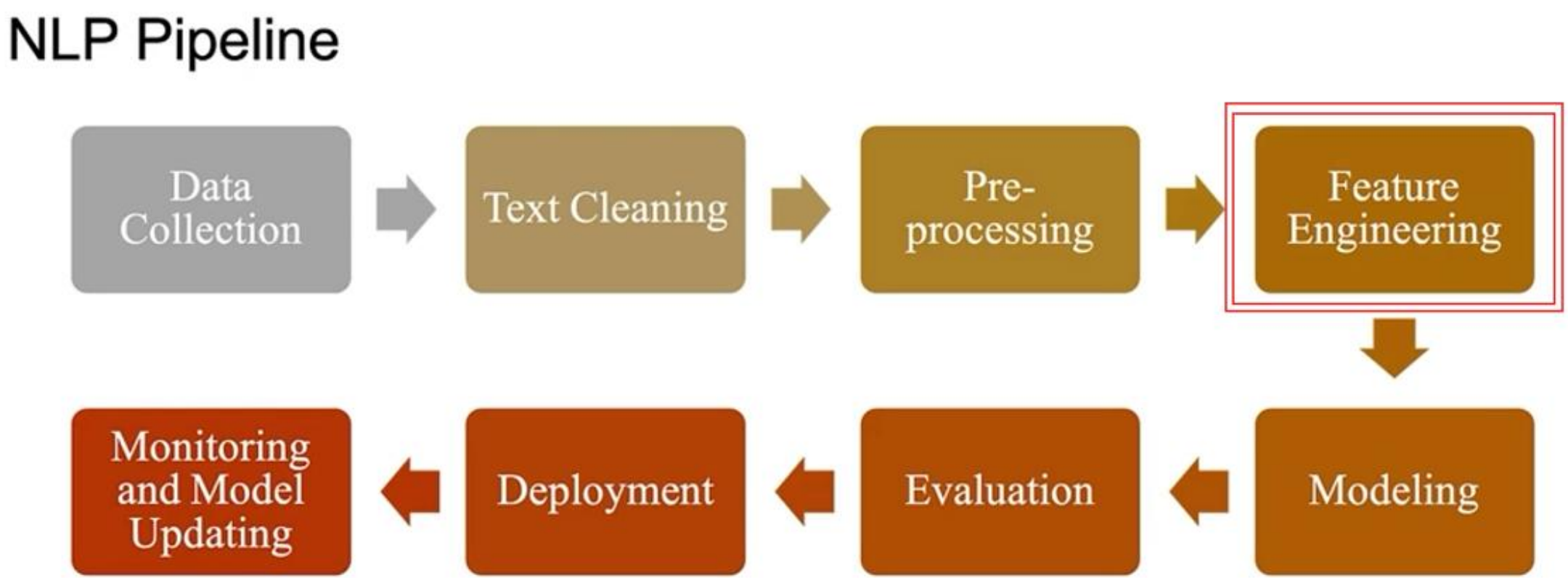
Metin Temsili

- Metin temsili, bir metni sayısal veya başka türde bir formatta temsil etme işlemidir.



Metin Temsili Neden Yapılmalıdır?

Metin temsili (text representation), doğal dilde yazılmış ham metinleri sayısal ve makine tarafından işlenebilir biçime dönüştürme işlemidir. Çünkü modern makine öğrenimi ve NLP modelleri yalnızca sayılarla çalışır; kelimeler, cümleler ve bağlam doğrudan işlemeye uygun değildir. İyi bir metin temsili, metnin anlamsal bilgisini, bağlamını ve önemli özelliklerini koruyarak modellerin daha doğru, genellenebilir ve verimli öğrenmesini sağlar. Ayrıca veri sıkıştırma, gürültü azaltma ve benzerlik ölçümleri gibi pratik faydalar sunar.

Başlık	*Kısa Açıklama*
Bilgisayarların Anlayabilmesi	Metinler önce sayısal vektörlere (örn. bag-of-words, TF-IDF, word embeddings) dönüştürülür; böylece algoritmalar matematiksel işlemlerle metin üzerinde çalışabilir.
Öznitelik Çıkarımı	Temsil, kelime sıklığı, n-gram, sözdizimsel/semantik özellikler, embedding'ler gibi faydalı özniteliklerin çıkarılmasını ve seçilmesini sağlar; bu öznitelikler model için bilgi taşır.
Model Eğitimi	Sınıflandırma, etiketleme, özetleme veya dil modelleme gibi görevlerde modelin giriş verisi temsildir; temsil kalitesi doğrudan model başarımını etkiler (daha iyi temsil → daha iyi performans).

