

HỆ KHUYẾN NGHỊ

BÀI TẬP THỰC HÀNH TUẦN 4

CẢI THIẾN CONTENT-BASED BẰNG WORD-EMBEDDING

ĐÁNH GIÁ KHUYẾN NGHỊ

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: **BTTH4_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

- Sử dụng các bộ pre-trained word-embedding để chuyển nội đối tượng sang không gian vector, để cải thiện kết quả của khuyến nghị dựa trên nội dung.

Link WE: <https://github.com/sonvx/word2vecVN>

Hoặc link: <https://drive.google.com/drive/folders/1DPqDm2TpeT060Iadop6q-GrjNyuFd9o8?usp=sharing>

- Đánh giá hệ khuyến nghị dựa trên lọc cộng tác: MSE, RMSE, MAE, NMAE.
- Đánh giá hệ khuyến nghị dựa trên nội dung: P@K, R@K, F1@K, NDCG.

2.2. Yêu cầu về nhà

- Cài đặt lại bài tập thực hành số 2 (CF) trên bộ dữ liệu MovieLens.
 - + Cắt từ tập ratings.csv khoảng 200 dòng để làm tập test.
 - + Sử dụng phương pháp khuyến nghị dựa trên người dùng để dự đoán rating cho 200 dòng ở trên (Sử dụng lại source code trong Lab3)
 - + Sử dụng rating vừa dự đoán để tính toán mức độ hiệu quả của hệ khuyến nghị.
 - + Đánh giá theo các độ đo: **MSE, RMSE, MAE, NMAE.**

- Cài đặt lại bài tập thực hành số 3 (CB).

Với userID = 201, các bạn lấy ra danh sách hết tất cả những bộ phim mà userID này đã đánh giá. (Kết quả đúng là 110 bộ phim)

- + Cắt ra 100 bộ phim đầu tiên làm tập test.
- + Sử dụng overview của 10 bộ phim còn lại để biểu diễn nội dung của người dùng userID = 201 này.
- + Sử dụng lại code khuyến nghị dựa trên nội dung của của lab 3 để dự đoán danh sách 100 bộ phim liên quan nhất với userID = 201. (tập dự đoán)
- + Đánh giá theo các độ đo: P@K, R@K, F1@K. (K=100) (trên tập test và tập dữ đoán)
- + Sắp xếp thứ tự các bộ phim trong tập test theo thứ tự rating, đánh giá độ đo MRR, NDCG trên tập test đã sắp xếp này và tập dự đoán.

*** Chú ý: Sử dụng TF-IDF và Word-embedding để vector hóa, sau đó nhận xét kết quả mô hình nào tốt hơn.