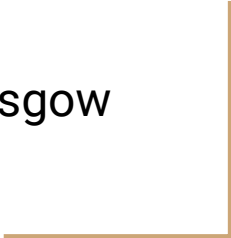



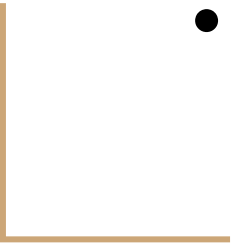


# **Doğal Dil İşleme**

## Kısa Bir Giriş

Alara Dirik  
University of Glasgow



- 
- Doğal Dil İşleme ürünleri ve kullanım alanları
  - Cümlelerin anlamlandırılması
  - Türkçe'ye özgü sıkıntılar
  - Türkçe veri setleri ve açık kaynak projeler
  - Varlık İsmi Tanıma (workshop)
  - Metin Kümeleme (workshop)
- 

# Doğal Dil İşleme Ürünleri ve Kullanım Alanları

# Doğal Dil İşleme Nedir?

**Kısaca:** Yazılı ve sesli metinlerin anlamlandırılması

- İnsanlar arası iletişim
- İnsan ve makine arasında iletişim

**Nasıl?**

- İstatistik
- Makine öğrenmesi
- Dilbilim
- Kural bazlı yaklaşımlar
- Yazılım

# Kullanım Alanları

İngilizce	↔	Türkçe
Neural machine translation (NMT) is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modeling entire sentences in a single integrated model.	×	Nöral makine çevirisi (NMT), bir kelime dizisinin olasılığını tahmin etmek için yapay bir sinir ağı kullanan ve genellikle tüm cümleleri tek bir entegre modelde modelleyen makine çevirisine bir yaklaşımdır.

**Source Text:** Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

**Summary:** Elizabeth was hospitalized after attending a party with Peter.



**Hız: 10 | Servis: 10 | Lezzet: 9**

...7 Patatesler ince geldi ve çok soğuktu ama burger efsane.

# Kullanım Alanları

- Çeviri - Neural Machine Translation
- Metin Bazlı Tavsiye Sistemleri
- Metin Sınıflandırma
  - Duygu Analizi - Sentiment Analysis
- Ses İşleme
- Özet Çıkarma
- Bilgi Çıkarımı
  - Varlık İsmi Tanıma
  - Kelimeler Arası Bağlılık
  - Soru Cevaplama
  - Soru Oluşturma
  - Benzer Metinleri Bulma
- Metin Kümeleme
- Otomatik Metin Üretme




Cümleleri Nasıl Anlamlandırırız?



# Doğal Dil İşleme Neden Zor?

- Mantıksız ve çelişkili cümleler
- Yazım hataları ve sistematik hatalar
- Deyimler
- Çok anlamlı kelimeler
- Yoruma açık cümleler
- Bağlam içinde anlamlı olan söz öbekleri ve cümleler

Hız: 10 | Servis: 10 | Lezzet: 9

 ...7 Patatesler ince geldi ve çok soğuktu ama burger efsane.

Bu interneti yavaslatinca noluyo!!!! Butun isleri net uzerinden olan insanalari magdur ediosunuz!!! E yeter artik acin ya!! Nefret ettirdinz

"Onu tanımasam çok mutsuz biri derdim."



Bülent Ecevit kimdir? Kaç yılında doğmuştur?

Kim?



# Metin Temizleme ve İşleme

[Colab Linki](#)

- Sistematik hataların düzeltilmesi

*"Hava çok güzel!" -> "Hava çok güzel!"*

- Metni standart hale getirme: küçük harfe çevirme, ünlem işaretlerinin çıkarılması

*"Hava çok güzel!" -> "hava çok güzel"*

- Normalizasyon

*"Dünde böyle güneşli bi gündü." -> "Dün de böyle güneşli bir gündü."*

- Yüksek frekanslı sözcüklerin çıkarılması

*"Dün de böyle güneşli bir gündü." -> "Dün de böyle güneşli gündü."*

- Kelimeleri kök haline çevirme

*"dün de böyle güneşli bir gündü" -> "dün de bu gün gün" →*

Her zaman iyi bir fikir olmayabilir!

