DATE : 13.03.2024

DT/NT : NT

LESSON : MACHINE LEARNING

SUBJECT : RECOMMENDATION ENGINE PROJECT

BATCH : 223

TECHPRO EDUCATION

techproeducation.com

+1 (585) 304 29 59

MOVIE RECOMMENDATION AND ANALYSIS

Tavsiye-Öneri (Recommendation) Sistemleri

- Recommendation system'ler, büyük veri kümeleri arasından kullanıcılara en uygun içeriği, ürünü veya hizmeti önermek için tasarlanmıştır.
- Kullanıcıların tercihlerine ve davranışlarına dayanarak kişiselleştirilmiş öneriler sunmayı amaçlar.
- Bu sistemler, kullanıcı memnuniyetini artırarak, kullanıcı etkileşimini güçlendirir ve sonuç olarak dönüşüm oranlarını ve satışları artırabilir.
- Bu sistemler, e-ticaretten video akış platformlarına, sosyal medyadan içerik sağlayıcılarına kadar birçok alanda kullanılır.

Ana Yaklaşımlar

Recommendation system'ler genellikle üç ana yaklaşım kullanır: içerik tabanlı filtreleme, işbirlikçi filtreleme ve hibrit modeller.

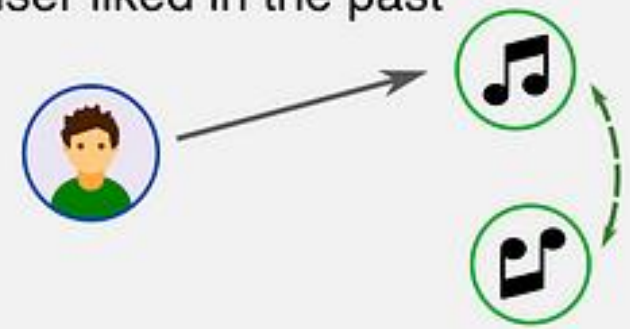
- **Content-Based (İçerik Tabanlı) Filtreleme:** Kullanıcının geçmişte ilgilendiği öğelerle benzer özelliklere sahip öğeleri önerir. Örneğin, bir kullanıcı belirli bir yazarın kitaplarını okuduysa, sistem o yazarın diğer kitaplarını veya benzer tarzda kitapları önerir.

- **Collaborative (İşbirlikçi) Filtreleme:** Kullanıcıların geçmiş etkileşimlerine dayanarak öneriler yapar. Bu model, benzer tercihlere sahip kullanıcıları bulup, bir kullanıcının henüz deneyimlemediği ama benzer kullanıcılar tarafından beğenilen öğeleri önerir.

- **Hibrit Modeller:** İçerik tabanlı ve işbirlikçi filtreleme yöntemlerinin kombinasyonunu kullanarak önerilerde bulunur. Bu yaklaşım, her iki modelin avantajlarını birleştirerek daha doğru ve kişiselleştirilmiş öneriler sunmayı hedefler.

Content-Based

- Use items metadata / tags
- Suggest items similar to what user liked in the past



Recommended

MORE LIKE
Billie Eilish

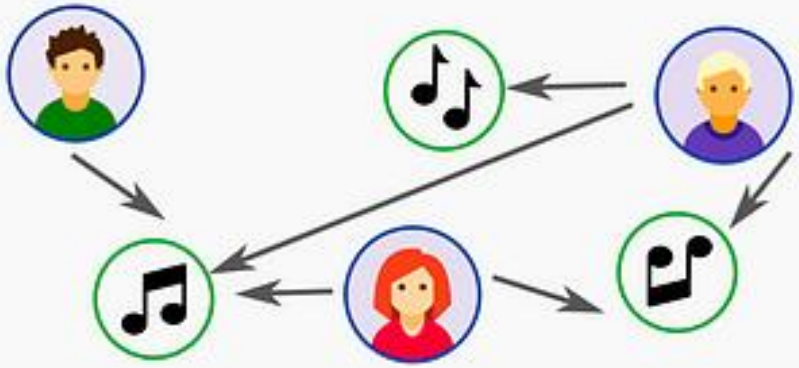
American Teen
Khalid

Reflection
Fifth Harmony

Lo Vas A Olvidar
Rosalía

Collaborative Filtering

- Use all feedbacks from all users
- Similiar users like similar items



Recommended

Made For You

Discover Weekly
Enjoy new discoveries chosen just for you!

