# Doğal Dil İşleme ve Metin Madenciliği

## Yapay Zekâ ve Veri Dersleri Rehberi · Kısım II: Yapay Zekânın Dalları

Metin ön işleme | Metin temsili | Dil modelleri | Dönüştürücüler | Metin madenciliği | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü: https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_06_Dogal_Dil_Isleme.html

GİRİŞ

# Ders Hangi Soruyu Sorar?

> **Bilgisayar insan dilini nasıl çözümler, anlar ve üretir?**

Doğal dil işleme, bilgisayarın insan dilini çözümlemesini, anlamasını ve üretmesini inceler; metin madenciliği ise büyük metin yığınlarından bilgi ve örüntü çıkarır. Dersin merkezinde metni sayılara çeviren **metin temsili** ve sıradaki sözcüğü tahmin eden dil modelleri vardır.

## Bir bakışta

| | | |
|---|---|---|
| **YAKLAŞIM**<br>dilbilim, istatistik ve derin öğrenme | **GİRDİ**<br>metin | **ÇIKTI**<br>etiket, çeviri, özet, yanıt |
| **BAŞARI ÖLÇÜTÜ**<br>doğruluk, BLEU, insan değerlendirmesi | **MATEMATİK YÜKÜ**<br>orta | **VERİ İHTİYACI**<br>çok metin (derlem) |

# Kısa Tarihçe

**1954**
Georgetown-IBM deneyi: altmışı aşkın Rusça cümle otomatik olarak İngilizceye çevrilir.

**1966**
Weizenbaum'un ELIZA programı basit kalıp eşlemeyle bir terapisti taklit eder.

**1990'lar**
İstatistiksel yöntemler kural tabanlı sistemlerin yerini alır; IBM çeviri modelleri öne çıkar.

**2013**
Mikolov ve arkadaşlarının word2vec yöntemi sözcükleri anlamı yansıtan vektörlere gömer.

**2017**
Dönüştürücü (transformer) mimarisi tanıtılır; ardından BERT ve GPT gibi önceden eğitilmiş modeller gelir.

**2022**
Büyük dil modelleri sohbet arayüzleriyle geniş kitlelere ulaşır.

DERS

# Temel Kavramlar

**Belirteçleme**
Metni sözcük ya da alt sözcük birimlerine bölmek.

**Biçimbilim**
Kök ve ek çözümleme; Türkçe gibi eklemeli dillerde kritik.

**Sözcük torbası**
Sırayı yok sayıp sözcük sayılarıyla metni temsil etmek.

**TF-IDF**
Belgede sık, derlemde seyrek sözcüklere yüksek ağırlık.

**Sözcük gömmeleri**
Anlamca yakın sözcükler vektör uzayında yakın durur.

**Dil modeli**
Sıradaki sözcüğün olasılığını tahmin eden model.

**Varlık tanıma**
Metindeki kişi, yer ve kurum adlarını bulmak.

**Konu modelleme**
LDA gibi yöntemlerle belgelerdeki gizli konuları bulmak.

# Tipik İş Akışı




**Örnek:** Bir e-ticaret sitesi ürün yorumlarını toplar, metni belirteçlere ayırıp ekleri çözümler, yorumları önceden eğitilmiş bir dönüştürücüyle vektöre çevirir, olumlu ve olumsuz duygu sınıflandırıcısı eğitir ve F1 puanıyla ölçer. Sonuç, şikâyetlerin hangi konularda toplandığını gösterir.