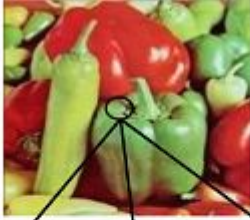
# Kelimeleri Anlamlandırma

## ASCII

a	97	n	110
b	98	o	111
c	99	p	112
d	100	q	113
e	101	r	114
f	102	s	115
g	103	t	116
h	104	u	117
i	105	v	118
j	106	w	119
k	107	x	120
l	108	y	121
m	109	z	122

- Kelimeler tek başına bir şey ifade ediyor mu?
- Her kelime aynı derecede önemli mi?
- Modellediğimiz dilde kaç kelime var?
- Kelime türetebilir miyiz?
- Yazılımsal ve donanımsal kısıtlamalar

# Kelimeleri Anlamlandırma



Kelimeleri makinelerin anlayabileceği bir forma sokmalıyız.

- Dil modelleri:
  - Kelimelere olasılık, frekans veya skor atar
  - Kelimelerle sınırlı değil, kelime grubu, kelime sıralamaları, cümle ve döküman bazında modeller mevcut

240 241 241	207 199 196	234 231 225
240 237 238	183 163 195	223 213 225
239 240 240	183 166 184	219 211 195
238 237 240	176 172 181	176 205 189
240 240 239	184 167 176	168 141 117
239 240 240	182 180 170	160 142 117

# Popüler Dil Modelleri ve İşleme Metodları

# One-Hot Encoding

## Vocabulary

index:	Word:
0	aardvark
1	able
...	...
2409	black
2410	bling
...	...
3202	candid
3203	cast
3204	cat
...	...
5281	is
5282	island
...	...
8676	the
8677	thing
...	...
9999	zombie

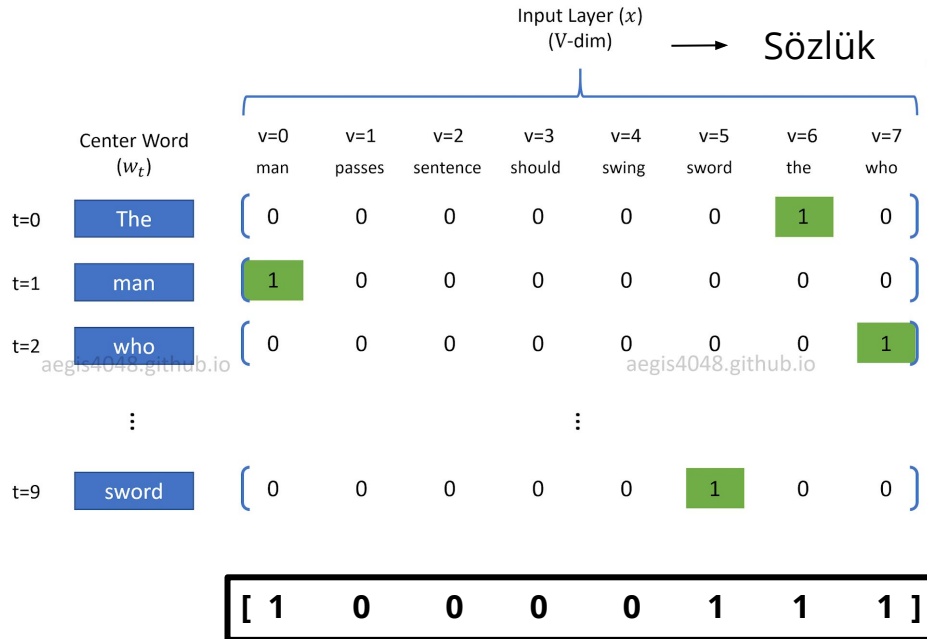
the cat is black

- Basit
- Yorumlaması kolay

Ancak:

- Devasa ve seyrek vektörler
- Ayrık - bağlamsal anlamı yok

# BoW: Bag of Words



One-hot vektörleri sütun bazlı toplamı

- Devasa ve seyrek vektörler
  - Vektör uzunluğu = sözlük uzunluğu
- Kelime sırasının ve cümledeki yerinin önemi yok
- Kelime anlamını kaydetmiyor

→ BoW vektörü

# N-grams

Kelimelerin grup bazlı olasılığını  
hesaplar: 2-grams, 3-grams, vs.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

2-grams modeli:

Birinci kelime "*Okula*" ise ikinci kelimenin "*gidiyorum*" olma olasılığı nedir?

