



BTK
AKADEMİ

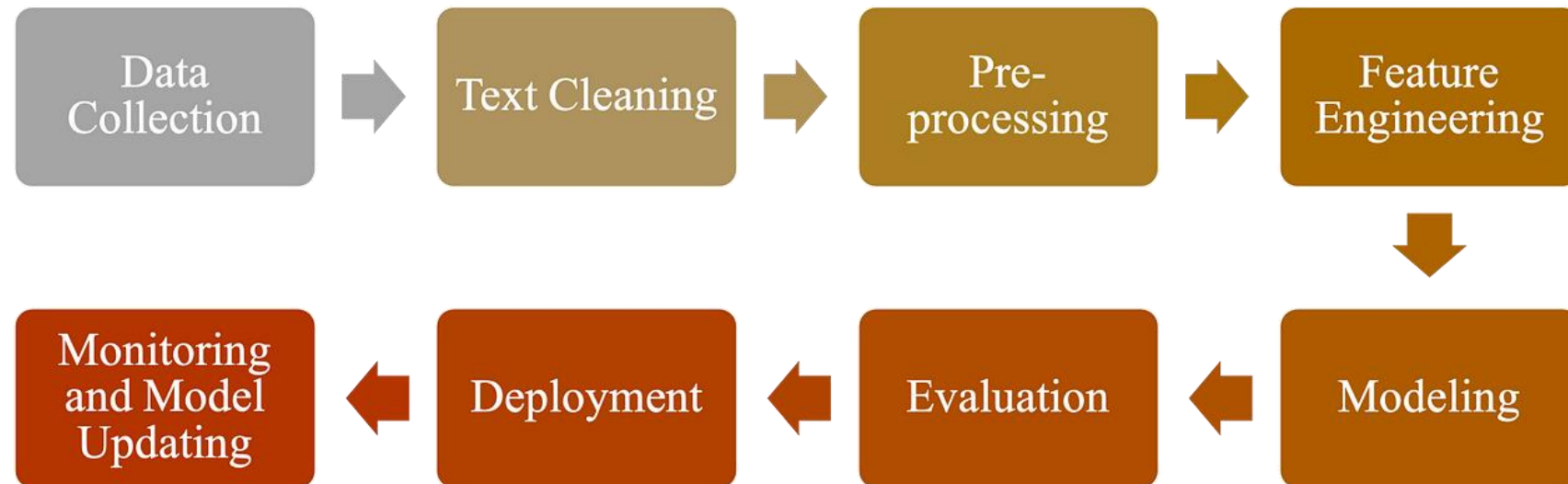
Bölüm
4

Metin Ön İşleme

Metin Ön İşleme

- Metin ön işleme, ham metin verilerini analiz ve işleme için daha uygun bir formata dönüştürme sürecidir.

NLP Pipeline

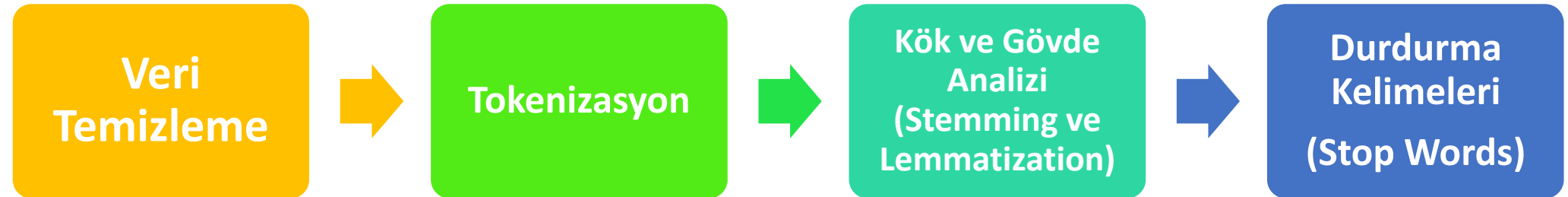


https://alvinntnu.github.io/NTNU_ENC2045_LECTURES/nlp/nlp-pipeline.html

Metin Ön İşleme Neden Yapılmalıdır?

- Veri Kalitesini Artırır
 - "Bu ürün çok *mükemmel!!!*"
- Model Performansını Artırır
- Veri Boyutunu Yönetir
 - "Bu uygulama çok güzel, ama biraz pahalı olabilir. Herkesin beğeneceği bir uygulama!"

Metin Ön İşleme Adımları





BTK
AKADEMİ

Veri Temizleme

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:** "<div> Hello World!</div>" → "Hello World!"



BTK
AKADEMİ

Tokenizasyon

Tokenizasyon

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





BTK
AKADEMİ

Kök ve Gövde Analizi

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", "koştı", "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", "koştı", "koşmak" → "koşmak"
- "evde", "evler", "evimiz" → "ev"



BTK
AKADEMİ

Durdurma Kelimeleri

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