Film Öneri Sistemi - Durum Raporu

1. Veri Toplama Durumu Proje, aşağıdaki CSV dosyalarında saklanan MovieLens veri setini kullanmaktadır:

rating.csv: Kullanıcıların filmler için verdiği puanları içerir (userId, movieId, rating, timestamp) movie.csv: Film bilgilerini içerir (movieId, title, genres) tag.csv: Kullanıcı tarafından oluşturulan etiketleri içerir (userId, movieId, tag, timestamp) genome_scores.csv: Filmler için etiket alaka puanlarını içerir (movieId, tagId, relevance) genome_tags.csv: Etiket bilgilerini içerir (tagId, tag) Mevcut Durum:

Veri dosyaları c:/Users/efeba/Desktop/archive konumunda bulunmaktadır Verilerin ilk yüklemesi uygulanmıştır Puanlama veri seti 20.000.263 satır ve 4 sütun içeren büyük bir veri setidir Film veri seti 27.278 satır ve 3 sütun içermektedir Veri doğrulama ve ön işleme fonksiyonları uygulanmıştır Karşılaşılan Sorunlar:

Veri doğrulama sürecinde, özellikle puanlama veri setindeki zaman damgası alanlarının işlenmesinde hatalar bulunmaktadır Film verilerinin işlenmesinde regexp_extract fonksiyonunun tanımlanmamış olmasıyla ilgili bir hata vardır Sonraki Adımlar:

Zaman damgası alanlarını düzgün şekilde işlemek için veri doğrulama fonksiyonunu düzeltmek Eksik regexp_extract fonksiyonu için import eklemek Tüm veri setleri için veri ön işlemeyi tamamlamak Veri kalitesini ve bütünlüğünü doğrulamak 2. Platform ve Sistemlerin Durumu Mevcut Kurulum:

Programlama Dili: Python Büyük Veri İşleme: Apache Spark ve PySpark Makine Öğrenimi: Spark MLlib (özellikle işbirlikçi filtreleme için ALS) Grafik İşleme: GraphFrames (planlanmış ancak tam olarak uygulanmamış) Görselleştirme: Matplotlib, Seaborn, NetworkX Geliştirme Ortamı: Yerel Windows makinesi Uygulama Durumu:

Veri Yükleme Modülü (data_loader.py): Temel yapı uygulanmıştır Tüm veri setlerini yüklemek için fonksiyonlar oluşturulmuştur Veri doğrulama ve ön işleme kısmen uygulanmıştır İstatistiksel analiz fonksiyonları oluşturulmuştur Model Eğitim Modülü (model_trainer.py): ALS öneri modeli uygulanmıştır Çapraz doğrulama çerçevesi kurulmuştur Değerlendirme metrikleri (RMSE, MAE) uygulanmıştır Model performansı için görselleştirme fonksiyonları oluşturulmuştur Grafik Görselleştirme Modülü (graph_visualizer.py): Temel yapı oluşturulmuştur Uygulama tamamlanmamıştır Ana Uygulama (main.py): İş akışı düzenlemesi uygulanmıştır Spark oturum yapılandırması kurulmuştur Hata işleme uygulanmıştır Mevcut İlerleme:

Sistemin temel mimarisi yerindedir İlk veri yükleme ve ön işleme uygulanmıştır ancak hatalar içermektedir Model eğitim çerçevesi uygulanmıştır Görselleştirme bileşenleri kısmen uygulanmıştır 3. Kurulum Sorunları, Teknik Bilgi, Demo Çalıştırmaları Kurulum Sorunları:

Spark Yapılandırması: Proje, findspark kullanılarak kurulan uygun Spark yapılandırması gerektirmektedir Depolama sorunlarını önlemek için çalışma zamanında Spark geçici dizini oluşturulur GraphFrames Entegrasyonu: GraphFrames paketi Spark yapılandırmasına dahil edilmiştir ancak uygulama tamamlanmamıştır Hata İşleme: Çeşitli çalışma zamanı hataları ile karşılaşılmıştır: Veri doğrulamada zaman damgası alanlarıyla ilgili tip uyuşmazlığı hatası regexp_extract fonksiyonu için eksik import Sistemin başarıyla çalışabilmesi için bu hataların düzeltilmesi gerekmektedir Gerekli Teknik Bilgi:

