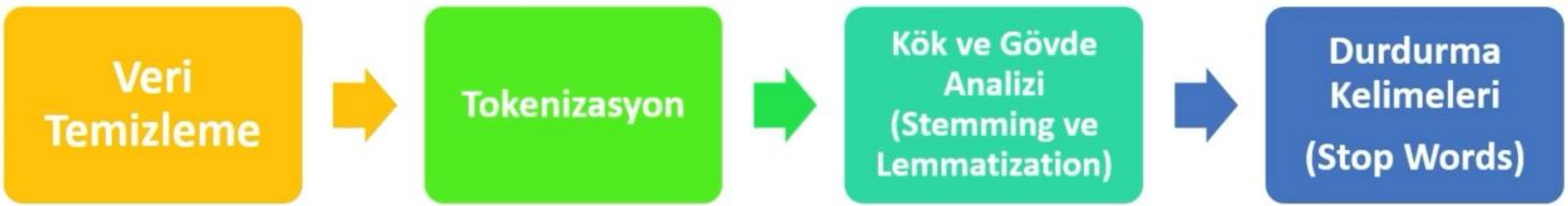
Metin Ön İşleme Adımları



Veri Temizleme

- Veri temizleme, metin verilerini analiz edilebilir hale getirmek için yapılan bir dizi işlemi içerir.
- Bu adımda, metinlerde bulunan hatalı, gereksiz, düzensiz veya model performansını olumsuz etkileyebilecek öğeler kaldırılır veya düzeltilir.
- **Veri Temizleme Adımları:**
  - **Boşlukların Temizlenmesi:** "Hello World" → " Hello World"
  - **Büyük-Küçük Harf Dönüşümleri:** "HeLlO WorlD " → " hello world"
  - **Noktalama İşaretlerinin Kaldırılması:** "Hello, World!" → " Hello World"
  - **Özel Karakterlerin Kaldırılması:** "Bu #yazı $100 dolara mal oldu!" → "Bu yazı 100 dolara mal oldu"
  - **Yazım Hatalarının Düzeltilmesi:** "Hillo Wirld" → "Hello World"
  - **HTML ve URL Temizleme:** " Hello World! " → " Hello World!"

Tokenizasyon

- Tokenization, bir metni daha küçük parçalara ayırma işlemidir.
- Bu küçük parçalar genellikle "token" olarak adlandırılır.
- Tokenlar, kelimeler, cümleler veya hatta karakterler olabilir.

Kök ve Gövde Analizi

- **Stemming (Kök Bulma)**
  - Stemming, kelimelerin kök formunu (yani temel anlamını) bulmak için kelimenin sonundaki eklerin (suffix) çıkarılması işlemidir.
  - Stemming işlemi, kelimenin anlamını tamamen doğru bir şekilde elde etmeyi amaçlamaz; daha ziyade, kelimenin en basit formunu bulmaya odaklanır.
  - Örnek:
    - "koşuyor", "koştu", "koşmak" → "koş"
    - "evde", "evler", "evimiz" → "ev"

Kök ve Gövde Analizi

- **Lemmatization (Gövdeleme)**
  - Lemmatization, kelimeleri sözlükteki temel formlarına (lemma) dönüştürme işlemidir.
  - Lemmatization, kelimenin anlamını ve dilbilgisel yapısını dikkate alarak doğru bir kök bulmaya çalışır.
  - Bu nedenle, lemmatization sonrası elde edilen kelime dilbilgisel olarak anlamlı ve sözlükte yer alan bir kelime olur.
  - **Örnek:**
    - "koşuyor", "koştu", "koşmak" → "koşmak"
    - "evde", "evler", "evimiz" → "ev"

Durdurma Kelimeleri (Stop Words)

- Durdurma kelimeleri (stop words), metinlerde genellikle anlamı çok az olan veya metnin analizi sırasında çok faydalı olmayan kelimelerdir.
- Bu kelimeler genellikle bağlaçlar, edatlar, zamirler ve diğer dil bilgisel işlevi olan kelimelerdir.
- Metin işleme süreçlerinde bu kelimeleri kaldırmak, analizlerin doğruluğunu artırabilir ve metin üzerinde daha anlamlı sonuçlar elde edilmesine yardımcı olabilir.
- **Örnek:**
  - **Türkçe Stop Words:** ve, bir, bu, ile, da, de, mi, o, çok, gibi
  - **İngilizce Stop Words:** and, the, is, in, to, of, it, that