

**PRACTICAL MACHINE LEARNING**  
**COURSE ASSIGNMENTS**  
**Content Based - Recommender System**  
**CLASS I1**



**list of group member:**

<b>Student name</b>	<b>Student ID</b>
Batara Agung Pramudito Trilaksono	082011633055
Fikri Yoma Rosyidan	082011633076
Aristo Riza Muhammad	082011633077
Rais Ilham Nustara	082011633090
Aida Filza Binti Aidrus	502122060005

**Lecturer :**

Endah Purwanti, S.Si., M.Kom.  
197812172005012001

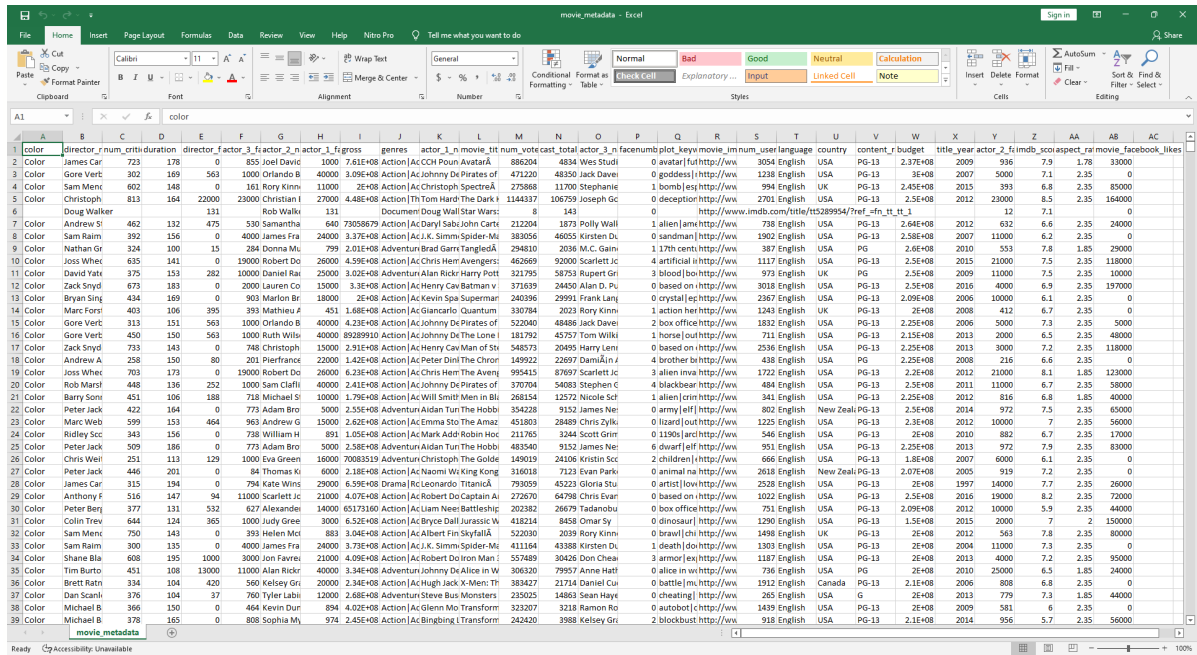
**INFORMATION SYSTEMS STUDY PROGRAM**  
**FACULTY OF SCIENCE AND TECHNOLOGY**  
**AIRLANGGA UNIVERSITY**  
**SURABAYA**  
**2021/2022**

