Veri Seti Hazırlama

Bu projede kullandığım veri seti, film öneri sistemleri alanında en çok kullanılan ve açık kaynak olarak paylaşılan **MovieLens** veri setidir. MovieLens, kullanıcıların izledikleri filmlere verdikleri puanlar üzerinden oluşturulmuş bir veri tabanıdır ve GroupLens Research tarafından araştırma ve eğitim amaçlı olarak herkesin erişimine sunulmaktadır.

Projeyi geliştirirken, sıfırdan bir veri toplamak yerine hazır, güvenilir ve akademik çalışmalarda sıkça tercih edilen bu veri setinden yararlandım. Bu sayede öneri sisteminin temelinde yer alacak puanlama, film türleri ve başlık bilgilerini doğrudan elde edebildim.

Veri setini https://grouplens.org/datasets/movielens/ adresinden indirdim. Kullandığım sürüm, yaklaşık **100.000 kullanıcı değerlendirmesi** içeren **MovieLens 100K** veri setidir. Bu sürümü tercih etmemin sebebi, hem boyut olarak yönetilebilir olması hem de film öneri sistemini eğitmek için yeterli çeşitliliğe sahip olmasıdır.

Veri Setinin İçeriği

MovieLens iki temel dosyadan oluşuyor:

- **movies.csv:**
  Her bir filmle ilgili bilgileri içeriyor. Burada film kimliği, adı ve tür bilgileri yer alıyor.
  Örnek sütunlar:
    - movieId: Her filme ait benzersiz ID
    - title: Filmin adı ve yılı
    - genres: Filmin türleri (örnek: Action, Comedy, Drama)
- **ratings.csv:**
  Kullanıcıların filmlere verdikleri puanları içeriyor.
  Örnek sütunlar:
    - userId: Kullanıcı kimliği
    - movieId: Film kimliği
    - rating: 0.5 ile 5.0 arasında verilen puan
    - timestamp: Oylamanın zamanı

Bu iki dosyayı, ortak movieId sütunu üzerinden birleştirerek tek bir veri tablosu elde ettim. Böylece her film için kullanıcıların ortalama puanlarını hesaplayabildim.

Veri Hazırlama Aşamaları

Veri setini kullanmadan önce, birkaç temel ön işleme adımı uyguladım. Bunlar:

1. Film isimlerinin içindeki yıl bilgisini ayırarak ayrı bir year sütunu oluşturdum.
2. Türler arasında yer alan | karakterini, okunabilirlik açısından virgül (,) karakteriyle değiştirdim.
3. Filmlerin ortalama puanlarını hesaplayarak avg_rating sütununa ekledim.
4. Eksik veya hatalı verileri kontrol ederek uygun biçimde temizledim.
5. Veri setini sadece gerekli sütunlarla sınırlayarak sistemin performansını artırdım.

Bu işlemler sonunda her film için başlık, yıl, tür ve ortalama puan bilgileri hazır hale geldi.

TMDb Entegrasyonu

MovieLens veri seti film isimleri ve tür bilgilerini içeriyor, ancak özet veya görsel gibi detayları bulunmuyor.
Bu yüzden veriyi daha zengin hale getirmek için **The Movie Database (TMDb)** API'sini entegre ettim.

Bu API sayesinde her film için aşağıdaki ek bilgileri çektim:

- Türkçe film adı (varsa),
- Filmin kısa özeti (overview),
- Film afişi (poster) bağlantısı.

Bu bilgiler, uygulamadaki chatbotun film önerilerini daha açıklayıcı ve görsel olarak daha etkileyici hale getirmesini sağladı.

Veri Setini Tercih Etme Sebebim

MovieLens veri setini tercih etmemin birkaç nedeni var:

- Akademik çalışmalarda yaygın olarak kullanıldığı için güvenilir bir kaynaktır.
- Filmlere ait puan, tür ve isim gibi temel özellikleri içerir.
- Gerçek kullanıcıların değerlendirmelerine dayandığı için modelin daha doğru öneriler yapmasını sağlar.
- TMDb API ile kolayca entegre edilerek içerik açısından zenginleştirilebilir.

Sonuç olarak, bu veri seti hem projenin amacı olan **"film önerisi yapan akıllı chatbot"** sistemine tamamen uygun hem de teknik olarak güçlü bir temel sağlamaktadır.