Your Daily Drive
Fans like you also like these songs!

Diğer ML/DL Modelleri

- **Matris Faktorizasyonu Yöntemleri:** Genellikle yüksek boyutlu matrisleri, daha küçük ve yönetilebilir boyutlardaki matrislerin çarpımı şeklinde ifade etmeyi amaçlar. Veri içerisindeki gizli yapıları ve ilişkileri ortaya çıkarmak için kullanılır. Popüler yöntemler arasında SVD (Singular Value Decomposition) ve ALS (Alternating Least Squares) bulunur.

- **Kümeleme Algoritmaları:** Kümeleme algoritmaları, benzer özelliklere sahip kullanıcıları veya öğeleri gruplara ayırır. Bu gruplar üzerinden öneriler yapılır. Popüler kümeleme yöntemleri arasında K-means ve DBSCAN bulunur.

- **Derin Öğrenme Yöntemleri:** CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), Autoencoders gibi yöntemler, karmaşık öğe ve kullanıcı profillerini modelleyebilir ve kişiselleştirilmiş öneriler sunabilir. Bu modeller, özellikle zengin kullanıcı etkileşim verileri olduğunda ve karmaşık öğrenme görevlerinde etkilidir.

Önemli Kavramlar ve Teknolojiler

Similarity (Benzerlik):

- İki öğe veya kullanıcı arasındaki benzerliği hesaplayan bir metriktir.
- Benzerlik hesaplamaları, öneri sistemlerinin doğruluğunu artırmak için kullanılır.
- Örnek metrikler: kosinüs benzerliği, Pearson korelasyon katsayısı, Jaccard benzerliği.

User Profile (Kullanıcı Profili):

- Kullanıcının tercihlerini, geçmiş davranışlarını ve demografik bilgilerini içeren bir veri setidir.
- Bu profiller, öneri yapılırken karşılaştırılır.

Item Profile (Öğe Profili)

- Öğenin özelliklerini, içeriğini ve diğer meta verilerini içeren bir veri setidir.

Cold Start Problem (Soğuk Başlangıç Sorunu)

- Yeni bir kullanıcı veya öğe sisteme eklendiğinde, yeterli etkileşim verisi olmadığından doğru önerilerde bulunmak zorlaşır.
- Çözüm yolları arasında popüler öğeleri önermek veya demografik bilgilere dayanarak önerilerde bulunmak yer alır.

Uygulama Alanları

- **E-ticaret:** Kullanıcılara alışveriş tercihlerine göre ürün önermek.
- **Video ve Müzik Akış Hizmetleri:** Kullanıcıların izleme veya dinleme alışkanlıklarına göre içerik önermek.
- **Sosyal Medya:** Kullanıcılara ilgi alanlarına göre içerik veya diğer kullanıcıları önermek.
- **Haber Siteleri ve Bloglar:** Okur ilgisine göre haber veya makale önermek.

Problem Açıklaması

Kullanıcıların tercih ettiği benzer türlere ve filmlere dayalı analiz ve temel öneriler

Üzerinde odaklanacağımız bazı kritik noktalar:

- Filmlerin karlılığı
- Film diline dayalı gross analizi
- Farklı film türleri için gross ve profit karşılaştırması
- Oyunculara, filmlere, türlere dayalı öneri sistemleri

Bu proje, bu faktörler arasındaki korelasyonu anlamamıza yardımcı olacak.

Bir Filmin Karını Hesaplama

Budget: Yapımcıların bir filmi üretmek için harcadıkları, yapım, oyuncu ve reklam maliyetlerini içeren tutardır.