Bag of Words Nedir?

- Bag of Words (BoW), doğal dil işleme (NLP) ve metin madenciliğinde kullanılan temel bir metin temsili yöntemidir.
- BoW, metinlerdeki kelimeleri sayısal verilere dönüştürür ve metinlerin analizini sağlar.

https://dataaspirant.com/bag-of-words-bow/

BoW Yönteminin İşleyişi

1. Kelime Kümesi Oluşturma
2. Kelime Frekansı Hesaplama
3. Vektör Temsili

- **Kelime Kümesi Oluşturma**

  Metin 1: Kedi evde

  Metin 2: Kedi bahçede

  Kelime Kümesi: ["Kedi", "evde", "bahçede"]

- **Kelime Frekansı Hesaplama**

  Metin 1: "Kedi evde"

  "Kedi": 1

  "evde": 1

  "bahçede": 0

  Metin 2: "Kedi bahçede"

  "Kedi": 1

  "evde": 0

  "bahçede": 1

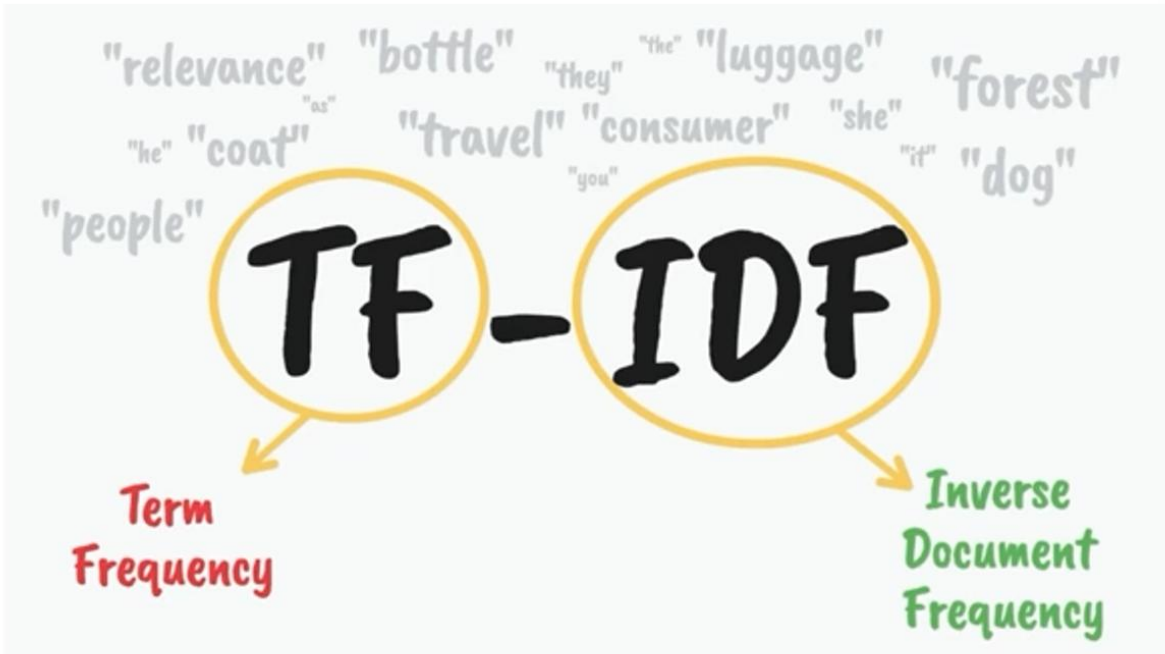
- **Vektör Temsili**

  Metin 1: [1, 1, 0]

  Metin 2: [1, 0, 1]

TF-IDF (Term Frequency-Inverse Document Frequency)

- TF-IDF (Term Frequency-Inverse Document Frequency), metin madenciliğinde ve bilgi erişiminde sıkça kullanılan bir özellik çıkarım yöntemidir.
- TF-IDF, kelimelerin belgeler içinde ne kadar önemli olduğunu belirlemek için kullanılır.



https://www.kdnuggets.com/2022/10/tfidf-defined.html

TF-IDF (Term Frequency-Inverse Document Frequency)

