1. **Giới thiệu về đề tài:**

Đề tài của nhóm chúng em là Tìm hiểu thuật toán recommendation, search để xây dựng trang web tìm kiếm phim.

Tổng quan về đề tài:

Đề tài là một công cụ tìm kiếm và gợi ý với mục đích là truy xuất phim dựa vào thông tin được cung cấp của người dùng, dựa trên thuật toán Okapi-BM25. Và thuật toán recommendation lựa chọn ra 30 bộ phim đề xuất sử dụng thuật toán knn. Dữ liệu được lưu trữ trong Mongodb local và code được viết bằng python với front-end là HTML. Trước khi được tải lên dữ liệu đánh giá về một bộ phim đã được phân chia và gán nhãn những đánh giá tích cực và tiêu cực. Do dữ liệu thu thập review về các phim có hạn nên nhóm chúng em chỉ load được 500 phim được review.

Điểm nổi bật của đề tài: truy vấn, tìm kiếm nhanh một bộ phim dựa vào nhiều yếu tố (tên diễn viên, tên đạo diễn, nhà sản xuất phim, nội dung phim, tiêu đề phim, thể loại phim) từ đó đưa ra các phim đề xuất cho người dùng, phần review chúng em lấy từ trang imdb là những review đánh giá có chiều sâu từ phía người dùng

Đề tài của nhóm em chú trọng vào nghiên cứu và tìm hiểu thuật toán.

1. **Datasets:**

* tmdb\_5000\_movies: Datasets được thu thập thông tin từ IMDB có sẵn trên Kaggle. Chứa khoảng 5000 bộ phim gồm các thuộc tính: tiêu đề (title), tổng quan (overview), ngày phát hành (release date), đạo diễn (main\_actor), công ty sản xuất (production companies), nước sản xuất (production countries), thể loại (genres), ngôn ngữ (original languages), mức độ phổ biến (popularity), thời lượng (runtime), điểm trung bình imdb (IMDb average score), số phiếu bầu (votes), doanh thu (budget)
* MovieGenre: là tập dữ liệu gồm link để download các poster. Nhóm em code 1 đoạn script python để download các poster từ link này
* movies\_metadata: Bao gồm 50.000 bộ phim từ imdb là siêu dữ liệu gồm tất cả các thông tin chi tiết về một bộ phim được lấy từ Kaggle nhóm em dùng tập dữ liệu này để lấy imdb\_id của từng bộ phim để load các review từ trang imdb

[TMDB 5000 Movie Dataset | Kaggle](https://www.kaggle.com/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv)

1. **Thuật toán BM25:**

Trong tìm kiếm thông tin, Okapi BM25 là hàm tính thứ hạng (xếp hạng) được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định. Hàm xếp hạng này dựa trên mô hình xác suất, được phát minh ra vào những năm 1970 – 1980. Phương pháp có tên BM25 (BM – best match), nhưng người ta thường gọi "Okapi BM25", vì lần đầu tiên công thức được sử dụng trong hệ thống tìm kiếm Okapi, được sáng lập tại trường đại học London những năm 1980 và 1990.

BM25 là một phương pháp xếp hạng được sử dụng rộng rãi trong tìm kiếm. Trong Web search những hàm xếp hạng này thường được sử dụng như một phần của các phương pháp tích hợp để dùng trong machine learning.

Áp dụng vào đề tài:

Ban đầu ta có một hàm f. Sau đó nhận đầu vào là một văn bảng x, hàm này sẽ chia văn bản thành các từ đơn chuyển nó thành chữ thường loại bỏ các ký tự đặc biệt và các stopword. Sau đó sẽ return về một danh sách các từ, tiếp theo sẽ so sánh các từ đó với tiêu đề phim, nội dung, đạo diễn, diễn viên, nhà sản xuất để chọn ra các bộ phim liên quan.

Ở đây thuật toán BM-25 được chọn để truy xuất các bộ phim dựa trên truy vấn của người dùng truy vấn từ cơ sở dữ liệu chứa ít nhất một từ liên quan đến truy vấn của người dùng, trong đó thuộc tính tiêu đề phim sẽ được ưu tiên hơn khi tìm kiếm.

Đối với các phim được tìm kiếm lần đầu sẽ được lưu vào bộ nhớ cache, tăng tốc độ truy xuất cho mỗi lần tiếp theo

Ví dụ: ta tìm kiếm phim Avengers heroes and Kung Fu

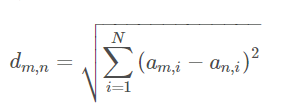
|  |  |  |
| --- | --- | --- |
| Search | Avengers heroes | Kung Fu |
| 1 | The Avengers | Kung Fu Panda |
| 2 | Avengers: Age of Ultron | Kung Fu Panda 2 |
| 3 | The Toxic Avenger | Kung Fu Panda 3 |
| 4 | The Toxic Avenger Part II | Legend of a Rabbit |
| 5 | Last Action Hero | Bulletproof Monk |

1. **Thuật toán recommendation**:

Mục tiêu của công cụ đề xuất là đề xuất n phim có nội dung tương tự đến một bộ phim mục tiêu (content-based). Trong đề tài của chúng em sẽ đề xuất 15 bộ phim. Để tính toán sự giống nhau giữa phim mục tiêu và các phim được đề xuất trước tiên chúng em xây dựng ma trận tương tự (similar matrix) với các cột là tên tiêu đề phim(title), đạo diễn (director), diễn viên (actor), các thể loại (geners) và từ khóa của phim (keywords). Mỗi hàng là mỗi bộ phim.