# RECOMMENDER SYSTEM (CONTENT-BASED FILTERING IN PYTHON)

## Dataset :

Movie\_metadata CSV

Berisi informasi tentang Film



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	color	director	num_critics	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	gross	genres	actor_1_name	movie_title	num_votes	cast_total	actor_3_name	num_votes	plot_keywords	movie_imdb_num_votes	language	country	content_rating	budget	title_year	actor_2_name	imdb_score	aspect_ratio	movie_facebook_likes			
2	Color	James Car	723	178	0	855	Joel David	1000	7.61E+08	Action	[Ac CCH Poun Avatar	886204	4834	Wes Studi	0	avatar	1	English	USA	PG-13	2.37E+08	2009	536	7.9	1.78	33000			
3	Color	Gore Ver	302	169	563	1000	Orlando B	40000	3.09E+08	Action	[Ac Johnny De Pirates of	471220	48350	Jack Davi	0	godde	1	English	USA	PG-13	3E+08	2007	5000	7.1	2.35	0			
4	Color	Sam Men	602	148	0	161	Rory Kinn	11000	2E+08	Action	[Ac Christoph Spectre	275868	11700	Stephanie	1	bomb	1	English	UK	PG-13	2.45E+08	2015	393	6.8	2.35	85000			
5	Color	Christop	813	164	22000	29000	Christan I	27000	4.48E+08	Action	[Th Tom Hard The Dark	1144337	106759	Joseph Gc	0	deceptio	1	English	USA	PG-13	2.5E+08	2012	23000	8.5	2.35	164000			
6		Doug Walk			131		Rob Walk	331			[Document Doug Walk Star Wars	8	345		0		1	http://www.imdb.com/title/tt0208954/?ref=fn_h1_1					12	7.1	0				
7	Color	Andrew Si	462	132	475	530	Samantha	640	7305879	Action	[Ac Daryl Sabi John Cart	212204	1873	Polly Wall	1	alien	1	English	USA	PG-13	2.64E+08	2012	632	6.6	2.35	24000			
8	Color	Sam Raim	392	156	0	4000	James Fra	24000	3.37E+08	Action	[Ac J.K. Simm Spider-Me	383056	46055	Kirsten Dc	0	sandman	1	English	USA	PG-13	2.58E+08	2007	11000	6.2	2.35	0			
9	Color	Nathan Gr	324	100	15	284	Donna Mu	799	2.01E+08	Adventure	[Brad Garri Tangle	294810	2038	M.C. Gain	1	17th cent	1	English	USA	PG	2.6E+08	2010	553	7.8	1.85	29000			
10	Color	Joss Whee	635	141	0	19000	Robert Do	26000	4.39E+08	Action	[Ac Chris Men Avengers	462689	92000	Scarlett Jc	4	artifici	1	English	USA	PG-13	2.5E+08	2015	21000	7.5	2.35	118000			
11	Color	David Yate	375	153	282	10000	Daniel Rai	25000	3.02E+08	Adventure	[Alan Ricki Harry Pott	321795	58753	Rupert Gri	3	blood	1	English	UK	PG	2.5E+08	2009	11000	7.5	2.35	10000			
12	Color	Jack Snyd	673	183	0	2000	Lauren Co	15000	3.3E+08	Action	[Ac Henry Cav Batman v	371639	24450	Alan D. Pu	0	based on	1	English	USA	PG-13	2.5E+08	2016	4000	6.9	2.35	197000			
13	Color	Bryan Sing	434	169	0	903	Marlon Br	18000	2E+08	Action	[Ac Kevin Spa Supermar	240396	29993	Frank Lang	0	crystal	1	English	USA	PG-13	2.09E+08	2006	10000	6.1	2.35	0			
14	Color	Marc For	403	106	395	393	Mathieu J	451	1.88E+08	Action	[Ac Giancarlo Quantum	330784	2023	Rory Kinn	1	action her	1	English	UK	PG-13	2E+08	2008	412	6.7	2.35	0			
15	Color	Gore Ver	313	151	563	1000	Orlando B	40000	4.23E+08	Action	[Ac Johnny De Pirates of	522040	48486	Jack Davi	2	box office	1	English	USA	PG-13	2.25E+08	2006	5000	7.3	2.35	5000			
16	Color	Gore Ver	450	150	563	1000	Ruth Willi	40000	8928910	Action	[Ac Johnny De The Lone	181792	45757	Tom Willi	1	horse	1	English	USA	PG-13	2.25E+08	2013	2000	6.5	2.35	48000			
17	Color	Jack Snyd	733	143	0	748	Christoph	15000	2.91E+08	Action	[Ac Henry Cav Man of St	548573	20495	Harry Leni	0	based on	1	English	USA	PG	2.5E+08	2013	3000	7.2	2.35	118000			
18	Color	Andrew A	258	150	80	201	Pierfrance	22000	1.42E+08	Action	[Ac Peter Dini The Chron	149922	22697	Dani	4	brother	1	English	USA	PG	2.25E+08	2008	216	6.6	2.35	0			
19	Color	Joss Whee	703	173	0	19000	Robert Do	26000	6.23E+08	Action	[Ac Chris Men The Aven	95415	87697	Scarlett Jc	3	alien inva	1	English	USA	PG-13	2.2E+08	2012	21000	8.1	1.85	123000			