- **Term Frequency (TF):** Bir kelimenin bir belgede ne kadar sık geçtiğini ölçer.
- **Inverse Document Frequency (IDF):** Bir kelimenin tüm belgelerdeki yaygınlığını ölçer. Bir kelimenin çok belgede geçiyorsa, o kelime çok fazla bilgi sağlamaz.

$$\mathrm{TF}(t,d) = \frac{\mathrm{Sayac}(t,d)}{\mathrm{Toplam\ Kelime\ Sayısı}(d)}$$

Burada, $\mathrm{Sayac}(t,d)$ kelimenin t bir belgede d sayısıdır.

$$\mathrm{IDF}(t,D) = \log\left(\frac{\mathrm{Toplam\ Belgeler\ Sayısı}(D)}{1 + \mathrm{Belgelerde\ Geçen\ Sayısı}(t)}\right)$$

Burada, $\mathrm{Belgelerde\ Geçen\ Sayısı}(t)$ kelimenin t belgelerdeki sayısıdır.

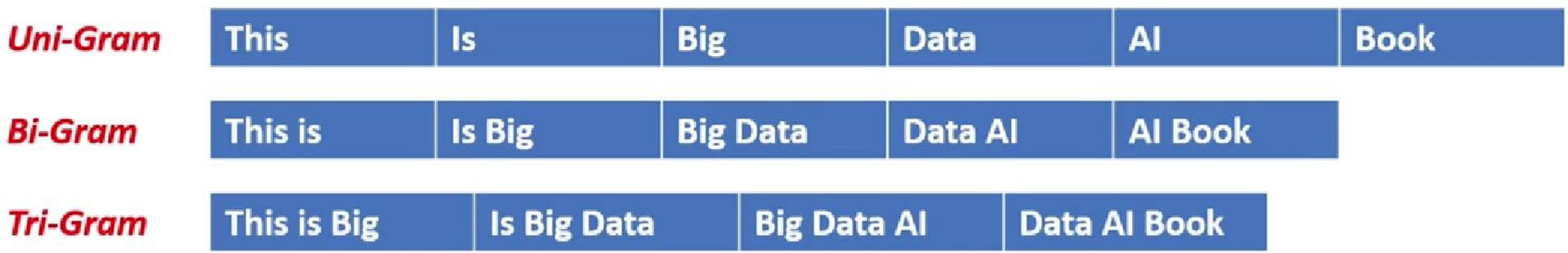
$$\text{TF-IDF}(t,d,D) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D)$$

Bu çarpım, bir kelimenin bir dokümandaki genel önemini gösterir.

N-Gram Modelleri Nedir?

- Bir dil modelinde kullanılan kelime veya karakter dizisinin uzunluğunu belirten bir terimdir.
- N-Gram modelleri, metinleri n kelimelik veya n karakterlik kısımlara bölerek analiz eder.

This is Big Data AI Book



https://devopedia.org/n-gram-model

N-Gram Modelleri

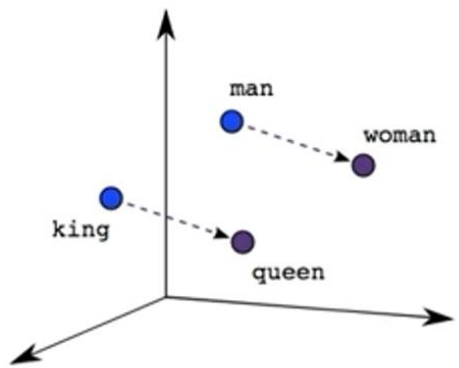
- "Bu bir örnek metindir"
- **Unigram (n=1)**
  - ['Bu', 'bir', 'örnek', 'metindir’]
- **Bigram (n=2)**
  - ['Bu bir', 'bir örnek', 'örnek metindir’]
- **Trigram (n=3)**
  - ['Bu bir örnek', 'bir örnek metindir’]

N-Gram Modelleri Kullanım Alanları

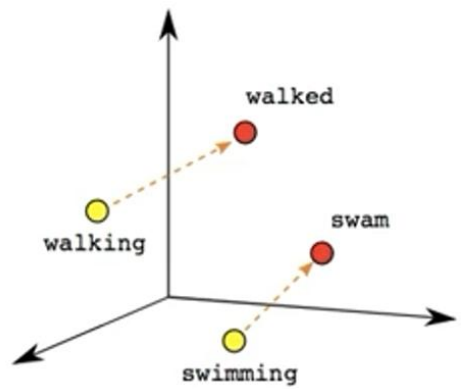
- Metin Modelleme
- Metin Sınıflandırma
- Metin Üretimi
- Metin Benzerliği