DERS

# Derste Neler İşlenir?

1. Düzenli ifadeler, belirteçleme, normalleştirme
2. Biçimbilim: kök bulma ve ek çözümleme
3. n-gram dil modelleri ve yumuşatma
4. Naive Bayes ve lojistik regresyonla metin sınıflama
5. Vektör anlambilimi: TF-IDF ve word2vec
6. Sözcük türü etiketleme ve gizli Markov modelleri
7. Varlık tanıma ve dizi etiketleme (CRF)
8. Sözdizimsel çözümleme: bağlılık ağaçları
9. Yinelemeli ağlarla dil modelleme
10. Dönüştürücüler, BERT ve önceden eğitme
11. Makine çevirisi ve BLEU ile değerlendirme
12. Konu modelleme (LDA) ve bilgi çıkarımı

## Kazandırdığı beceriler

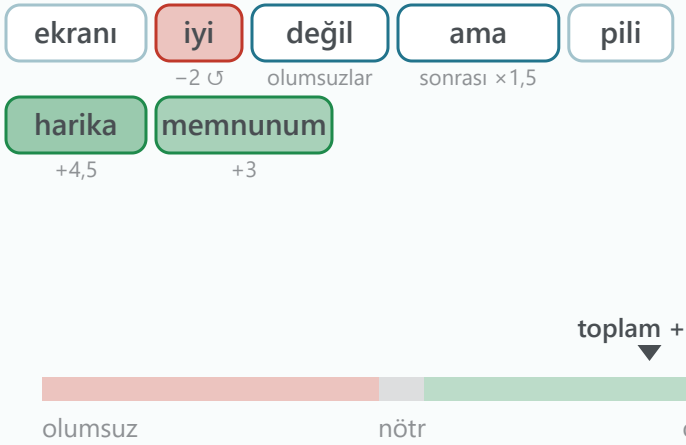
- spaCy ve NLTK ile metin işleme
- düzenli ifadelerle metin temizlemek
- Zemberek ile Türkçe biçimbilim
- Gensim ile sözcük gömmeleri
- Hugging Face modeline ince ayar
- F1 ve BLEU ile değerlendirme

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

# Müşteri Yorumu Olumlu mu?

Bir e-ticaret sitesine günde binlerce yorum geliyor; olumsuzlar hemen müşteri hizmetlerine gitmeli. Hazır bir yorum seçin ya da kendiniz yazın; model sözcükleri tek tek puanlar.




5. Ekranı iyi değil ama pili harika, ...

Ekranı iyi değil ama pili harika, memnunum.

**“değil” ve “-me” etkisini hesapla**

7 sözcük, 3 tanesi duygu taşıyor.
Toplam puan **+5,5** → karar: **olumlu**

Bilgisayar metni önce **sözcüklere** böler, sonra her sözcüğe bir ağırlık verir. Türkçede “beğenmedim” gibi ekler anlamı tersine çevirir; dil işleme bu yapıyı modele öğretmek zorundadır.

KONUM

# Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

| Ders | Neden gerekir |
|---|---|
| Yapay Zekâ | dil anlama ve üretme |
| Programlama | Python ile dizgi işleme, dosya okuma ve düzenli ifadeler |
| Veri Yapıları | sözlük, hash tablosu ve trie: sözcük dağarcığını tutmak için |