20	Color	Rob Mar	448	136	252	1000	Sam Claffi	40000	2.41E+08	Action	[Ac Johnny De Pirates of	370704	54083	Stephen C	4	blackbe	1	English	USA	PG-13	2.5E+08	2011	11000	6.7	2.35	50000			
21	Color	Barry Soni	451	106	188	718	Michael Si	10000	1.79E+08	Action	[Ac Will Smi Men in Bil	268154	12572	Nicole Sc	1	alien	1	English	USA	PG-13	2.25E+08	2012	816	6.8	1.85	40000			
22	Color	Peter Jack	422	164	0	773	Adam Bro	5000	2.55E+08	Adventure	[Aidan Tun The Hobbi	354228	9152	James Nei	0	army	1	English	New Zeal	PG-13	2.5E+08	2014	972	7.5	2.35	65000			
23	Color	Marc Web	599	153	464	963	Andrew G	15000	2.62E+08	Action	[Ac Emma Sto The Amaz	451803	28489	Chris Zylu	0	liard	1	English	USA	PG-13	2.3E+08	2012	10000	7	2.35	56000			
24	Color	Ridley Sc	343	156	0	733	William H	891	1.05E+08	Action	[Ac Mark Add Robin Hec	211765	3244	Scott Grr	0	1150s	1	English	USA	PG-13	2E+08	2010	882	6.7	2.35	17000			
25	Color	Peter Jack	509	186	0	773	Adam Bro	5000	2.58E+08	Adventure	[Aidan Tun The Hobbi	483340	9152	James Nei	6	dwar	1	English	USA	PG-13	2.25E+08	2013	972	7.9	2.35	83000			
26	Color	Chris Wei	251	113	129	1000	Eva Green	16000	70083519	Adventure	[Christoph The Golde	149019	24106	Kristin Sc	2	children	1	English	USA	PG-13	1.8E+08	2007	6000	6.1	2.35	0			
27	Color	Peter Jack	446	201	0	84	Thomas Ki	6000	2.18E+08	Action	[Ac Naom Wing Kong	316018	7123	Evan Park	0	animal	1	English	New Zeal	PG-13	2.07E+08	2005	919	7.2	2.35	0			
28	Color	James Car	315	194	0	794	Kate Wind	29000	6.39E+08	Drama	[Ac Leonardo Titanic	793059	45223	Gloria Stu	0	art	1	English	USA	PG-13	2E+08	1997	14000	7.7	2.35	26000			
29	Color	Anthony F	516	147	94	11000	Scarlett Jc	21000	4.07E+08	Action	[Ac Robert Do Captain A	272670	64798	Chris Evar	0	based on	1	English	USA	PG-13	2.5E+08	2016	19000	8.2	2.35	72000			
30	Color	Peter Ber	377	131	532	627	Alexand	14000	65173160	Action	[Ac Liam Neer Battleship	203282	26679	Tadanobu	0	box office	1	English	USA	PG-13	2.09E+08	2012	10000	5.9	2.35	44000			
31	Color	Colin Trev	644	124	365	1000	Judy Gree	3000	6.52E+08	Action	[Ac Bryce Dall Jurassic W	418214	8458	Omar Sy	0	dinosaur	1	English	USA	PG-13	1.5E+08	2015	2000	7	2	150000			
32	Color	Sam Men	750	140	0	393	Helen Mc	883	3.04E+08	Action	[Ac Albert Fin Skyfall	522830	3059	Rory Kinn	0	brav	1	English	UK	PG-13	2E+08	2012	563	7.8	2.35	80000			
33	Color	Sam Raim	300	135	0	4000	James Fra	24000	3.73E+08	Action	[Ac J.K. Simm Spider-Me	411164	43388	Kirsten Dc	1	death	1	English	USA	PG-13	2E+08	2004	11000	7.3	2.35	0			
34	Color	Shane Bla	608	195	1000	3000	Jon Favre	21000	4.09E+08	Action	[Ac Robert Do Iron Man	557489	30426	Don Chea	3	armor	1	English	USA	PG-13	2E+08	2013	4000	7.2	2.35	95000			
35	Color	Tim Burto	451	108	13000	11000	Alan Rickr	40000	3.34E+08	Adventure	[Johnny De Alice in W	306320	79957	Anne Hat	0	alice in w	1	English	USA	PG	2E+08	2010	25000	6.5	1.85	24000			
36	Color	Brett Ratn	334	104	420	560	Kelsey Gri	20000	2.34E+08	Action	[Ac Hugh Jack N Men Th	383427	23744	Daniel Cu	0	battle	1	English	Canada	PG	2.1E+08	2006	808	6.8	2.35	0			
37	Color	Dan Scanl	376	104	37	760	Tyler Labi	12000	2.68E+08	Adventure	[Steve Bus Monsters	255025	14865	Sean Hay	0	cheatin	1	English	USA	G	2E+08	2013	779	7.3	1.85	44000			
38	Color	Michael B	366	150	0	464	Kevin Dur	884	4.02E+08	Action	[Ac Glenn Mo Transform	323207	3218	Ramon Ro	0	autob	1	English	USA	PG-13	2E+08	2009	581	6	2.35	0			
39	Color	Michael B	378	165	0	880	Sophia My	974	2.45E+08	Action	[Ac Bingbing L Transform	244240	3988	Kelsey Gri	2	blockbu	1	English	USA	PG-13	2.1E+08	2014	956	5.7	2.35	56000			