Word Embeddings

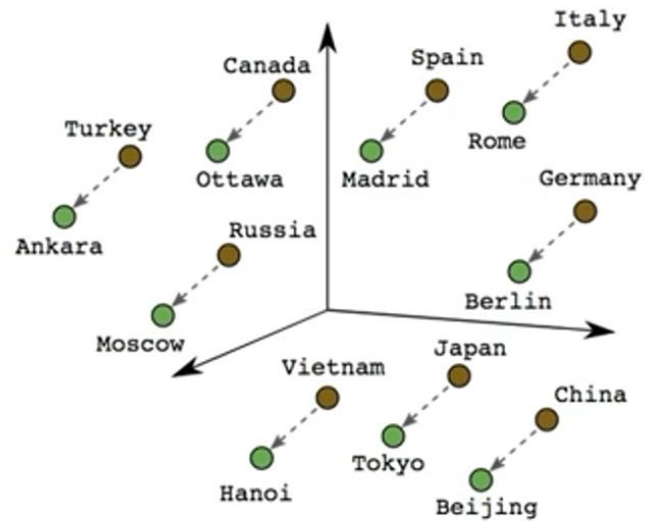
- Word Embeddings (Kelime Gömme), doğal dil işleme (NLP) ve makine öğreniminde kullanılan bir tekniktir.
- Kelimeleri, genellikle sürekli bir vektör uzayında anlamlı temsil edecek şekilde sayısal vektörlere dönüştürür.
- Bu temsiller, kelimeler arasındaki anlamsal ve dilbilgisel ilişkileri yakalamayı hedefler.



Male-Female



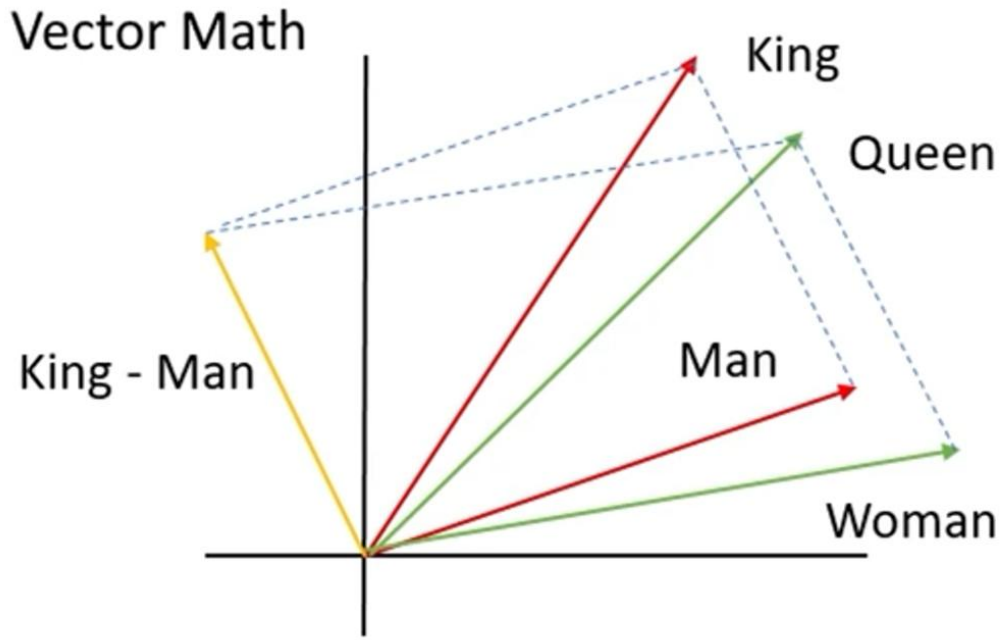
Verb Tense



Country-Capital

Word Embeddings Özellikleri

- Anlamsal Benzerlik
    - Örneğin, "king" ve "queen" kelimeleri benzer vektörler alabilir.
- Matematiksel İşlemler
    - Örneğin, "king" - "man" + "woman" = "queen" hesaplaması yapılabilir.
- Kapsamlılık



