Model Eğitim Öncesi Analiz

Model eğitiminden önceki aşamalar, başarılı bir makine öğrenimi projesinin temel taşlarını oluşturur. Bu süreçte yapılan çalışmalar, modelin performansını doğrudan etkiler ve elde edilen sonuçların güvenilirliğini sağlar. Veri toplama süreci, özellikle kritik bir aşamadır; çünkü modelin öğrenmesi için gerekli olan verinin kalitesi, modelin başarısını belirleyen en önemli faktörlerden biridir. Bu bağlamda, öncelikle henüz oluşturulmamış bir veri setinde hangi özelliklerin yer alacağını belirlemek, yani Özellik Tanımlama (Feature Definition) sürecini tamamlamak gereklidir. Bu süreç, veri seti oluşturulurken toplanacak verilerin hangi niteliklere sahip olacağını belirlemek anlamına gelir ve daha sonra yapılacak feature engineering ve model geliştirme aşamalarını doğrudan etkiler. Özellik tanımlama süreci, veri setinin temelini oluşturur ve veri toplama ile model geliştirme aşamalarının başarısını büyük ölçüde etkiler. Yanlış veya alakasız özellikler seçilirse, modelin performansı düşebilir, hatta modelin yararlı sonuçlar üretmesi engellenebilir. Bu nedenle, veri setini tasarlarken bu sürece yeterince zaman ayırmak ve kapsamlı bir araştırma yapmak, sonrasında yapılacak feature engineering ve model eğitim süreçlerinin kalitesini artırır.

Website kategorizasyonu :

- Haber Siteleri
- Blog Siteleri
- Sosyal Medya Platformları
- E-Ticaret Siteleri
- Eğitim Siteleri
- Devlet/Kamu Kurum Siteleri
- Sağlık Siteleri
- Forum Siteleri
- Eğlence/Medya Siteleri
- Sanat/Kültür Siteleri
- Kişisel Siteler
- Yetişkin İçerik Siteleri
- Kumar/Bahis Siteleri
- İnşaat/Emlak Siteleri
- Reklam/Pazarlama Siteleri

Kategorizasyon Özellikleri(Features) :

Metin Temelli Özellikler :

Metin temelli özellikler, bir web sitesinin ana içeriğini anlamaya yardımcı olur. Metin, web sitelerinin ne hakkında olduğunu ve hangi tür içerikleri sunduğunu belirlemek için en temel bileşendir. Anahtar kelimeler, metin uzunluğu gibi özellikler, web sitesinin kategorisini doğru bir şekilde belirlemek için kritik öneme sahiptir.

1) **Anahtar Kelimeler:**

   Bir haber sitesinde "politika", "ekonomi", "son dakika" gibi kelimeler sıkça geçebilir.

   Bir teknoloji blogunda "yapay zeka", "gadget", "inovasyon" gibi kelimeler öne çıkabilir.

```
import requests
from bs4 import BeautifulSoup
from collections import Counter
import re


def get_website_text(url):
    # Web sayfasını indirme
```

```
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')

    # Sayfadaki tüm metni çekme
    text = soup.get_text(separator=' ')
    return text


def find_keywords(text, keywords):
    # Metni küçük harfe çevirme ve noktalama işaretlerinden arındırma
    text = text.lower()
    text = re.sub(r'[^\w\s]', '', text)

    # Metni kelimelere bölme
    words = text.split()

    # Anahtar kelimelerin sayımını yapma
    keyword_count = Counter(word for word in words if word in keywords)

    return keyword_count


# Örnek kullanım
url = "https://www.hurriyet.com.tr/"  # Örnek Website
keywords = [
    'yapay zeka', 'gadget', 'inovasyon', 'politika', 'ekonomi', 'son
dakika',
    'spor', 'sağlık', 'teknoloji', 'bilim', 'eğitim', 'kültür', 'sanat',
    'müzik', 'sinema', 'finans', 'borsa', 'yatırım', 'kripto', 'çevre',
    'iklim', 'tarım', 'enerji', 'ulaşım', 'otomobil', 'yazılım', 'donanım',
    'mobil', 'siber güvenlik', 'veri bilimi', 'yapay öğrenme', 'derin
öğrenme',
    'robotik', 'blockchain', 'oyun', 'oyun geliştirme', 'e-spor', 'sosyal
medya',
    'girişimcilik', 'startup', 'pazarlama', 'dijital pazarlama', 'SEO',
    'e-ticaret', 'medya', 'haber', 'blog', 'podcast', 'video', 'film',
    'dizi', 'kitap', 'yazar', 'makale', 'rapor', 'analiz'
]


text = get_website_text(url)
keyword_count = find_keywords(text, keywords)
print(keyword_count)
```