KONUM

# Bu Ders Sana Göre mi?

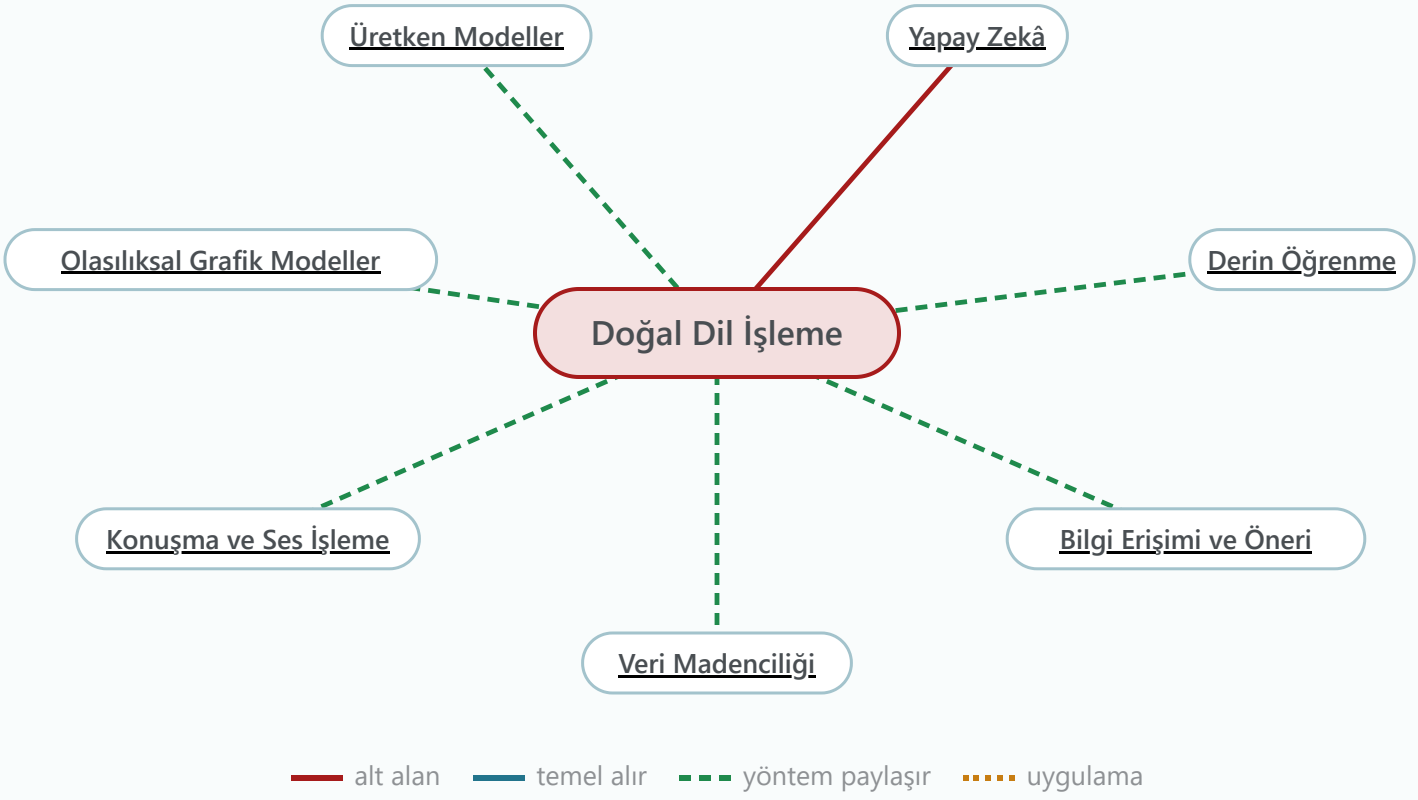
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

| Kendine sor | Hayırsa önce |
|---|---|
| Python'da dizgi işleyip düzenli ifadeyle metin arayabiliyor musun? | Programlamaya giriş, ardından düzenli ifadelere kısa bir giriş |
| Sözlük (hash tablosu) kullanarak sözcük sıklığı sayabiliyor musun? | Veri Yapıları dersindeki hash tablosu konusu |
| Koşullu olasılığı ve Bayes kuralını açıklayabiliyor musun? | Olasılık ve İstatistik |
| Bir sınıflandırıcıyı kesinlik, duyarlılık ve F1 ile ölçebiliyor musun? | Makine Öğrenmesi |