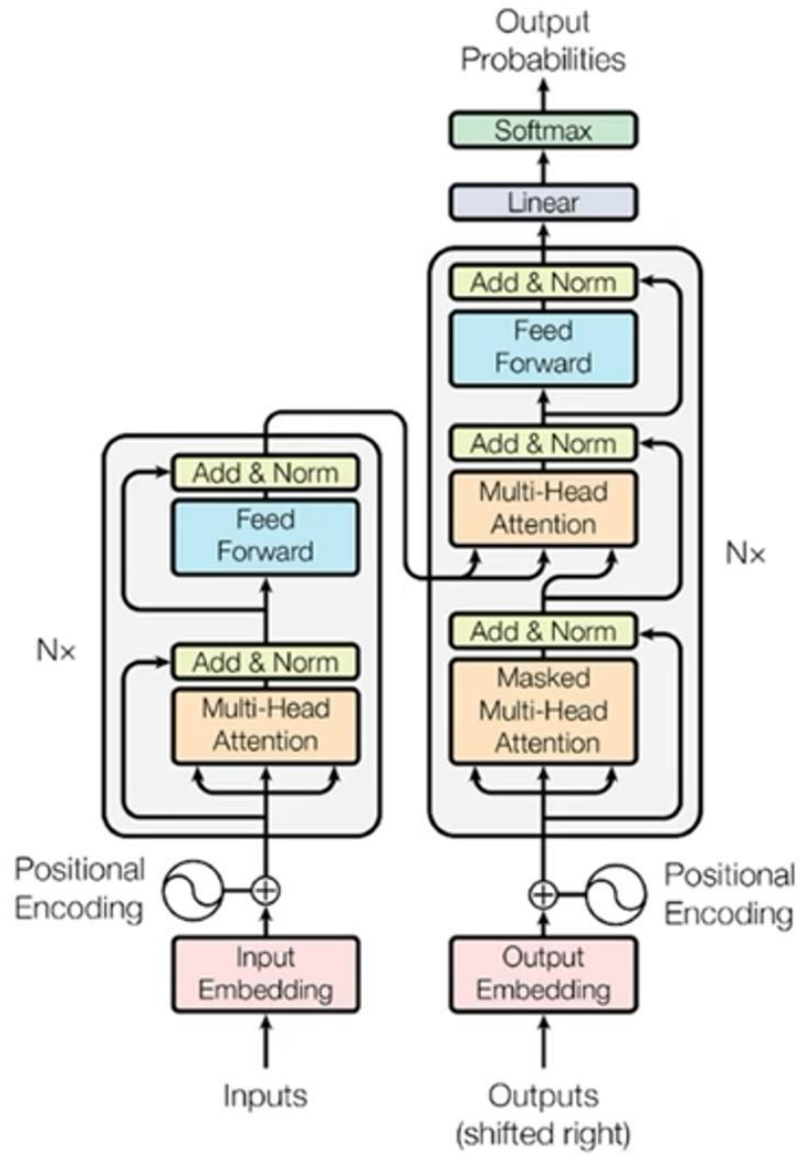
Word Embeddings Modelleri

- **Word2Vec:** Google tarafından geliştirilen, kelimeleri vektörlere dönüştüren ve bu vektörleri dildeki ilişkileri yakalayacak şekilde eğiten bir modeldir.
- **GloVe (Global Vectors for Word Representation):** Stanford Üniversitesi tarafından geliştirilen, kelime gömme temsillerini kelime ortaklıklarını yakalayacak şekilde hesaplayan bir modeldir.
- **FastText:** Facebook tarafından geliştirilen ve kelime gömme temsillerini kelime alt-birimlerini de dikkate alarak hesaplayan bir modeldir.

Transformers Tabanlı Metin Temsili

- Transformers, doğal dil işleme (NLP) ve diğer yapay zeka alanlarında son yıllarda devrim niteliğinde yenilikler getiren bir mimaridir.
- İlk olarak 2017 yılında Google tarafından yayınlanan "Attention is All You Need" adlı makalede tanıtılmıştır.
- Neden Transformers
    - Bağlamı Daha İyi Anlama
    - Paralel İşleme Yeteneği
    - Çeşitli NLP Görevlerinde Kullanım
    - Önceden Eğitilmiş Modellerin Yeniden Kullanımı
- En Bilindik Transformers Modelleri
    - BERT (Bidirectional Encoder Representations from Transformers)
    - GPT (Generative Pre-trained Transformer)

