HỆ KHUYẾN NGHỊ

BÀI TẬP THỰC HÀNH TUẦN 3

KHUYẾN NGHỊ DỰA TRÊN NỘI DUNG

1. **Quy định về việc nộp bài**
   - Thời gian: Được giảng viên thiết lập trên hệ thống Moodle.
   - Hình thức nộp: Trên Moodle.
   - Bài nộp bao gồm các file .**ipynb** trong một folder và nén lại thành một tập tin (**.zip**).
   - Cách đặt tên: **BTTH3_MSSV.zip**
   - Công cụ thực hành. **Google colab**
   - Lưu ý: Sai quy định thì sẽ nhận 0 điểm.

2. **Nội dung thực hành**

   2.1. Theo dõi giảng viên hướng dẫn

   DL mẫu:

```
train_set = ["Bộ phim kể về cuộc chiến <html> đấu chống lại <html> những linh hồn ác quỷ.",
          "Cuộc đột nhập # $ của Thom một kỹ sư thiên tài vào một ngân hàng bí ẩn tại Hồng Kông.",
          "Phim về những rắc rối hài hước đời thường và thế giới tình yêu trong độ tuổi 30.",
          "Về các siêu <url> anh hùng cùng hợp tác và (chống) # lại mối nguy hiểm mang quy mô quốc tế."]


test_set = ["Phim kể về tình yêu hài hước và quá trình sống chung rắc rối giữa một biên kịch với một diễn viên nam."]
```

- Tiền xử lý dữ liệu
  - Tách từ trên tiếng Anh và tiếng Việt
  - Loại bỏ stopword từ trên tiếng Anh và tiếng Việt

    Link stopword tiếng Việt: https://www.kaggle.com/mpwolke/vietnamese-stopwords-w2v/notebook
  - Chuẩn hóa từ

   - Chuyển chuỗi về hoa, thường, xóa các thẻ HTML, dấu câu, … từ biểu thức chính quy (Regular Expression)
   - Tùy vào miền dữ liệu chúng ta có thể xử lý emoji.
- Xây dựng ma trận vector TF-IDF
- Khuyến nghị sản phẩm cho người dùng nhờ vào độ đo Cosine

2.2.Yêu cầu về nhà

Dựa trên overview của các bộ phim trong bộ dữ liệu "The Movies Dataset" cài đặt thuật toán khuyến nghị dựa trên dung.