Apache Spark ve PySpark: Spark DataFrame işlemleri anlayışı Öneri sistemleri için Spark MLlib bilgisi Spark yapılandırma ve optimizasyon deneyimi Öneri Sistemleri: İşbirlikçi filtreleme algoritmaları anlayışı ALS (Alternating Least Squares) algoritması bilgisi Hiperparametre ayarlama ve çapraz doğrulama deneyimi Grafik Analizi: Grafik teorisi ve algoritmaları bilgisi Grafik işleme için GraphFrames veya NetworkX deneyimi Grafikler için görselleştirme teknikleri anlayışı Demo Çalıştırma Durumu:

İlk demo çalıştırmaları veri yükleme aşamasında hatalarla karşılaşmıştır Sistem henüz model eğitim aşamasına ilerlememiştir Henüz tam bir uçtan uca çalıştırma başarılı olmamıştır 4. Makale Taslağı Özet Bu makale, Apache Spark üzerinde inşa edilmiş, büyük ölçekli film puanlama veri setlerini işlemek için tasarlanmış ölçeklenebilir bir film öneri sistemi sunmaktadır. Sistem, Alternating Least Squares (ALS) algoritmasını kullanarak işbirlikçi filtreleme tekniklerini kullanır ve film ilişkilerinin grafik tabanlı analizi yoluyla önerileri geliştirir. Dağıtılmış hesaplama yeteneklerinden yararlanarak, uygulamamız 20 milyondan fazla puanlama içeren MovieLens veri setini verimli bir şekilde işler. Sistem, kapsamlı veri ön işleme, çapraz doğrulama ile model eğitimi ve hem tahmin analizi hem de grafik tabanlı film ilişkileri için görselleştirme bileşenlerini içerir. Yaklaşımımız, büyük veri teknolojilerinin öneri sistemlerine nasıl etkili bir şekilde uygulanabileceğini göstermekte, kullanıcı tercihlerine ve film benzerliklerine ölçekli bir şekilde içgörüler sağlamaktadır.

İlgili Çalışmalar Öneri Sistemleri Öneri sistemleri, kullanıcıların geniş seçenekler arasında ilgili içeriği keşfetmelerine yardımcı olarak çevrimiçi platformların önemli bileşenleri haline gelmiştir. Özellikle matris faktörizasyon teknikleri olmak üzere işbirlikçi filtreleme yaklaşımları, bu alanda önemli başarılar göstermiştir (Koren vd., 2009). Netflix Ödülü yarışması (Bennett & Lanning, 2007) bu yöntemleri popüler hale getirmiş ve film önerileri için etkinliklerini göstermiştir.

Öneri Sistemleri için Dağıtılmış Hesaplama Veri setleri büyüdükçe, geleneksel öneri algoritmaları ölçeklenebilirlik zorluklarıyla karşılaşmaktadır. Apache Spark gibi dağıtılmış hesaplama çerçeveleri, öneri algoritmalarının ölçekli bir şekilde uygulanmasını sağlamıştır (Meng vd., 2016). Spark MLlib kütüphanesi, büyük ölçekli öneri görevlerine başarıyla uygulanan ALS dahil olmak üzere işbirlikçi filtreleme algoritmalarının ölçeklenebilir uygulamalarını sağlar (Zadeh vd., 2020).

Grafik Tabanlı Öneri Grafik tabanlı yaklaşımlar, kullanıcılar ve öğeler arasındaki karmaşık ilişkileri yakalayarak öneri sistemleri için güçlü teknikler olarak ortaya çıkmıştır (Huang vd., 2018). Bu yöntemler, öneri problemini iki parçalı bir grafikteki bağlantı tahmini görevi olarak modelleyebilir. Apache Spark üzerine inşa edilmiş GraphFrames, işbirlikçi filtreleme yaklaşımlarıyla entegre edilebilen grafik işleme için dağıtılmış bir çerçeve sağlar (Dave vd., 2016).