Gross: Yapımcıların filmlerini sinemalarda gösterime sokarak, uydu haklarını TV'ye satarak, Prime, Hulu, Disney+Hotstar, Netflix vb. gibi OTT (Over-the-top medya servisleri) platformlarından kazandıkları tutardır.

Profit: Gross - Budget

Tüm zamanların en karlı filmlerini hesaplamak için bu formülü kullanacağız.

Sosyal Medya Popülerliğini Hesaplama

Sosyal medya popülerliğini belirlemek için önemli faktörler şunları içerir:

- Filme oy veren kişi sayısı.
- Filmi değerlendiren kişi sayısı.
- Film sayfasındaki Facebook beğeni sayısı.

Bu metrikleri kullanarak, bu filmlerin sosyal medya popülerliğini hesaplamak için bir formül oluşturduk.

(No. of People Reviewed for Movie / No. Of People Voted for Movie) * No. Of Facebook Likes)

Aktörler için Rapor Hazırlama

Bir aktörün aşağıdaki özet bilgilerini içeren bir fonksiyon.

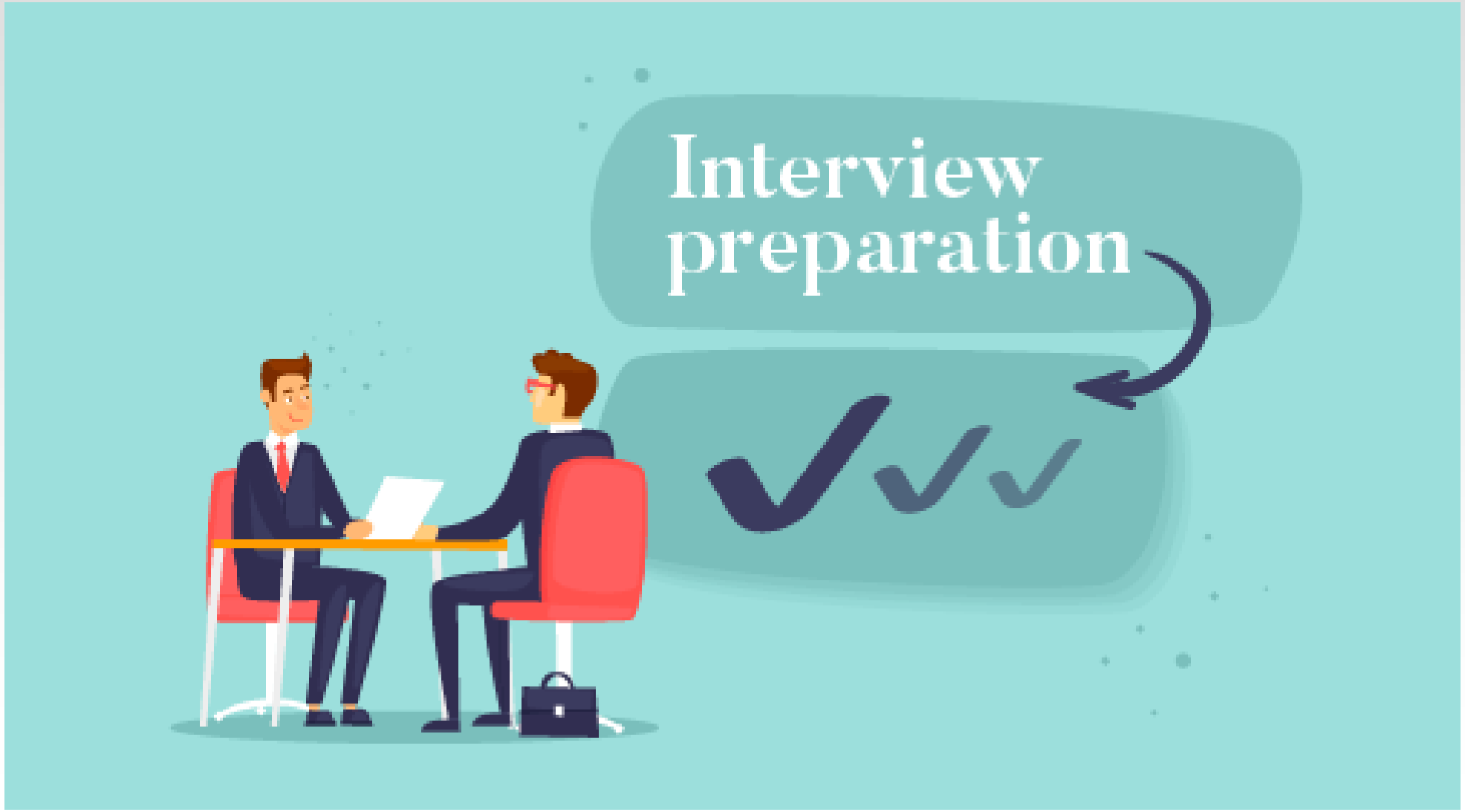
- The Time Period of Actor
- Maximum Gross Amount
- Minimum Gross Amount
- Average IMDB Ratings for the Movie
- Most Common Genres