- Link download: https://www.kaggle.com/rounakbanik/the-movies-dataset

	A	B	C	D	E	F	G	H	I	J	K	
1	adult	belongs_to_collection	budget	genres	homepage	id	imdb_id	original_language	original_title	overview	popularity	poster_path
2	FALSE	{'id': 10194, 'name': 'Toy S	30000000	[{'id': 16, 'nan	http://toystory	862	tt0114709	en	Toy Story	Led by Woody, Andy's toys live happily in his room until Andy's birthday bring	21.946.943	/rhIRbceoE9lR
3	FALSE		65000000	[{'id': 12, 'name': 'Adventure'		8844	tt0113497	en	Jumanji	When siblings Judy and Peter discover an enchanted board game that opens	17.015.539	/vzmL6fP7aPK
4	FALSE	{'id': 119050, 'name': 'Gru	0	[{'id': 10749, 'name': 'Roman		15602	tt0113228	en	Grumpier Old Men	A family wedding reignites the ancient feud between next-door neighbors ai	117.129	/6ksm1sjKMFL
5	FALSE		16000000	[{'id': 35, 'name': 'Comedy'},		31357	tt0114885	en	Waiting to Exhale	Cheated on, mistreated and stepped on, the women are holding their breath	3.859.495	/16XOMpEaLV
6	FALSE	{'id': 96871, 'name': 'Fath	0	[{'id': 35, 'name': 'Comedy'}]		11862	tt0113041	en	Father of the Bride Part	Just when George Banks has recovered from his daughter's wedding, he rec	8.387.519	/e64sOI48hQX
7	FALSE		60000000	[{'id': 28, 'name': 'Action'}, {'i		949	tt0113277	en	Heat	Obsessive master thief, Neil McCauley leads a top-notch crew on various ins	17.924.927	/zMyfPUelumi
8	FALSE		58000000	[{'id': 35, 'name': 'Comedy'},		11860	tt0114319	en	Sabrina	An ugly duckling having undergone a remarkable change, still harbors feeling	6.677.277	/jQh15y5YB7b
9	FALSE		0	[{'id': 28, 'name': 'Action'}, {'i		45325	tt0112302	en	Tom and Huck	A mischievous young boy, Tom Sawyer, witnesses a murder by the deadly In	2.561.161	/sGO5Qa55p7
10	FALSE		35000000	[{'id': 28, 'name': 'Action'}, {'i		9091	tt0114576	en	Sudden Death	International action superstar Jean Claude Van Damme teams with Powers I	523.158	/eoWvKD60lTS
11	FALSE	{'id': 645, 'name': 'James I	58000000	[{'id': 12, 'nan	http://www.m	710	tt0113189	en	GoldenEye	James Bond must unmask the mysterious head of the Janus Syndicate and p	14.686.036	/5c0ovjT41KnY
12	FALSE		62000000	[{'id': 35, 'name': 'Comedy'},		9087	tt0112346	en	The American President	Widowed U.S. president Andrew Shepherd, one of the world's most powerfu	6.318.445	/lymPNGLZgPI
13	FALSE		0	[{'id': 35, 'name': 'Comedy'},		12110	tt0112896	en	Dracula: Dead and Lovin	When a lawyer shows up at the vampire's doorstep, he falls prey to his char	5.430.331	/xve4cgfYItnOI
14	FALSE	{'id': 117693, 'name': 'Balt	0	[{'id': 10751, 'name': 'Family'		21032	tt0112453	en	Balto	An outcast half-wolf risks his life to prevent a deadly epidemic from ravagin	12.140.733	/gV5PCAVCPN
15	FALSE		44000000	[{'id': 36, 'name': 'History'}, {'		10858	tt0113987	en	Nixon	An all-star cast powers this epic look at American President Richard M. Nixo	5.092	/cICkmCEiXRhy
16	FALSE		98000000	[{'id': 28, 'name': 'Action'}, {'i		1408	tt0112760	en	Cutthroat Island	Morgan Adams and her slave, William Shaw, are on a quest to recover the t	7.284.477	/odM9973klv9
17	FALSE		52000000	[{'id': 18, 'name': 'Drama'}, {'		524	tt0112641	en	Casino	The life of the gambling paradise ‚Äì Las Vegas ‚Äì and its dark mafia undert	10.137.389	/xo517ibXBDd
18	FALSE		16500000	[{'id': 18, 'name': 'Drama'}, {'		4584	tt0114388	en	Sense and Sensibility	Rich Mr. Dashwood dies, leaving his second wife and her daughters poor by	10.673.167	/lA9HTy84Bb6