Transformers Tabanlı Metin Temsili



Transformers Tabanlı Metin Temsili

- **Attention**, modelin belirli girdi parçalarına farklı derecelerde dikkat göstermesine olanak tanır.
- Özellikle, bir kelimenin diğer kelimelerle olan ilişkisini anlamak için kullanılır.
- Örneğin, "Kedi hızlıdır" cümlesinde, "hızlıdır" kelimesi "Kedi" kelimesine olan dikkat skorlarını hesaplar:
    - **Sorgu (Query) ve Anahtar (Key) Çarpımı:** "Kedi" kelimesinin "hızlıdır" kelimesi ile olan ilişki skoru hesaplanır.
    - **Dikkat Skoru:** Bu skora göre "Kedi" kelimesinin temsilini günceller.

- **Input Embedding,** girdi verilerini modelin işleyebileceği bir formata dönüştürmek için kullanılan bir tekniktir.
- Örnek: Bir cümle düşünelim: "Kedi hızlıdır.«
- Kelime Vektörleri: Bu cümledeki kelimeler, word2vec, GloVe veya BERT gibi bir embedding tekniği kullanılarak sayısal vektörlere dönüştürülür.
    - "Kedi" → [0.21, -0.32, 0.87, ...]
    - "hızlıdır" → [-0.13, 0.45, -0.20, ...]

Transformers Tabanlı Metin Temsili

- **Multi-Head Attention,** attention mekanizmasının birden fazla başlıkla (head) çalıştığı bir tekniktir.
- Bir cümledeki her kelime, diğer kelimelerle olan ilişkilerini farklı açılardan öğrenmek isteyebilir.
- Örneğin, "Kedi" kelimesinin "hızlıdır" kelimesiyle ilişkisini anlamak için birden fazla dikkat başlığı kullanılır.
    - Başlık 1: "Kedi" ve "hızlıdır" arasındaki anlam ilişkisini öğrenir.
    - Başlık 2: "Kedi" ve "hızlıdır" arasındaki gramatik ilişkileri öğrenir.
    - Başlık 3: "Kedi" kelimesinin cümledeki konumunu öğrenir.

- **Masked Multi-Head Attention**, modelin gelecekteki kelimeleri görmesini engeller, yani model sadece geçmiş bilgileri kullanarak tahminde bulunur.
- Örneğin, "Kedi _____ hızlıdır" cümlesinde, model "Kedi" ve "hızlıdır" arasındaki ilişkilere dayanarak boşluğa "oldukça" kelimesini tahmin eder.

- **Add & Norm**, bir katman çıktı ile giriş arasındaki kısa yolu (residual connection) ekleyip ardından layer normalization uygulayan bir adımdır.
- **Feed-Forward Network**, her encoder ve decoder katmanında bulunan bir ağdır.
- **Output Embedding**, modelin çıktısını temsil eden ve genellikle bir dil modelinde kullanılan bir tekniktir.

Metin Temsili Yöntemlerinin Karşılaştırılması

Yöntem	Temel Özellikler	Kullanım Kolaylığı	Sonuçların Başarı Durumu
Bag of Words (BoW)	Kelime frekanslarına dayalı, sıklık matrisleri oluşturur.	Basit, doğrudan uygulanabilir.	Genellikle düşük, bağlam bilgisinden yoksundur.
TF-IDF	Kelime sıklığına ek olarak, kelimenin belgelerdeki önemini ölçer.	Kolay, standart kütüphaneler mevcut.	Orta, bağlam bilgisi kısıtlıdır ama bilgiye değer katar.
N-grams	Kelime ya da karakter n-gramlarını kullanarak bağlamı yakalar.	Orta, işlem gücü ve bellek kullanımı artar.	Orta, bağlam bilgisi artırılabilir ancak model karmaşıklığı artar.
Word Embeddings (GloVe, Word2Vec, FastText)	Kelimeleri vektörlere dönüştürür, anlam ilişkilerini yakalar.	Orta, önceden eğitilmiş modeller mevcut.	Yüksek, bağlamı daha iyi anlar, semantik ilişkiler sağlar.
Transformers (BERT, GPT-3, vb.)	Derin öğrenme temelli, bağlamı dikkat mekanizması ile yakalar.	Orta-İleri, genellikle yüksek hesaplama gücü gerektirir.	Çok yüksek, bağlamı derinlemesine anlar, çeşitli NLP görevlerinde başarılı.