Genome tag trong hệ thống gợi ý

1. Ý tưởng:
   1. Dùng thư viện lightfm:
      * lightfm không có sẵn implementation cho genome tag
      * Coi bảng *genome-scores.csv* là interaction matrix với movieId đóng vai trò là userId, tagId đóng vai trò itemId và relevance là rating của từng film cho từng tag một
      * Em không implement theo cách này vì:
        + overfit do dataset bao gồm relevance của tất cả các tag cho từng item một
   2. Không dùng thư viẹn:
      * Coi tất cả relevance của từng bộ phim như một điểm rồi tính khoảng cách giữa các điểm đó. Lọc lấy các điểm gần với item mốc nhất.
2. Cách triển khai:
   * Tạo một bảng có kich thước (số tag) \* (số item). Mỗi hàng trong bảng nói trên (tạm gọi là temp) ứng với một item (vì bảng cho sẵn viết theo cột dọc)
   * Tạo một ma trận có kích thước (số item) \* (số item). Mỗi ô (i, j) trong ma trận tương ứng với khoảng cách giữa 2 điểm tượng trưng cho itemId = i + 1 và itemId = j + 1 (itemId bắt đầu từ 1)
   * Khi cần gợi ý cho itemId = id thì ta lọc ra cột (hoặc hàng) id - 1
   * Trong hàng ta vừa lọc ra (tạm gọi là tmp) sắp xếp các giá trị khoảng cách theo thứ tự tăng dần, lọc lấy 5 id có giá trị khoảng cách thấp nhất (ngoại trừ item id có khoảng cách 0).
3. Khó khăn:
   * Ban đầu train với điểm có relevance của tất cả các tag (1128) dẫn đến overfit và thời gian chạy lớn.
   * Sử dụng thước đo khoảng cách là công cụ không mạnh (bằng đo góc giữa 2 vecto từ 2 điểm đó)
   * Việc chuyển qua lại giữa id và index dễ gây nhầm lẫn
4. Cái tiến lần một:
   * Chuyển sang coi mỗi hàng trong temp như một vecto, chuyển sang dùng cos giữa các vecto để đánh giá.
   * tmp sẽ được lọc theo thứ tự giảm dần, tạo một dataframe từ tmp (để giữ index) , sort giảm dần theo giá trị của cos, lọc lấy 5 giá trị lớn nhất khác 1)
   * Áp dụng thêm PCA để giảm chiều của các vecto. Thông số n\_components lấy theo reseach paper Mohd Ali, Syed & Nayak, Dr & Barik, Dr. Rabindra & Lenka, Rakesh. (2017). Movie Recommendation System using Genome Tags and Content-based Filtering.