Değerlendirme Metrikleri Öneri sistemlerini değerlendirmek, Kök Ortalama Kare Hatası (RMSE), Ortalama Mutlak Hata (MAE), kesinlik, geri çağırma ve sıralama tabanlı metrikler dahil olmak üzere çeşitli metrikleri içerir (Herlocker vd., 2004). Çapraz doğrulama teknikleri, sağlam değerlendirme sağlamak ve aşırı uyumu önlemek için yaygın olarak kullanılır (Cremonesi vd., 2010).

Önerilen Uygulama Sistem Mimarisi Film öneri sistemimiz dört ana bileşenden oluşmaktadır:

Veri Yükleme ve Ön İşleme Modülü: Film veri setlerinin yüklenmesi, doğrulanması ve ön işlenmesini yönetir Model Eğitim Modülü: İşbirlikçi filtreleme için çapraz doğrulama ile ALS algoritmasını uygular Grafik Analiz Modülü: Film ilişki grafiklerini oluşturur ve analiz eder Görselleştirme Modülü: Model performansı ve film ilişkileri için görselleştirmeler oluşturur Veri İşleme Hattı Veri işleme hattı şunları içerir:

CSV dosyalarından ham verilerin yüklenmesi Veri kalitesinin doğrulanması ve eksik değerlerin işlenmesi Özelliklerin çıkarılması (örn. başlıktan film yılı) Kategorik özelliklerin (örn. türler) uygun formatlara dönüştürülmesi Verilerin eğitim ve test setlerine bölünmesi Öneri Algoritması Spark MLlib'den ALS algoritmasını kullanarak işbirlikçi filtrelemeyi uyguluyoruz. Uygulama şunları içerir:

Çapraz doğrulama yoluyla hiperparametre ayarı RMSE ve MAE metrikleri kullanarak değerlendirme Tahmin doğruluğunun görselleştirilmesi Kişiselleştirilmiş film önerilerinin oluşturulması Grafik Tabanlı Analiz Sistemimizin grafik tabanlı bileşeni:

Kullanıcı puanlarına dayalı bir film benzerlik grafiği oluşturur Topluluk tespit algoritmaları kullanarak benzer filmler kümelerini tanımlar Grafik özelliklerini kullanarak önerileri geliştirir İçgörüler sağlamak için film ilişkilerini görselleştirir Değerlendirme Çerçevesi Değerlendirme çerçevemiz şunları içerir:

Sağlam model değerlendirmesi sağlamak için K-katlı çapraz doğrulama Çoklu değerlendirme metrikleri (RMSE, MAE) Farklı model konfigürasyonlarının karşılaştırılması Farklı kullanıcı segmentleri için öneri kalitesinin analizi Görselleştirme Bileşenleri Sistem şunlar için görselleştirmeler içerir:

Model performansı (gerçek vs. tahmin edilen puanlar) Hata dağılımı Film benzerlik grafikleri Kullanıcı puanlama desenleri Sonuç ve Sonraki Adımlar Film öneri sistemi projesi, temel mimarinin yerinde olmasıyla sağlam bir temel oluşturmuştur. Veri yükleme, model eğitimi ve görselleştirme bileşenleri uygulanmıştır, ancak sistem tam olarak işlevsel hale gelmeden önce çözülmesi gereken sorunlar hala vardır.

Acil Sonraki Adımlar:

Veri yükleme ve ön işleme modüllerinde tanımlanan hataları düzeltmek Grafik görselleştirme bileşeninin uygulamasını tamamlamak Tam veri seti ile uçtan uca testler çalıştırmak Daha iyi performans için Spark yapılandırmasını optimize etmek Ek metriklerle değerlendirme çerçevesini geliştirmek Uzun Vadeli Hedefler:

Karşılaştırma için ek öneri algoritmaları uygulamak Önerilerin interaktif keşfi için bir web arayüzü geliştirmek Hibrit bir öneri sistemi oluşturmak için içerik tabanlı özellikleri dahil etmek Gerçek zamanlı öneri yeteneklerini keşfetmek Bu durum raporu, projenin mevcut durumunun kapsamlı bir genel bakışını sağlayarak, hem kaydedilen ilerlemeyi hem de karşılaşılan zorlukları vurgulamaktadır. Proje, gelişmiş özelliklere sahip ölçeklenebilir bir film öneri sistemi sunma yolunda ilerlemektedir, ancak kısa vadede ele alınması gereken çeşitli teknik sorunlar bulunmaktadır.