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

# Dersin Komşuları



ÖRTÜŞME

# Örtüştüğü Dersler

| İlişki | Ders | Ortak konu |
|---|---|---|
| alt alan | Bu ders, Yapay Zekâ dersinin alt alanıdır | dil anlama ve üretme |
| yöntem paylaşır | Derin Öğrenme ile yöntem paylaşır | sözcük gömmeleri, dönüştürücü (transformer) mimarisi |
| yöntem paylaşır | Olasılıksal Grafik Modeller ile yöntem paylaşır | saklı Markov modeli, koşullu rastgele alan |
| yöntem paylaşır | Üretken Modeller ile yöntem paylaşır | metin üretimi |
| yöntem paylaşır | Konuşma Tanıma ve Ses İşleme ile yöntem paylaşır | dil modeli, konuşmadan metne |
| yöntem paylaşır | Veri Madenciliği ile yöntem paylaşır | metin sınıflandırma, konu modelleme |
| yöntem paylaşır | Bilgi Erişimi ve Öneri Sistemleri ile yöntem paylaşır | metin temsili, TF-IDF, anlamsal arama |

# Ayrıştığı Noktalar (1/2)

| Ölçüt | Doğal Dil İşleme | Konuşma ve Ses İşleme | Bilgi Erişimi ve Öneri |
|---|---|---|---|
| **Temel soru** | Bilgisayar insan dilini nasıl çözümler, anlar ve üretir? | Ses dalgasından sözcükler, konuşmacı ve olaylar nasıl çıkarılır? | Kullanıcıya en ilgili belge ya da ürün milyonlarca aday içinden nasıl getirilir? |
| **Yaklaşım** | dilbilim, istatistik ve derin öğrenme | akustik sinyal ve dil modeli | eşleme ve sıralama |
| **Girdi** | metin | ses dalgası | sorgu, belgeler, kullanıcı geçmişi |
| **Çıktı** | etiket, çeviri, özet, yanıt | metin, konuşmacı, ses olayı | sıralı sonuç listesi |
| **Başarı ölçütü** | doğruluk, BLEU, insan değerlendirmesi | sözcük hata oranı (WER) | kesinlik, duyarlılık, NDCG |
| **Matematik yükü** | orta | orta-yüksek | orta: vektör uzayı |
| **Veri ihtiyacı** | çok metin (derlem) | çok saatlik kayıt | çok etkileşim verisi |

- **Konuşma ve Ses İşleme:** Konuşma tanıma ses dalgasından metne ulaşır; NLP metin elde edildikten sonra anlamı çözümler.
- **Bilgi Erişimi ve Öneri:** Bilgi erişimi sorguya en ilgili belgeyi getirir; NLP belgenin içindeki dili ayrıntılı olarak çözümler ve üretir.

# Ayrıştığı Noktalar (2/2)

| Ölçüt | Doğal Dil İşleme | Veri Madenciliği | Üretken Modeller |
|---|---|---|---|
| **Temel soru** | Bilgisayar insan dilini nasıl çözümler, anlar ve üretir? | Büyük veri yığınlarındaki gizli ve işe yarar örüntüler nasıl bulunur? | Bir model gerçek veriye benzeyen yeni örnekleri nasıl üretir? |
| **Yaklaşım** | dilbilim, istatistik ve derin öğrenme | keşfe dayalı çözümleme | veri dağılımını öğrenme |
| **Girdi** | metin | büyük tablo ya da işlem verisi | çok büyük, çoğu etiketsiz veri |
| **Çıktı** | etiket, çeviri, özet, yanıt | kural, küme, aykırı değer | yeni örnek: görüntü, metin, ses |
| **Başarı ölçütü** | doğruluk, BLEU, insan değerlendirmesi | ilginçlik, destek, güven | gerçekçilik ve çeşitlilik (FID vb.) |
| **Matematik yükü** | orta | orta | yüksek |
| **Veri ihtiyacı** | çok metin (derlem) | çok | çok fazla |

- **Veri Madenciliği:** Veri madenciliği her tür veride örüntü arar; bu ders dilin kendine özgü yapısını (sözcük, ek, cümle) işler.
- **Üretken Modeller:** Üretken modeller görüntü ve ses dahil her türlü örnek üretir; NLP dili anlama ve çözümleme görevlerini de kapsar.

