HỆ KHUYẾN NGHỊ

BÀI TẬP THỰC HÀNH TUẦN 3

KHUYÉN 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:

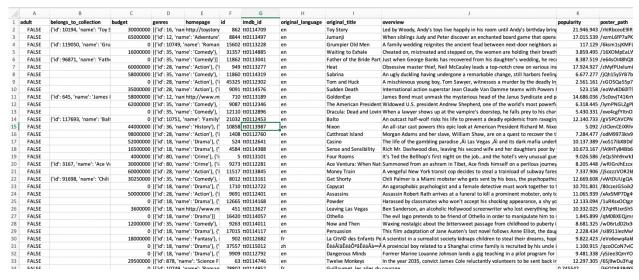
- Tiền xử lý dữ liệu
 - o Tách từ trên tiếng Anh và tiếng Việt
 - Loại bỏ stopword từ trên tiếng Anh và tiếng Việt
 Link stopword tiếng Việt: https://www.kaggle.com/mpwolke/vietnamese-stopwords-w2v/notebook
 - Chuẩn hóa từ

- Chuyển chuỗi về hoa, thường, xóa các thẻ HTML, dấu câu, ... từ biểu thức chính quy (Regular Expression)
- O 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



- 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