- Yazım düzeltme, otomatik yazı tamamlama (auto-completing), metin üretimi, ses tanıma
- Kullanılan veri setine bağımlı
- Ayırık - bağlamsal anlamı yok

# TF-IDF

$$w_{x,y} = \text{tf}_{x,y} \times \log \left( \frac{N}{\text{df}_x} \right)$$

## TF-IDF

Term  $x$  within document  $y$

$\text{tf}_{x,y}$  = frequency of  $x$  in  $y$

$\text{df}_x$  = number of documents containing  $x$

$N$  = total number of documents

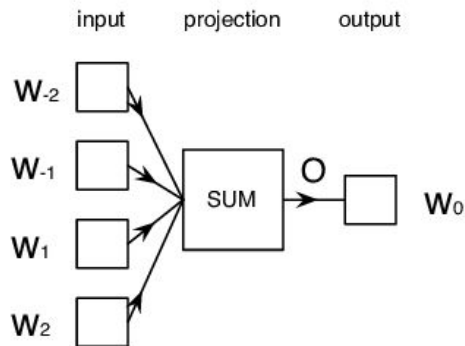
Term Frequency - Inverse Document Frequency

- Nedir
  - Sıralama algoritması
  - Farklı konularda dökümanlardan oluşan bir veri setinde her kelimenin her konu için önemini hesapla
- Nasıl
  - Kelime döküman içerisinde kaç kere kullanılıyor?
  - Kelime kaç dökümanda geçiyor?



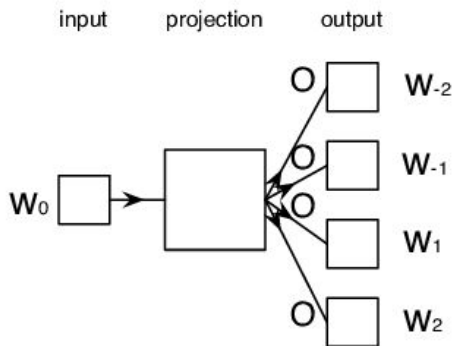
# Word2Vec

## CBOW



Bugün      ...      gitmedim  
            *okula*  
            *işe*

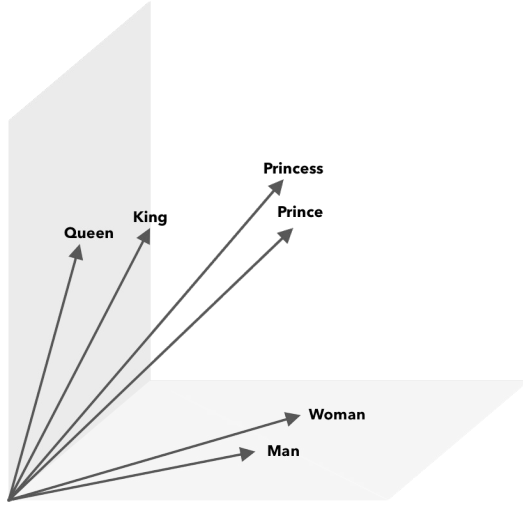
## Skip-Ngram



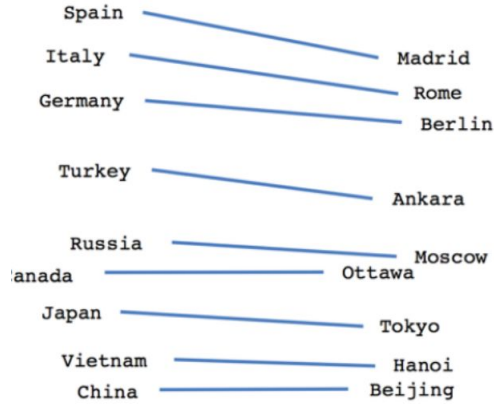
            ...      okula      ....  
*Bugün*                      *gitmedim*  
*Dün*                        *gittim*

- Nedir
  - Kelimelerin bağlamsal olasılığı
  - Benzer kelimeler benzer bağlamlarda kullanılır
- Nasıl
  - Sığ Nöral Ağ (Neural Network) modeli

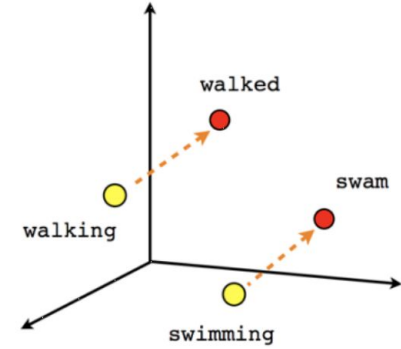
# Word2Vec



Kadın - Erkek



Baş şehir - Ülke



Zaman Ekleri

# Word2Vec

## **CBOW**

- Bir kelimenin bir kelime grubunun ortasında olma olasılığı nedir?