Önemli Çıkarımlar

Bir projeyi tamamladıktan sonra temel çıkarımları analiz etmek çok önemlidir. Şimdi başlıca temel çıkarımları tartışalım.

- Eksik değerlerin işlenmesini anlama
- Feature Engineering nasıl yapılır
- Veriler gereksinimlere göre nasıl manipüle edilir
- İçeriğe (content) dayalı tavsiyeler nasıl yapılır
- Benzerliğe (similarity) dayalı tavsiyeler nasıl yapılır

INTERVİEW PREPARATION



EDA (Keşifsel Veri Analizi) sürecinde tipik olarak hangi adımları takip edersiniz?

- Genellikle **veri setini anlama**, **veri temizleme**, **veri görselleştirme** ve **veri ön işleme** adımlarını içerir.
- İlk olarak veri setinin yapısını, eksik değerleri, aykırı değerleri ve veri türlerini incelerim. Ardından, eksik veya hatalı verileri düzeltmek için temizleme işlemleri yaparım.
- Görselleştirme için matplotlib, seaborn gibi kütüphaneleri kullanarak verilerin dağılımını, korelasyonunu ve trendlerini incelerim.
- Son olarak, modelleme için veriyi uygun hale getirmek amacıyla feature engineering ve scaling işlemleri gerçekleştiririm.

EDA sırasında kullandığınız Python kütüphaneleri nelerdir ve neden tercih edersiniz?

- **Numpy** ve **Pandas** veri manipülasyonu ve analizi için, **matplotlib** ve **seaborn** ise veri görselleştirme için tercih ettiğim kütüphanelerdir.
- Pandas'ın DataFrame yapısı veri üzerinde kolaylıkla işlem yapmamı sağlarken, matplotlib ve seaborn kütüphaneleri ise veri setinin görsel olarak keşfedilmesine olanak tanır.
- Ayrıca, seaborn'un istatistiksel görselleştirmeler için sunduğu yüksek seviye arayüz, karmaşık veri ilişkilerini anlamayı kolaylaştırır.

Bir veri setindeki değişkenleri incelemek için hangi görselleştirme yöntemlerini kullanırsınız?

- Kategorik değişkenler için **barplot** veya **countplot**;
- Sürekli değişkenler için **histogram**, **boxplot** veya **violin plot**;
- İki değişken arası ilişkiler için **scatter plot** veya **pairplot**;
- Korelasyon için **heatmap** kullanırım.

EDA sürecinde veri kalitesi sorunlarını nasıl belirlersiniz ve bu sorunları nasıl ele alırsınız?

