bayes - Youtube video](#)
- ☐ [Terim TF\\*IDF - Youtube Videos](#)
- ☐ [Örnek sınıflandırma](#)





# Referanslar

---

<https://nlp.stanford.edu/fsnlp/>

[https://happyharrycn.github.io/CS540-Fall20/lectures/statistics\\_nlp.pdf](https://happyharrycn.github.io/CS540-Fall20/lectures/statistics_nlp.pdf)

2010