PROJE RAPORU: İÇERİK BAZLI FİLM ÖNERİ SİSTEMİ

Ders: Veri Madenciliği **Öğrenci Adı/No:** Alaattin Buğra DURMUŞ / 231118013 **Tarih:** 07.12.2025 **Kullanılan Ortam:** Google Colab (Python, Pandas, Scikit-learn)

---

1. PROJENİN AMACI

Bu projenin temel amacı, veri madenciliği teknikleri ve doğal dil işleme (NLP) yöntemleri kullanılarak kullanıcıların geçmiş tercihlerine veya bir filmin içeriğine (tür, açıklama vb.) dayalı olarak kişiselleştirilmiş film önerileri sunan bir sistem geliştirmektir. Projede, kullanıcıdan bağımsız olarak ürünün özelliklerine odaklanan **"İçerik Bazlı Filtreleme" (Content-Based Filtering)** yaklaşımı benimsenmiştir.

2. VERİ SETİ VE KAYNAĞI

Projede, akademik çalışmalarda ve öneri sistemleri literatüründe standart bir kıyaslama (benchmark) veri seti olarak kabul edilen **MovieLens** veri seti kullanılmıştır.

- **Veri Kaynağı:** GroupLens Research (University of Minnesota).
- **Versiyon:** MovieLens Latest Small Dataset.
- **Veri Hacmi:** 600'den fazla kullanıcıdan, 9.000'den fazla film için toplanmış yaklaşık 100.000 derecelendirme (rating) ve 3.600 etiket (tag) içermektedir.
- **Seçim Nedeni:** Veri setinin temizlenmiş olması, akademik güvenilirliği ve Google Colab ortamında bellek (RAM) sorunu yaratmadan işlenebilir boyutta olması nedeniyle tercih edilmiştir.

3. KULLANILAN YÖNTEM VE METODOLOJİ

Sistemin geliştirilmesinde aşağıdaki veri madenciliği adımları izlenmiştir:

3.1. Yaklaşım: İçerik Bazlı Filtreleme (Content-Based Filtering)

Bu projede İşbirlikçi Filtreleme (Collaborative Filtering) yerine İçerik Bazlı yaklaşım tercih edilmiştir.

Neden Bu Yöntem Seçildi?

1. **Soğuk Başlangıç (Cold-Start) Sorununa Çözüm:** İşbirlikçi filtreleme, yeni bir film sisteme eklendiğinde o film henüz kimse tarafından oylanmadığı için önerilemez.

Ancak projemizde kullandığımız İçerik Bazlı yöntem, filmin meta verilerini (tür, açıklama) kullandığı için hiç izlenmemiş filmleri dahi başarıyla önerebilir.
2. **Açıklanabilirlik (Explainability):** Kullanıcıya "Bu filmi önerdik çünkü X filmini sevdiniz ve bu iki film tür olarak çok benziyor" şeklinde şeffaf bir açıklama sunulabilir.

3.2. Özellik Çıkarımı: TF-IDF (Term Frequency-Inverse Document Frequency)

Filmlerin tür (genre) bilgileri metin tabanlı verilerdir. Bilgisayarın bu veriyi işleyebilmesi için metnin Sayısal Vektörlere dönüştürülmesi gerekir. Bu aşamada basit kelime sayma (Bag of Words) yerine TF-IDF vektörleştirme tekniği kullanılmıştır.

Neden TF-IDF Seçildi?

- Basit sayma yöntemleri, çok sık geçen ama ayırt edici olmayan kelimelere gereksiz ağırlık verir.
- TF-IDF ise bir terimin (örneğin "Drama" veya "Sci-Fi") veri setindeki özgül ağırlığını hesaplar. Nadir bulunan bir tür eşleşmesi, sık bulunan bir tür eşleşmesinden matematiksel olarak daha değerlidir. Bu da önerilerin isabet oranını artırır.

3.3. Benzerlik Metriği: Kosinüs Benzerliği (Cosine Similarity)

Oluşturulan film vektörleri arasındaki ilişkiyi ölçmek için Kosinüs Benzerliği kullanılmıştır.

Neden Öklid Mesafesi Değil de Kosinüs Benzerliği?

- Yüksek boyutlu veri uzaylarında (metin madenciliği gibi), vektörlerin boyu (magnitude) değil, birbirlerine olan açıları (yönleri) daha anlamlıdır.
- Öklid mesafesi, doküman uzunluğundan etkilenebilirken, Kosinüs benzerliği iki filmin içerik örtüşmesini 0 ile 1 arasında normalize edilmiş bir skorla (açısal olarak) verir.

4. UYGULAMA ADIMLARI VE TEKNOLOJİLER

Proje **Python** programlama dili ile **Google Colab** bulut ortamında geliştirilmiştir. Kullanılan temel kütüphaneler şunlardır:

1. **Pandas:** `csv` dosyalarının okunması (Dataframe), veri temizliği (pipe '|' karakterlerinin temizlenmesi) ve veri manipülasyonu için kullanıldı.
2. **Scikit-learn (sklearn):**
    - `TfidfVectorizer`: Metin verilerini matematiksel matrislere dönüştürmek için.
    - `linear_kernel`: Kosinüs benzerlik matrisini hesaplamak için (Büyük matrislerde `cosine_similarity` fonksiyonuna göre daha hızlı çalıştığı için tercih edildi).

5. SONUÇ VE DEĞERLENDİRME

Geliştirilen sistem, kullanıcıdan bir film ismi aldığında (Örn: "Toy Story"), veri tabanındaki diğer tüm filmlerle olan kosinüs benzerliğini hesaplamakta ve en yüksek skora sahip 10 filmi listelemektedir.

Yapılan testlerde, sistemin aynı türe ve alt kategorilere sahip filmleri başarıyla eşleştirdiği gözlemlenmiştir. Proje, veri madenciliği dersi kapsamında; **veri ön işleme, özellik çıkarımı (vektörleştirme)** ve **benzerlik madenciliği** süreçlerini başarıyla örneklendirmektedir.