19	FALSE		4000000	[{'id': 80, 'name': 'Crime'}, {'i		5	tt0113101	en	Four Rooms	It's Ted the Bellhop's first night on the job...and the hotel's very unusual gue	9.026.586	/eQs5hh9rxrk1
20	FALSE	{'id': 3167, 'name': 'Ace Ve	30000000	[{'id': 80, 'name': 'Crime'}, {'i		9273	tt0112281	en	Ace Ventura: When Nat	Summoned from an ashram in Tibet, Ace finds himself on a perilous journey	8.205.448	/wRlGnJhEzcx
21	FALSE		60000000	[{'id': 28, 'name': 'Action'}, {'i		11517	tt0113845	en	Money Train	A vengeful New York transit cop decides to steal a trainload of subway fares	7.337.906	/jSozzzVOR2kl
22	FALSE	{'id': 91698, 'name': 'Chili	30250000	[{'id': 35, 'name': 'Comedy'},		8012	tt0113161	en	Get Shorty	Chili Palmer is a Miami mobster who gets sent by his boss, the psychopathic	12.669.608	/vWtDUUgQA:
23	FALSE		0	[{'id': 18, 'name': 'Drama'}, {'		1710	tt0112722	en	Copycat	An agoraphobic psychologist and a female detective must work together to t	10.701.801	/80czeJGSoik2
24	FALSE		50000000	[{'id': 28, 'name': 'Action'}, {'i		9691	tt0112401	en	Assassins	Assassin Robert Rath arrives at a funeral to kill a prominent mobster, only tc	11.065.939	/xAx5MP7Dg4
25	FALSE		0	[{'id': 18, 'name': 'Drama'}, {'		12665	tt0114168	en	Powder	Harassed by classmates who won't accept his shocking appearance, a shy yo	12.133.094	/1uRKsxOCtgz
26	FALSE		3600000	[{'id': 18, 'nan	http://www.m	451	tt0113627	en	Leaving Las Vegas	Ben Sanderson, an alcoholic Hollywood screenwriter who lost everything bec	10.332.025	/37qHRJxnSh5
27	FALSE		0	[{'id': 18, 'name': 'Drama'}]		16420	tt0114057	en	Othello	The evil Iago pretends to be friend of Othello in order to manipulate him to s	1.845.899	/qM0BXEQjmr
28	FALSE		12000000	[{'id': 35, 'name': 'Comedy'},		9263	tt0114011	en	Now and Then	Waxing nostalgic about the bittersweet passage from childhood to puberty i	8.681.325	/wD6rLdD2Ix3
29	FALSE		0	[{'id': 18, 'name': 'Drama'}, {'		17015	tt0114117	en	Persuasion	This film adaptation of Jane Austen's last novel follows Anne Elliot, the daug	2.228.434	/si8911lezMvA
30	FALSE		18000000	[{'id': 14, 'name': 'Fantasy'}, {		902	tt0112682	fr	La CitV© des Enfants Pe	A scientist in a surrealist society kidnaps children to steal their dreams, hopi	9.822.423	/eVo6ewq4akl
31	FALSE		0	[{'id': 18, 'name': 'Drama'}, {'		37557	tt0115012	zh	ÊëáÂïäÊëáÔºåÊëáÂà∞Â	A provincial boy related to a Shanghai crime family is recruited by his uncle i	1.100.915	/qcoOCoN7viC
32	FALSE		0	[{'id': 18, 'name': 'Drama'}, {'		9909	tt0112792	en	Dangerous Minds	Former Marine Louanne Johnson lands a gig teaching in a pilot program for	9.481.338	/y5Jee3QmYO
33	FALSE		29500000	[{'id': 878, 'name': 'Science F		63	tt0114746	en	Twelve Monkeys	In the year 2035, convict James Cole reluctantly volunteers to be sent back ir	12.297.305	/6Sj9wDu3Yug
34	FALSE		0	[{'id': 10749, 'name': 'Roman		78802	tt0114952	fr	Guillaumet, les ailes du courage		0.745542	/k6QDtR38dKF

- Sử dụng file **movies_metadata.csv**: Chứa thông tin của hơn 45.000 bộ phim có trong bộ dữ liệu Full MovieLens. Bộ dữ liệu gồm nhiều cột khác nhau, tuy nhiên chúng ta sẽ sử dụng cột "**Overview**" nội dung chính của phim để xây dựng hệ khuyến nghị của mình.
- Thử chọn ra top 10 bộ phim cho một người dùng đã từng xem một bộ phim có tên là "**Father of the Bride Part II**"

- Sử dụng file **Họ tên_MSSV.ipynb** được cung cấp sẵn trong phần nộp BTTH 3 và hoàn thành các yêu cầu ở phần #code here