**Import Library dan File Unloading**

Langkah pertama yang harus kita lakukan adalah melakukan import library yang dibutuhkan untuk pengerjaan project ini dan melakukan pembacaan dataset.

Notes :

* Library yang akan kita gunakan adalah **pandas** (as **pd**) dan **numpy**(as**np**)
* Dataset yang akan digunakan adalah **title.basics.tsv** dan **title.ratings.tsv**

Akses dataset :

1. title.basic.tsv = https://dqlab-dataset.s3-ap-southeast-1.amazonaws.com/title.basics.tsv
2. title.ratings.tsv = https://dqlab-dataset.s3-ap-southeast-1.amazonaws.com/title.ratings.tsv

## 5 Data teratas dari table movie

Hal pertama yang akan kita lakukan adalah menampilkan 5 data teratas yang ada pada table movie (movie\_df).

## Tipe Data dari Setiap Kolom

Setelah berhasil menampilkan 5 data teratas yang ada pada table movie (movie\_df), hal selanjutnya yang akan kita lakukan adalah melakukan pengecekan **tipe data dari masing-masing kolom** yang ada pada **table movie** (movie\_df) tersebut.

## Pengecekan Data dengan Nilai NULL

Merupakan suatu hal yang wajib untuk melakukan pengecekan terhadap nilai NULL yang ada di dalam dataset saat melakukan cleaning.

Oleh karena itu, hal selanjutnya yang akan kita lakukan adalah melakukan pengecekan apakah ada**data bernilai NULL pada masing-masing kolom yang ada pada table movie (movie\_df)**

## Analisis Kolom dengan data bernilai NULL - part 1

Dari hasil pengecekan nilai NULL yang sudah dilakukan sebelumnya, diketahui bahwa kolom 'primaryTitle' dan 'originalTitle' memiliki banyak data yang bernilai NULL.

Hal selanjutnya yang akan kita lakukan adalah melakukan pengecekan terhadap **bentuk data dari kolom primaryTitle dan originalTitle yang bernilai NULL.**

## Membuang Data dengan Nilai NULL - part 1

Setelah melihat hasil sebelumnya, dapat dilihat bahwa semua data tidak memiliki judul dan kita dapat membuang data-data tersebut.

Pekerjaan selanjutnya yang akan kita lakukan adalah membuang data dengan nilai NULL tersebut dan melihat jumlah data yang ada setelah data-data bernilai NULL tersebut dibuang.

## Analisis Kolom dengan data bernilai NULL - part 2

Selain kolom 'primaryTitle' dan 'originalTitle',masih terdapat kolom lain yang memiliki data bernilai NULL. Kolom tersebut adalah kolom 'genres'

Selanjutnya, kita akan melakukan hal yang sama seperti yang sudah kita lakukan pada kolom 'primaryTitle' dan 'originalTitle'.

Lakukan pengecekan terhadap **bentuk data dari kolom genres yang bernilai NULL.**

## Membuang Data dengan Nilai NULL - part 2

Setelah melihat hasil sebelumnya, dapat dilihat bahwa semua data tidak memiliki judul dan kita dapat membuang data-data tersebut.Pekerjaan selanjutnya yang akan kita lakukan adalah membuang data dengan nilai NULL tersebut dan melihat jumlah data yang ada setelah data-data bernilai NULL tersebut dibuang.

## Mengubah Nilai '\\N'

Jika kita perhatikan pada kolom 'startYear' , 'endYear', dan 'runtimeMinutes', terdapat data dengan nilai '\\N'

'\\N' berarti NULL.

Hal selanjutnya yang akan kita lakukan adalah mengubah nilai dari \\\N tersebur menjadi np.nan dan melakukan casting kolom startYear, endYear, dan runtimeMinut

## Mengubah nilai genres menjadi list

Selanjutnya, kita akan membuat sebuah function yang bernama **transform\_to\_list** untuk mengubah nilai genre menjadi list.

## Menampilkan 5 data teratas

Seperti yang sudah kita lakukan pada table movie (movie\_df) sebelumnya, sekarang kita akan menampilkan 5 data teratas dari table ratings (rating\_df)

## Menampilkan tipe data

Selanjutnya, kita akan menampilkan tipe data dari masing-masing kolom yang ada pada table rating (rating\_df)

## Inner Join table movie dan table rating

Mari kita melakukan **inner** join antara rating\_df dan movie\_df untuk mendapatkan rating pada setiap film yang tersedia, lalu tampilkan **5 data teratas** dan**tipe data** dari tiap kolom yang ada.

## Memperkecil ukuran Table

Hal selanjutnya yang akan kita lakukan adalah memperkecil ukuran table dengan menghilangkan semua nilai NULL dari kolom **startYear** dan **runtimeMinutes** karena tidak masuk akal jikalau film tersebut tidak diketahui kapan tahun rilis dan durasi nya.

## Pertanyaan 1: Berapa nilai C?

Hal pertama yang akan kita cari adalah nilai dari C yang merupakan rata-rata dari averageRating

## Pertanyaan 2: Berapa nilai m?

Mari kita ambil contoh film dengan numVotes di atas 80% populasi, jadi populasi yang akan kita ambil hanya sebesar 20%.

## Pertanyaan 3: Bagaimana cara membuat fungsi weighted formula?

Selanjutnya kita harus membuat sebuah fungsi dengan menggunakan dataframe sebagai variable.

## Pertanyaan 4: Bagaimana cara membuat simple recommender system?

Dari task yang sudah kita lakukan sebelumnya, telah terdapat field tambahan 'score'.

Pertama kita akan filter numVotes yang lebih dari m kemudian diurutkan score dari tertinggi ke terendah untuk diambil nilai beberapa nilai teratas

## Pertanyaan 5: Bagaimana cara membuat simple recommender system dengan user preferences?

Dari task yang sudah dilakukan sebelumnya, dapat dilihat sekarang daftar film telah diurutkan dari score tertinggi ke terendah.

Film dengan averageRating yang tinggi tidak selalu mendapat posisi yang lebih tinggi dibanding film dengan averageRating lebih rendah, hal ini disebabkan karena kita juga memperhitungkan faktor banyaknya votes

Sistem rekomendasi ini masih bisa ditingkatkan dengan menambah filter spesifik tentang titleType, startYear, ataupun filter yang lain

Pekerjaan selanjutnya yang akan kita lakukan adalah membuat function untuk melakukan filter berdasarkan isAdult, startYear, dan genres.