## Kode:

### 1. Import Library

```
import numpy as np
import pandas as pd
```

### 2. Import data

```
data = pd.read_csv('movie_metadata.csv')
data.describe()
```

Output

	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_1_facebook_likes	gross	num_voted_users
count	4993.000000	5028.000000	4939.000000	5020.000000	5036.000000	4.159000e+03	5.043000e+03
mean	140.194272	107.201074	686.509212	645.009761	6560.047061	4.846841e+07	8.366816e+04
std	121.601675	25.197441	2813.328607	1665.041728	15020.759120	6.845299e+05	1.384853e+05
min	1.000000	7.000000	0.000000	0.000000	0.000000	1.620000e+02	5.000000e+00
25%	50.000000	93.000000	7.000000	133.000000	614.000000	5.340988e+06	8.593500e+03
50%	110.000000	103.000000	49.000000	371.500000	988.000000	2.551750e+07	3.435900e+04
75%	195.000000	118.000000	194.500000	636.000000	11000.000000	6.230944e+07	9.630900e+04
max	813.000000	511.000000	23000.000000	23000.000000	640000.000000	7.605058e+08	1.689764e+06

### 3. Memilih kolom yang digunakan

```
# data yang digunakan
data = data[['movie_title', 'num_critic_for_reviews',
            'duration', 'gross', 'budget', 'title_year',
            'imdb_score']]
data.isna().sum()
```

movie_title	0
num_critic_for_reviews	50
duration	15
gross	884
budget	492
title_year	108
imdb_score	0
dtype: int64	

### 4. Menghapus data duplikat

```
# Menghapus data duplikat
data_no_duplicates = data.drop_duplicates()
data = data_no_duplicates
data_no_duplicates =
data.drop_duplicates(subset=['movie_title',
                              'title_year'])
data = data_no_duplicates
```

### 5. Menghapus spasi diawal dan diakhir judul

```
# menghapus spasi diawal dan diakhir judul
data['movie_title'] = data['movie_title'].str.strip()
```