Output :
Counter({'haber': 6, 'spor': 4, 'borsa': 3, 'sağlık': 3, 'teknoloji': 2, 'eğitim': 2, 'analiz': 1, 'kripto': 1, 'ekonomi': 1, 'video': 1, 'dizi': 1, 'kitap': 1, 'sanat': 1, 'yatırım': 1, 'startup': 1, 'enerji': 1})

2) **Metin Uzunluğu:**

Blog yazılarının ortalama uzunluğu (örneğin, 1000-2500 kelime) ve haber makalelerinin uzunluğu (örneğin, 300-1000 kelime) arasındaki fark.

```
import requests
from bs4 import BeautifulSoup
import re
```

```
def get_website_text(url):
    # Web sayfasını indirme
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')

    # Sayfadaki tüm metni çekme
    text = soup.get_text(separator=' ')
    return text


def calculate_text_length(text):
    # Metni küçük harfe çevirme ve noktalama işaretlerinden arındırma
    text = text.lower()
    text = re.sub(r'[^\w\s]', '', text)

    # Metni kelimelere bölme
    words = text.split()

    # Kelime sayısını hesaplama
    text_length = len(words)

    return text_length


# Örnek kullanım
url = "https://www.milliyet.com.tr/gundem/yunanistanda-orman-yangini-bakan-
yumakli-2-ucak-ve-1-helikopter-gonderildi-7170675"
text = get_website_text(url)
text_length = calculate_text_length(text)
print(f"Metin uzunluğu: {text_length} kelime")
```

Output :
Metin uzunluğu: 715 kelime

Yapısal Özellikler :

Bir web sitesinin HTML yapısı ve URL düzeni analiz edilerek, sitenin teknik özelliklerini anlayabiliriz. URL yapısı, HTML tag kullanımı gibi özellikler, sitenin içeriğinin nasıl sunulduğunu ve kullanıcı deneyiminin nasıl şekillendiğini gösterir.

1) **URL Yapısı:**

".edu" uzantısına sahip bir site büyük olasılıkla bir eğitim sitesidir.

```
from urllib.parse import urlparse


def analyze_url_structure(url):
    # URL'yi parçalarına ayırma
    parsed_url = urlparse(url)

    # Domain uzantısını belirleme
    domain_extension = parsed_url.netloc.split('.')[-1]

    # Domain uzantılarına göre site türünü belirleme
    if domain_extension == 'edu':
        return "Bu site büyük olasılıkla bir eğitim sitesidir."
    elif domain_extension == 'gov':
```

```
        return "Bu site büyük olasılıkla bir hükümet sitesidir."
    elif domain_extension == 'org':
        return "Bu site büyük olasılıkla bir sivil toplum kuruluşudur."
    elif domain_extension == 'com':
        return "Bu site büyük olasılıkla bir ticari (commercial) sitedir."
    elif domain_extension == 'net':
        return "Bu site büyük olasılıkla bir ağ (network) hizmeti sunan
sitedir."
    elif domain_extension == 'mil':
        return "Bu site büyük olasılıkla bir askeri (military) sitedir."
    elif domain_extension == 'int':
        return "Bu site büyük olasılıkla uluslararası kuruluşlara aittir."
    elif domain_extension == 'info':
        return "Bu site büyük olasılıkla bilgi sağlama amacı güden bir
sitedir."
    elif domain_extension == 'biz':
        return "Bu site büyük olasılıkla bir iş veya ticaret sitesi."
    elif domain_extension == 'io':
        return "Bu site büyük olasılıkla teknoloji veya startup ile ilgili
bir sitedir."
    elif domain_extension == 'tv':
        return "Bu site büyük olasılıkla medya veya yayıncılık ile
ilgilidir."
    elif domain_extension == 'co':
        return "Bu site büyük olasılıkla bir şirket veya ticaret sitesi."
    elif domain_extension == 'me':
        return "Bu site büyük olasılıkla kişisel bir site."
    else:
        return f"Bu site muhtemelen genel bir site, uzantısı:
.{domain_extension}"


# Örnek kullanım
url = "https://www.example.com"
url_analysis = analyze_url_structure(url)
print(url_analysis)
```