## **Skip-Ngram**

- Bir kelime grubunun (context) bir kelimenin etrafında olma olasılığı nedir?

- Benzer kelimeleri bulabiliyor
- Kompakt vektörler
- Milyonlarca döküman üzerinde eğitilip tekrar kullanılabilir

Ancak

- Kelimenin cümle içinde nerede olduğunun önemi yok
- Çok anlamlı kelimeleri göz ardı ediyor

# BERT

Use the output of the masked word's position to predict the masked word

Possible classes:  
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzyva

FFNN + Softmax



Randomly mask  
15% of tokens

Input

[CLS] Let's stick to improvisation in this skit

Kelimeleri modellemek için hem bağlamını (etrafındaki kelimeler) hem de cümle içindeki yerini kullan.

"Yüz gündür grevdeler."

"Akşama kadar yüz."

-> İki farklı "yüz" vektörü

# Başka Yaklaşımlar

Kelime bazlı, karakter bazlı ve alt-kelime (sub-word) bazlı pek çok yaklaşım mevcut

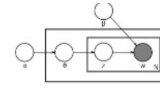
- fastText: <https://fasttext.cc/>
- ELMo: <https://allennlp.org/elmo>
- GloVe: <https://nlp.stanford.edu/projects/glove/>
- GPT-2 <https://github.com/openai/gpt-2>
- Ve daha fazlası...

# Dil Modellerini Nasıl Kullanırız?

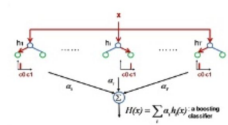
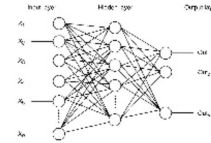
# Dil Modellerini Nasıl Kullanırız?

- Tek başına
  - Kelime, cümle, döküman bazlı benzerlik
  - Keşif, istatistiksel analiz
- Algoritma girdisi olarak
  - İstatistiksel modeller
  - Makine öğrenmesi
    - Derin öğrenme

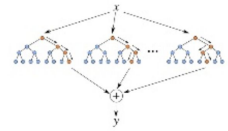
- Logistic Regression
- Elastic Nets
- Gradient Boosted Decision Trees
- Random Forests
- Neural Networks
- LambdaMART
- Matrix Factorization
- LDA
- ...



$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

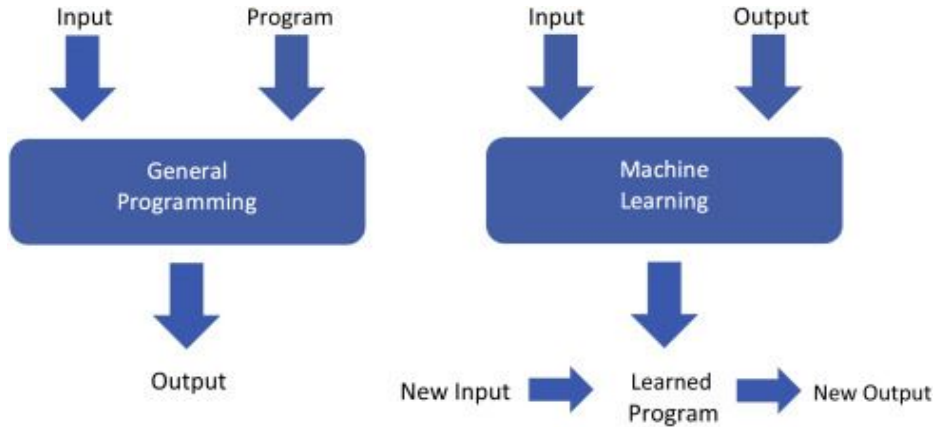


$$\begin{matrix} d \\ n \end{matrix} \begin{matrix} \mathbf{X} \end{matrix} = \begin{matrix} h \\ n \end{matrix} \begin{matrix} \mathbf{U} \end{matrix} \times \begin{matrix} d \\ h \end{matrix} \begin{matrix} \mathbf{V}^T \end{matrix}$$