- Veri kalitesi sorunları, **eksik değerler**, **tutarsız girdiler**, **aykırı değerler** ve **gereksiz veriler** gibi sorunları içerebilir.
- Bu sorunları pandas kütüphanesindeki **isnull()**, **describe()**, **value_counts()** ve **unique()** gibi fonksiyonları kullanarak belirleyebiliriz.
- Eksik verileri doldurmak için **fillna()**, aykırı değerleri ele almak için **IQR yöntemi** veya **Z-skoru**, ve gereksiz verileri kaldırmak için **drop()** metodunu kullanabiliriz.

Eksik veri ile çalışırken hangi teknikleri kullanırsınız ve bu tekniklerin avantajları ve dezavantajları nelerdir?

- Eksik verileri ele almanın birkaç yolu vardır: eksik verileri silmek, ortalama veya medyan ile doldurmak veya modelleme yöntemleri kullanarak tahmin etmek.
- Bu yöntemlerin her birinin kendi avantajları ve dezavantajları vardır.
- Örneğin, veri silme yöntemi, veri kaybına neden olabilirken, imputasyon yöntemleri veri setinin yapısını bozmadan eksik verileri doldurmamıza izin verir.

Veri setinizdeki kategorik değişkenleri nasıl işlersiniz ve bu değişkenleri modelleme için nasıl hazırlarsınız?

- Kategorik değişkenleri işlemek için **one-hot encoding**, **label encoding**, **ordinal encoding** veya **binary encoding** gibi yöntemler kullanılır.
- One-hot encoding, her kategori için yeni bir sütun oluştururken, label encoding her kategoriyi benzersiz bir sayısal değere dönüştürür.
- Modelin gereksinimlerine ve kategorik değişkenin özelliklerine bağlı olarak en uygun yöntem seçilir.
- Son olarak, modelleme için veriyi uygun hale getirmek amacıyla feature engineering ve scaling işlemleri gerçekleştiririm.

Bir veri setinin aykırı değerlerini (outlier) nasıl tespit eder ve ele alırsınız?

- Aykırı değerler, box plot veya z-skorları gibi yöntemlerle tespit edilebilir.
- Aykırı değerlerin ele alınmasında kullanılan yöntemler ise, bu değerleri kaldırmak, sınır değerlerle değiştirmek veya daha gelişmiş yöntemler olan robust istatistikler veya model tabanlı yaklaşımlar kullanmak olabilir.

Zaman serisi verileri ile çalışırken karşılaşabileceğiniz zorluklar nelerdir ve bu zorlukların üstesinden gelmek için hangi yöntemleri kullanırsınız?

- Zaman serisi verileriyle çalışırken karşılaşılan zorluklar arasında sezonluluk, trendler, dönemsellik ve otokorelasyon bulunur.
- Ayrıca, zaman serisi verileri genellikle eksik verilere veya ani sapmalara (outlier) sahip olabilir.
- Bu zorlukların üstesinden gelmek için öncelikle veriyi görselleştirme ve zamana göre gruplama yaparak anlamaya çalışırım.
- Trend ve sezonluluk gibi bileşenleri modellemek için ARIMA, SARIMA gibi modelleri veya LSTM gibi derin öğrenme yöntemlerini kullanırım.
- Eksik veriler için ise interpolasyon, zaman serisi özgü imputasyon yöntemleri veya tahmin edici modeller kullanılabilir.

Pandas kütüphanesinde groupby fonksiyonunu kullanmanın avantajları nelerdir ve tipik bir kullanım senaryosu verebilir misiniz?

- groupby fonksiyonu, veriyi kategorilere ayırıp bu kategoriler üzerinde özet istatistikler hesaplamak için kullanılır.
- Veri setindeki kategorik değişkenlere göre gruplama yaparak, her gruba özel analizler ve işlemler yapmamızı sağlar.
- Tipik bir kullanım senaryosu olarak, bir e-ticaret sitesindeki kullanıcıların satın alma miktarlarını kullanıcı tipine göre gruplayıp, her grup için ortalama satın alma miktarını hesaplamak verilebilir.

Tea break...

10:00

Start | Stop | Reset | mins: 10 | secs: 0 | type: Tea

Breaktime for PowerPoint by Flow Simulation Ltd.

Pin controls when stopped