

TOBB ETÜ Bilgisayar Mühendisliği Bölümü
BİL 470/570 Dönem Projesi Ara Rapor

2021-22 Yaz Dönemi

Proje Üyesi: Mert Can GÖNEN

Date: 01/07/2022

1. Proje ismi

- The Content Selector for Disney Plus User

2. Proje tanımı ¹ (kısa özet, problemin açıklanması, beklenen sonuçlar)

- Günümüzde birçok insanın kullandığı, diziler ve filmler içeren bir video oynatma platformu olan Disney Plus'ta bulunan içerikler için kullanıcıya en uygun ve en doğru içeriği seçmeye yarayan bir uygulama projesidir. Proje kapsamında supervised learning metodolojisi ile bir prediction yapılacaktır.

3. Proje metodolojisi, daha önce benzer projelerde kullanılan metotlar hakkında bilgiler, literatür taraması

- Daha önce benzer projelerde TfIdfVectorizer kullanılarak recommendation sistemleri oluşturulmuş. Cosine Similarity ve Count Vectorizer kullanılarak da oluşturulan projeler bulunmakta. Cosine Similarity metodu TfIdfVectorizer'le de beraber kullanılmakta. Bu projenin amacı ise Cosine Similarity ve Count Vectorizer kullanarak bir recommendation sistemi oluşturmak.

4. Projede kullanılan veriler, dış bağlantılar hakkında bilgiler, verilerin tamamını veya birkaç örneği gösterebilirsiniz.

- Projede kullanılan veriler bu linkte bulunmakta:**
<https://www.kaggle.com/datasets/shivamb/disney-movies-and-tv-shows>
- Veri olarak filmlere ait sıralı id'ler, vizyon türü, başlık, yönetmen, cast ve vizyona giriş tarihi gibi türler bulunmaktadır. Veriler tavsiye işlemi için kullanılacakken sıralı halde değildir. Bazı öznitelikler bazı instance'lar için missing value şeklindedir. Bu yüzden öncelikle bu verilerin elden geçirilmesi ve bir işleme sokulması gerekli. Bu instance'ları drop etmek ya da tavsiye için öğretim işlemini yaparken o instance'a ait diğer öznitelikleri kullanarak bir sonuç çıkartmak düşünülen iki çözüm. Ancak bunun için henüz herhangi bir karar verilmedi. Bunun yanı sıra verilerde herhangi bir duplicate instance örneği yok. Ancak olur da karşılaşırsa bunlardan sadece bir tanesini dahil edilip diğeri drop edilecek. Veriler aşağıdaki formatta bulunmaktadır:

	show_id	type	title	director	cast	country	date_added	release_year	rating	duration	Genre	description	year_add	month_add
0	s1	Movie	Duck the Halls: A Mickey Mouse Christmas Special	Alonso Ramirez Ramos, Dave Wasson	Chris Diamantopoulos, Tony Anselmo, Tress MacN...	United States	November 26, 2021	2016	TV-G	23 min	Animation	Join Mickey and the gang as they duck the halls!	2021	November
1	s2	Movie	Ernest Saves Christmas	John Cherry	Jim Varney, Noelle Parker, Douglas Seale	United States	November 26, 2021	1988	PG	91 min	Comedy	Santa Claus passes his magic bag to a new St. ...	2021	November
2	s3	Movie	Ice Age: A Mammoth Christmas	Karen Disher	Raymond Albert Romano, John Leguizamo, Denis L...	United States	November 26, 2021	2011	TV-G	23 min	Animation	Sid the Sloth is on Santa's naughty list.	2021	November
3	s4	Movie	The Queen Family Singalong	Hamish Hamilton	Darren Criss, Adam Lambert, Derek Hough, Alexa...	United States	November 26, 2021	2021	TV-PG	41 min	Musical	This is real life, not just fantasy!	2021	November
5	s6	Movie	Becoming Cousteau	Liz Garbus	Jacques Yves Cousteau, Vincent Cassel	United States	November 24, 2021	2021	PG-13	94 min	Biographical	An inside look at the legendary life of advent...	2021	November

¹ Eğer gerekiyorsa ayrı bir sayfa kullanın

- Burada görülen özniteliklerde bulunan "cast, description ve Genre" tavsiye için temel alınacak özniteliklerdir. Görüldüğü gibi veri kümesinde hem sayısal hem de text tabanlı veriler bulunmaktadır. Karşılaştırılabilir verilerin yanı sıra karşılaştırma yapmak yerine gruplandırılacak veriler de var. Son raporda veriler belirli özniteliklere göre gruplanıp grafikler aracılığıyla temsil edilecektir. Missing values, rating, release year, duration, country main (filmin ait olduğu ülke) ve dal (drama, komedi, vb.) gibi özniteliklere ait grafikler oluşturulacak.

5. Projenin ayrıntılı planı, akış şeması, veri işleyişi vs., projede karşılaşılan sorunlar, projenin şu andaki durumu (varsa elde edilen sonuçlar veya ekran çıktıları), bundan sonra yapılacaklar

- Projenin kodlanmasına henüz başlanmadı. Ancak 1-2 hafta içerisinde tam hali hazır olmuş olacak. Ancak tavsiye işlemi için cast, description ve Genre öznitelikleri kullanılarak bir 'metadata soup' oluşturulacak, ardından count matrix CountVectorizer kullanılarak oluşturulacak. Son olarak Cosine Similarity hesaplanacak ve beğenilen film ismiyle beraber bu similarity değeri fonksiyona parametre olarak verilecek. Sonuçlar listelenecek. Bu işlemler başlamadan önce 4.maddede anlattığım işlemler veriyi daha elverişli hale getirmek için uygulanacak. Çıktı olarak ise önerilen filmlere ait isimler kullanıcıya verilecek.