$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} (\|y - X\beta\|^2 + \lambda_0 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

# Gözetimli Öğrenme

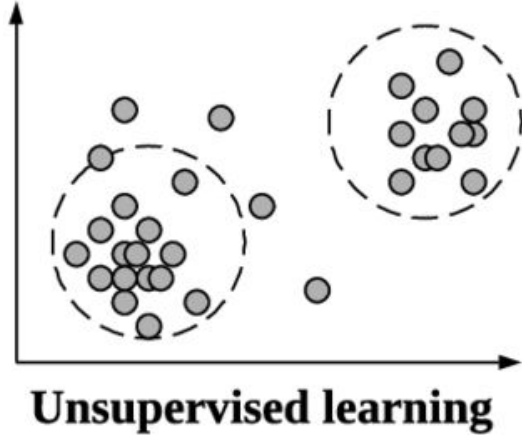


Yeterince girdi ve çıktınız varsa iki grup arasındaki karmaşık kuralları keşfedebilirsiniz.

- Ürün yorumu - etiket (pozitif/negatif)  
-> otomatik sınıflandırma
- İngilizce - Türkçe paralel cümleler  
-> otomatik çeviri



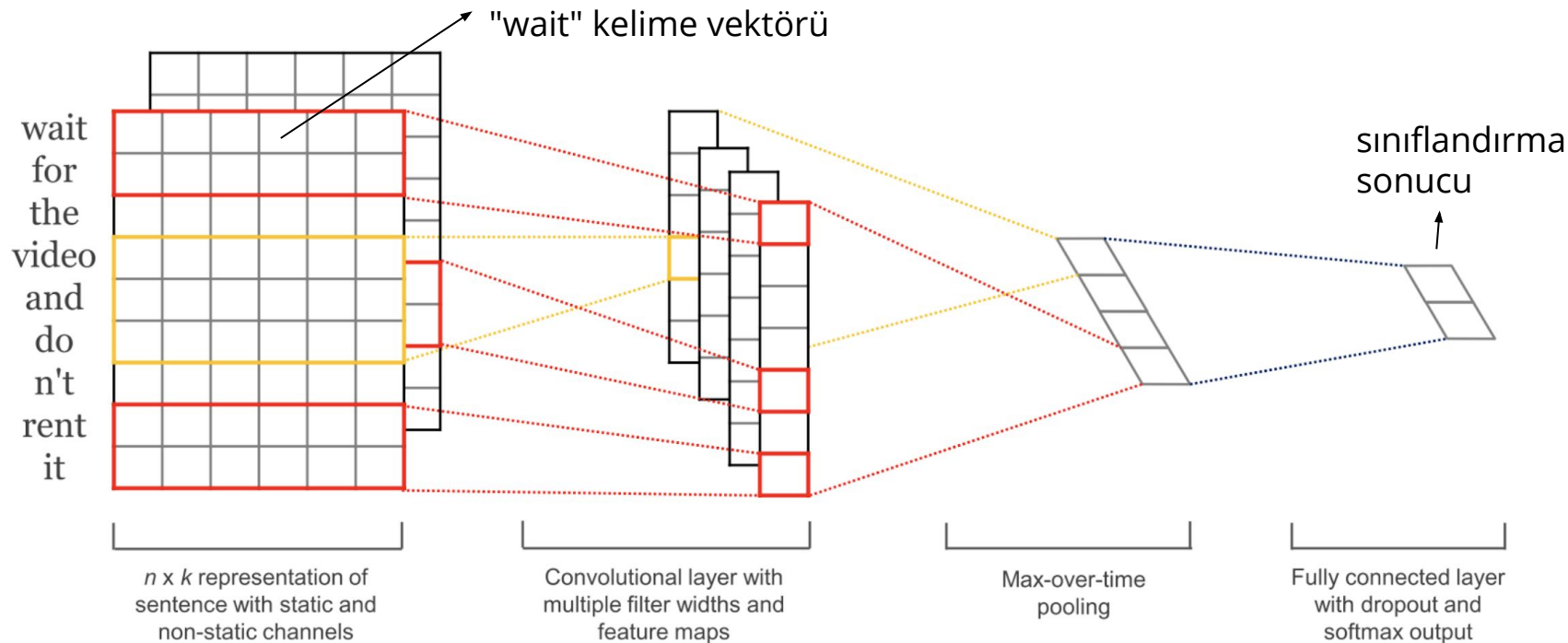
# Gözetimsiz Öğrenme



Çıktılar olmadan, sadece girdileri kullanarak veriyi ve tarif etmek için kullanılır.

- Kümeleme
- Birliktelik kuralları (association rules)