### 6. Handling missing value

```
# handling missing value
data.fillna(method="bfill", inplace=True)
```

Output

```
1 data.head(10)
```

✓ 0.4s

	movie_title	num_critic_for_reviews	duration	gross	budget	title_year	imdb_score
0	Avatar	723.0	178.0	760505847.0	237000000.0	2009.0	7.9
1	Pirates of the Caribbean: At World's End	302.0	169.0	309404152.0	300000000.0	2007.0	7.1
2	Spectre	602.0	148.0	200074175.0	245000000.0	2015.0	6.8
3	The Dark Knight Rises	813.0	164.0	448130642.0	250000000.0	2012.0	8.5
4	Star Wars: Episode VII - The Force Awakens	462.0	132.0	73058679.0	263700000.0	2012.0	7.1
5	John Carter	462.0	132.0	73058679.0	263700000.0	2012.0	6.6
6	Spider-Man 3	392.0	156.0	336530303.0	258000000.0	2007.0	6.2
7	Tangled	324.0	100.0	200807262.0	260000000.0	2010.0	7.8
8	Avengers: Age of Ultron	635.0	141.0	458991599.0	250000000.0	2015.0	7.5
9	Harry Potter and the Half-Blood Prince	375.0	153.0	301956980.0	250000000.0	2009.0	7.5

## 7. Buat dataframe yang berisi variabel / feature

```
movie_features = data.loc[:,  
"num_critic_for_reviews":].copy()  
movie_features.head()
```

Output

	num_critic_for_reviews	duration	gross	budget	title_year	imdb_score
0	723.0	178.0	760505847.0	237000000.0	2009.0	7.9
1	302.0	169.0	309404152.0	300000000.0	2007.0	7.1
2	602.0	148.0	200074175.0	245000000.0	2015.0	6.8
3	813.0	164.0	448130642.0	250000000.0	2012.0	8.5
4	462.0	132.0	73058679.0	263700000.0	2012.0	7.1

## 8. Buat hitung cosine similarity

```
from sklearn.metrics.pairwise import cosine_similarity  
cosine_sim = cosine_similarity(movie_features.values,  
movie_features.values)
```

Output

```
array([[1.          , 0.89253031, 0.83431836, ... , 0.96824449, 0.96306654,
        0.95820985],
       [0.89253031, 1.          , 0.99327609, ... , 0.78975726, 0.77281149,
        0.72665644],
       [0.83431836, 0.99327609, 1.          , ... , 0.71684148, 0.69745632,
        0.6422833 ],
       ...,
       [0.96824449, 0.78975726, 0.71684148, ... , 1.          , 0.99961049,
        0.97925231],
       [0.96306654, 0.77281149, 0.69745632, ... , 0.99961049, 1.          ,
        0.98209637],
       [0.95820985, 0.72665644, 0.6422833 , ... , 0.97925231, 0.98209637,
        1.          ]])
```

## 9. Memberi Nomor pada judul film

```
# memberi nomor index
```

```
index = pd.Series(data.index, index=data['movie_title'])
```

Output

```
1 index.head(11)
```

✓ 0.4s

```
movie_title
Avatar                                0
Pirates of the Caribbean: At World's End  1
Spectre                               2
The Dark Knight Rises                   3
Star Wars: Episode VII - The Force Awakens  4
John Carter                             5
Spider-Man 3                             6
Tangled                                 7
Avengers: Age of Ultron                  8
Harry Potter and the Half-Blood Prince   9
Batman v Superman: Dawn of Justice       10
dtype: int64
```

## 10. Membuat model Content Based RS

```
def get_recommendation(movie_title,
similarity=cosine_sim):
```

```
    idx = index[movie_title]
```

```
    # Get the pairwise similarity scores of all movie
    sim_scores = list(enumerate(cosine_sim[idx]))
```

```

    # Sort the movie based on the similarity scores
    sim_scores = sorted(sim_scores, key=lambda x: x[1],
reverse=True)

    # Get the scores of the 10 most similar movie
    sim_scores = sim_scores[0:10]

    # Get the movie indices
    movie_indices = [i[0] for i in sim_scores]

    # Return the top 20 most similar movie
    result = data[['movie_title']].iloc[movie_indices]
    return result

```

## 11. Output Hasil Recommender System

```

# Mencetak 20 Film rekomendasi yang mirip dengan Iron
Man
print('Film yang mirip dengan Hulk')
get_recommendation('Hulk')

```

Output

## Film yang mirip dengan Hulk

movie_title	
168	G.I. Joe: Retaliation
2673	The Marine
983	The Judge
3503	Bride & Prejudice
432	Jumper
1949	The Losers
1252	Autumn in New York
761	Joy
2245	Georgia Rule
1959	The Box