2) **HTML Tag Kullanımı:**

   Haber sitelerinde çok sayıda "h1" ve "h2" başlıkları, blog sitelerinde ise "p" ve "blockquote" etiketleri sıkça kullanılır.

Görsel Temelli Özellikler :

Görsel temelli feature'lar, bir web sitesinde bulunan görsel içeriklerin türünü ve yoğunluğunu analiz eder. Görseller, sitenin hangi tür içeriklere odaklandığını anlamak için önemli ipuçları sağlar. Ayrıca, görsel temelli siteler genellikle belirli bir kategoride yer alır, bu da sınıflandırmayı kolaylaştırır.

1) **Görsel İçerik Türü:**

   Sosyal medya platformlarında kullanıcılar tarafından yüklenen çok sayıda profil fotoğrafı ve video bulunabilir. Eğitim sitelerinde ise ders materyalleri ve diagramlar gibi daha çok bilgi amaçlı görseller yer alır.

2) **Görsel Sayısı ve Yoğunluğu:**

Bir haber sitesinde genellikle her makaleye eşlik eden bir görsel bulunurken, sosyal medya sitelerinde sayfa başına onlarca görsel olabilir.

Kullanıcı Etkileşimleri ile İlgili Özellikler :

Kullanıcıların sitede geçirdiği süre, tıklama oranları, sosyal paylaşım gibi veriler, sitenin hangi tür içeriklere odaklandığını ve kullanıcıların bu içeriklerle nasıl etkileşime geçtiğini anlamaya yardımcı olur.

1) **Ziyaretçi Demografisi:**

   Sosyal medya platformlarında genç kullanıcılar ağırlıklı olabilir, eğitim sitelerinde ise daha geniş bir yaş aralığına hitap edebilir.

2) **Bounce Rate :**

   Haber sitelerinde kullanıcılar genellikle tek bir makale okur ve çıkar, bu nedenle yüksek bir bounce rate gözlemlenebilir.

Teknik Özellikler :

Site hızı, mobil uyumluluk, HTTPS kullanımı gibi özellikler, sitenin kullanıcı deneyimini ve güvenliğini nasıl optimize ettiğini gösterir. Bu özellikler, sitenin kategorisini belirlemede dolaylı bir rol oynar.

1) **Site Hızı:**

   Bir teknoloji blogu, multimedya içeriklerine sahip bir eğlence sitesine göre daha hızlı yüklenebilir.

2) **Mobil Uyumluluk:**

   E-ticaret siteleri genellikle mobil uyumlu olacak şekilde tasarlanır, ancak bazı eski blog siteleri bu uyumu göstermeyebilir.

İçerik Temelli Özellikler :

Video içeriği, multimedya kullanımı gibi özellikler, sitenin hangi tür içeriklere odaklandığını anlamaya yardımcı olur. Bu özellikler, sitenin kullanıcıya sunduğu deneyimi ve içeriğin amacını belirler.

1) **Video İçeriği:**

   Bir haber sitesinde genellikle birkaç video bulunurken, bir video akış platformunda çok sayıda video olabilir.

2) **Multimedya Kullanımı:**

   Eğlence siteleri genellikle animasyonlar ve oyunlar içerirken, bir eğitim sitesi daha çok metin ve statik görseller kullanabilir.

Ek olarak Reklam yoğunluğu bir ek feature olabilir.