- Temel Bileşen Analizi (PCA)

# CNN: Convolutional Neural Networks



# CNN: Convolutional Neural Networks

1 <sub>x1</sub>	1 <sub>x0</sub>	1 <sub>x1</sub>	0	0
0 <sub>x0</sub>	1 <sub>x1</sub>	1 <sub>x0</sub>	1	0
0 <sub>x1</sub>	0 <sub>x0</sub>	1 <sub>x1</sub>	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

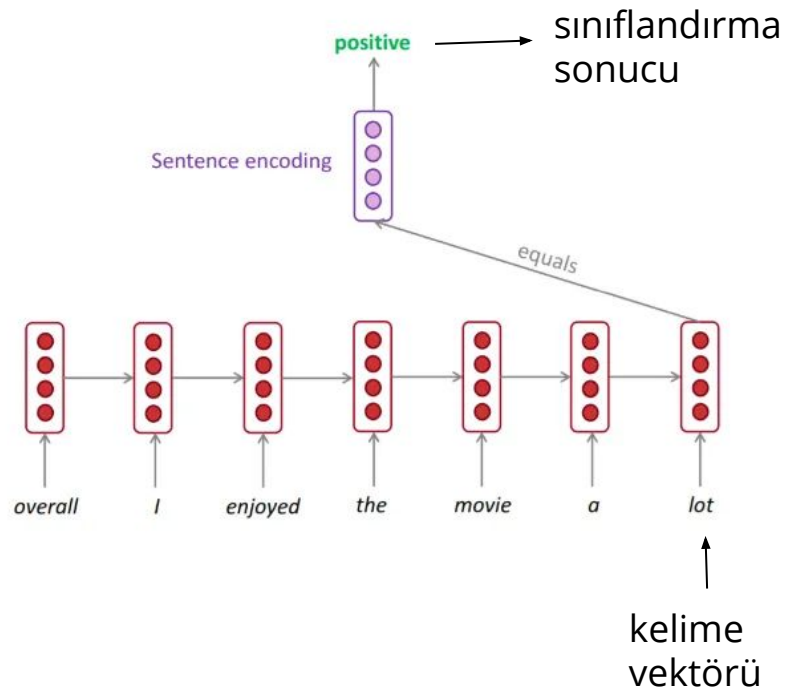
Convolved  
Feature

Doğal Dil İşleme için genelde 3 ve üzeri evrimsel / convolutional katman kullanılır.

Amaç

- En genel hatlardan başlayarak her katmanda daha sofistike temsilleri öğrenmek
- Girdiyi daha kompakt hale getirerek bilgi 'özeti' çıkarmak