UYGULAMA

# Uygulama Alanları

**Makine çevirisi**
Diller arasında otomatik çeviri.

**Duygu çözümlemesi**
Yorum ve paylaşımlardaki tutumu bulmak.

**Soru yanıtlama**
Belgeden ya da bilgi tabanından yanıt bulmak.

**Özetleme**
Uzun metni kısa ve öz hâle getirmek.

**Sohbet asistanları**
Kullanıcıyla doğal dilde konuşan sistemler.

**Belge sınıflandırma**
Haber, dilekçe ya da e-postayı konusuna ayırmak.

UYGULAMA

# Riskler ve Etik

**Uydurma (halüsinasyon)**
Dil modelleri var olmayan kaynakları akıcı biçimde uydurur; ABD'de avukatların mahkemeye uydurma içtihat sunduğu davalar oldu.

**Gömmelerdeki kalıp yargılar**
Gömmeler "erkek-programcı, kadın-ev hanımı" benzetmesi kurar; çeviri sistemleri cinsiyetsiz "o"yu kalıp yargıyla çevirir.

**Az kaynaklı diller**
Modeller en iyi İngilizcede çalışır; Türkçe ve verisi az dillerde aynı sistem belirgin biçimde daha çok hata yapar.

**Metinle gözetim**
E-posta, mesaj ve paylaşımlar kişisel veridir (KVKK); metin madenciliği kişileri izlemek ve fişlemek için de kullanılabilir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

# Araçlar ve Örnek Veriler

## Araçlar ve kütüphaneler

NLTK | spaCy | Hugging Face Transformers | Gensim | Stanza | Zemberek (Türkçe)

## Tanınmış veri setleri

Penn Treebank | CoNLL-2003 | SQuAD | GLUE | IMDb film yorumları | Wikipedia dökümleri

UYGULAMA

# Sık Karıştırılanlar

| Yaygın düşünce | Doğrusu |
|---|---|
| Model metni insan gibi anlar. | Dil modelleri istatistiksel örüntü öğrenir; akıcı metin her zaman doğru ya da tutarlı değildir. |
| İngilizce için yapılan her yöntem Türkçede de çalışır. | Eklemeli Türkçede bir kök yüzlerce biçim alır; belirteçleme ve biçimbilim ayrı ele alınır. |
| Büyük dil modelleri çıkınca klasik NLP bitti. | TF-IDF, düzenli ifadeler ve küçük modeller hâlâ hızlı, ucuz ve çoğu iş için yeterlidir. |
| Metin madenciliği sözcük saymaktır. | Sayım başlangıçtır; varlık, ilişki ve konu çıkarmak dilin yapısını gerektirir. |

SONRASI

# Bu Dersten Sonra

## Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

## İlgili meslekler

NLP mühendisi | Hesaplamalı dilbilimci | Veri bilimci | Arama ve öneri mühendisi | Araştırmacı

NLP'den sonra konuşma tanıma ile sözlü dile, bilgi erişimi ile anlamsal aramaya, üretken modeller ile büyük dil modellerinin iç yapısına geçilebilir. Açıklanabilir YZ ve etik dersi ise dil modellerinin yanlılık ve güvenilirlik sorunlarını ele alır.

SONRASI

# Kaynaklar


- Jurafsky ve Martin, *Speech and Language Processing*, 2. baskı, Prentice Hall, 2009.
- Manning ve Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
- Bird, Klein ve Loper, *Natural Language Processing with Python*, O'Reilly, 2009.
- Eisenstein, *Introduction to Natural Language Processing*, MIT Press, 2019.
- Vaswani ve ark., "Attention Is All You Need", *Advances in Neural Information Processing Systems* 30, 2017.

# Son

## Bölüm 6: Doğal Dil İşleme ve Metin Madenciliği

← Önceki bölüm · Alanın haritası · Sonraki bölüm →