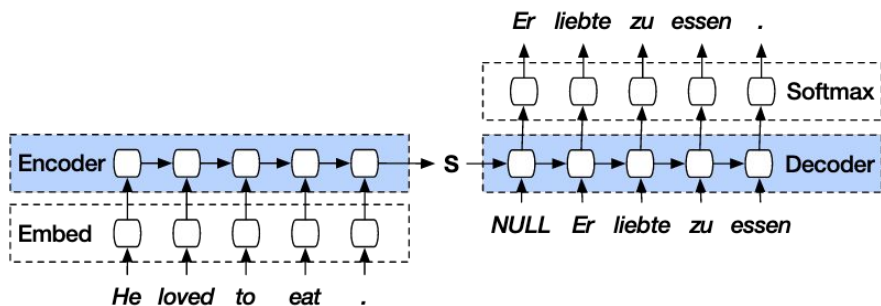
# RNN: Recurrent Neural Networks



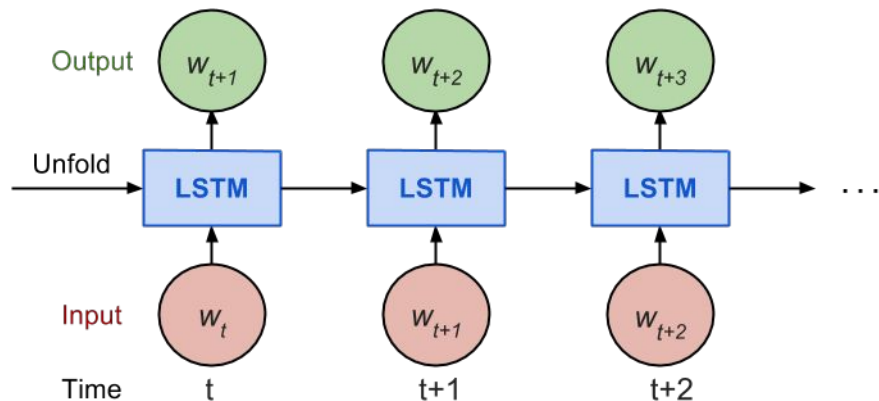
Feedforward (ileri doğru çalışan) ağ tipi, bilgi sadece ileri doğru işlenir.

- Her işlemin çıktısı bir sonraki işlemin girdisi olarak kullanılır
- Sıralı veri için ideal: yazı, konuşma, zaman serileri

# RNN: Recurrent Neural Networks



Çeviri - Neural Machine Translation



Metin üretimi

# Türkçe Doğal Dil İşleme

# Türkçe Doğal Dil İşleme

- Eklemeli yapı
  - gör+dü+m
- Kelime türetme
  - saat+çi, top+la+n+tı
- Ünlü/ünsüz uyumu
  - al+dı, git+ti, gör+dü
- Ünlü/ünsüz düşmesi
  - söyl(e)+üyor

*"I might go"*

*"Gidebilirim"* -> "Gid + eabil + i + rim"

- Yumuşama
  - tıp -> tıbbi
- Deyim açısından zengin
  - Can kulağıyla dinliyordu.

# Türkçe Doğal Dil İşleme

- Eklemeli yapı
  - gör+dü+m
- Kelime türetme
  - saat+çi, top+la+n+tı
- Ünlü/ünsüz uyumu
  - al+dı, git+ti, gör+dü
- Ünlü/ünsüz düşmesi
  - söyl(e)+üyor

*"I might go"*

*"Gidebilirim"* -> "Gid + eabil + i + rim"

- Yumuşama
  - tıp -> tıbbi
- Deyim açısından zengin
  - Can kulağıyla dinliyordu.



# Açık Kaynak Projeler

# Türkçe Doğal Dil İşleme İçin Kaynaklar

## Türkçe veri setleri

- [TWNERTC](#)
- Mozilla Common Voice
- OPUS
- SentiTurkNet
- TSCorpus

## Açık kaynak projeler

- [TRMorph](#)
- [Turkish Stemmer](#)
- [Turkish POS Tagger](#)
- [Deasciifier](#)
- [Turkish Morphology](#)
- [NER experiments in Turkish and English](#)

# Türkçe Doğal Dil İşleme İçin Kaynaklar

## Açık kaynak kütüphaneler

- Zemberek
- PolyGlott (kısmen)
- SpaCy (kısmen)
- NLTK
- fastText

## Eğitilmiş Türkçe Dil Modelleri

- [BERT](#)
- [Word2Vec](#)
- [fastText](#)

# Varlık İsmi Tanıma

# Varlık İsmi Tanıma

[Colab Linki](#)

## Kullanım Alanları

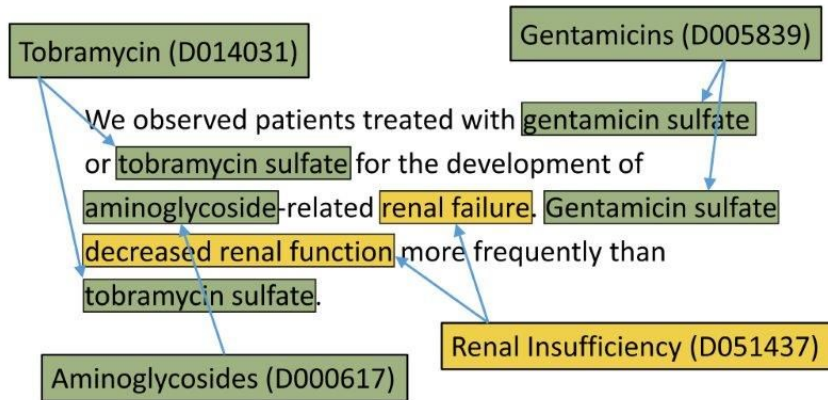
- Gizliliği koruma
- Döküman filtreleme
- İçerik tavsiyesi
- Bilgi çıkarma (müşteri hizmetleri, bankacılık, hukuk, vs.)
- Dil modellerinin performansını arttırma



Hatayspor x Galatasaray x maci oncesinde ev sahibi ekibin  
tarafarlari Galatasaray Teknik Direktoru Fatih Terim'i x tribune  
cagirip Kebapci Selo tezahuratinda bulundu.

# Varlık İsmi Tanıma

[Colab Linki](#)



- Kişi, yer ve organizasyon isminden çok daha fazlası
- Kalp hastalıkları, gezegen-uydu isimleri, bitki isimleri...

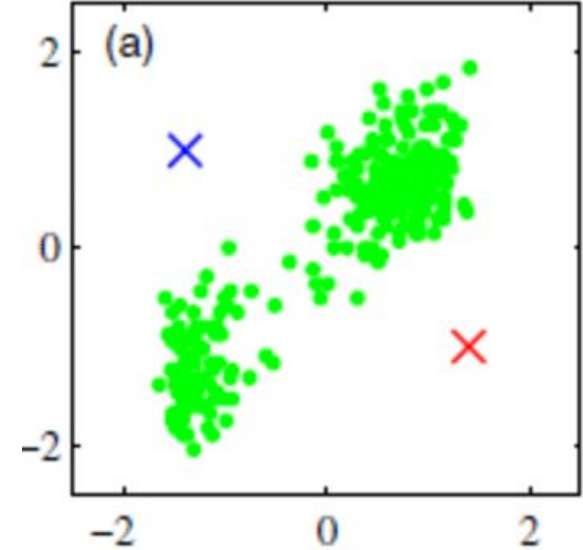
# Metin Kümeleme

# K-Means ve TF-IDF ile Metin Kümeleme

[Colab Linki](#)

Gözetimsiz Metin Kümeleme Kullanım Alanları

- Keşif ve istatiksel analiz
- Anormal girdi tespiti
- Gözetimli öğrenme için veri etiketlemeyi kolaylaştırma

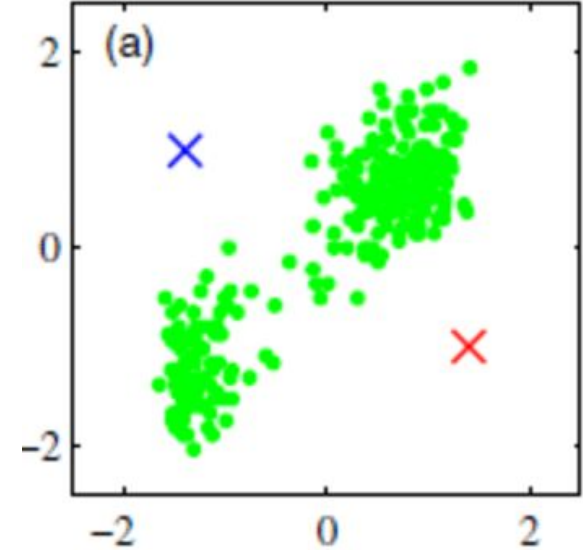




# K-Means ve TF-IDF ile Metin Kümeleme

[Colab Linki](#)

- Küme sayısını belirle
- Hedef
  - Kümeye ait noktaların küme merkezine uzaklığını minimize et
  - Kümeler arası uzaklığı maksimize et
- Nasıl
  - Hedef optimize olana kadar farklı küme merkezi koordinatlarını dene



Dinlediğiniz için teşekkürler!

::

Kod : Github://[alaradirik](#)

::

İletişim : LinkedIn://[alaradirik](#)