Ngoài tiêu đề phim có kiểu dữ liệu là string thì các cột còn lại là các giá trị Boolean. Nếu phim I có thể loại trùng với phim mục tiêu t thì giá trị trong ma trận của thể loại đó = 1 và ngược lại bằng 0.

Sau khi hình thành ma trận similar, tiếp theo sử dụng công thức tính khoảng cách (hay còn gọi là mức độ liên quan) của phim mục tiêu với phim các phim trong database sẽ chọn ra 15 phim có khoảng cách gần với phim mục tiêu nhất

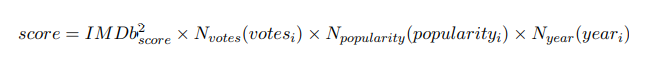


Với: m là phim mục tiêu, n là các phim được chọn, I là các thuộc tính

Khi đã có N phim giống nhau được đề xuất, tiếp theo em sẽ chọn ra 15 bộ phim dựa vào 3 tiêu chí:

* Điểm IMDB
* Số phiếu bầu Votes
* Độ phổ biến của phim (popularity)
* Năm sản xuất (release\_Date)

Đối với điểm IMDB em sẽ nhân đôi vì nó là biến quan trọng đến đề xuất điểm IMDB càng cao thì người dùng càng ưu tiên lựa chọn. Đối với các tiêu chí còn lại em **sử dụng hàm Gaussian để đặt nặng các thuộc tính** như điểm IMDB, số phiếu bầu (votes), Độ phổ biến của phim (popularity), và năm sản xuất (release\_Date)

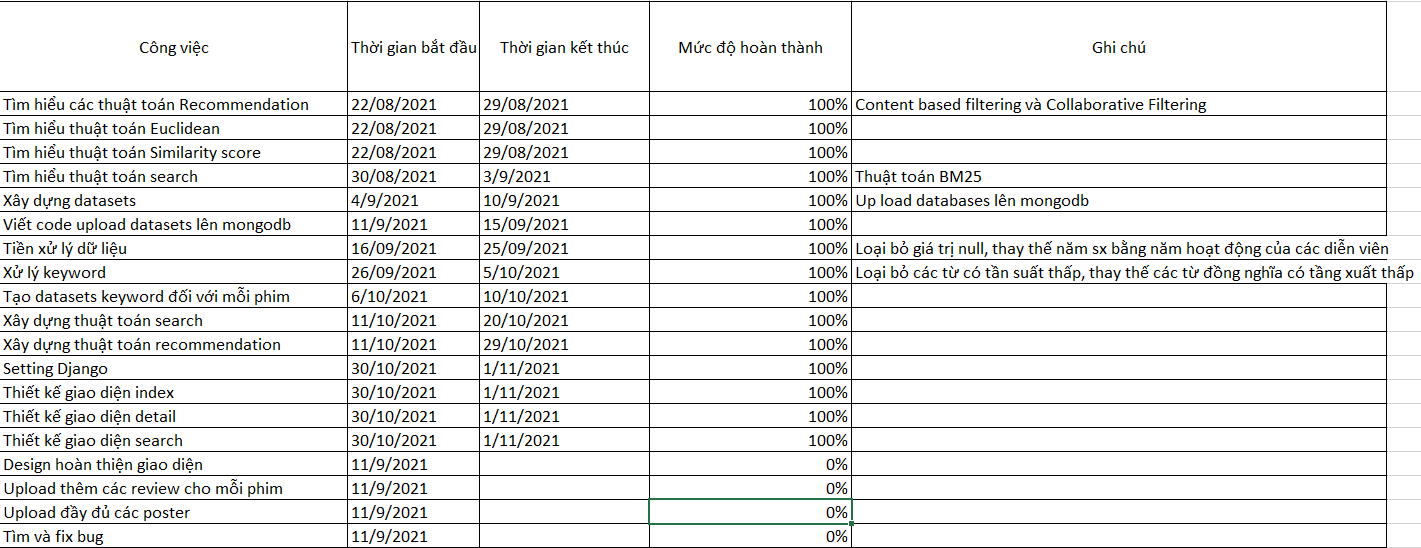


* Score càng cao thì được ưu tiên đề xuất hàng đầu, nếu chỉ dựa vào mức độ liên quan giữa các bộ phim thôi thì chưa đủ nên ở bước này em tính score và đặt nặng các thuộc tính đó thì hệ thống gợi ý sẽ hiệu quả hơn rất nhiều

1. **Phân loại review:**

Em có tìm được một trang web có ghi lại reviews của các user trên trang imdb.com và lưu lại thành các file text có gắn nhãn pos và neg tức là những bình luận tích cực và tiêu cực tiếp theo chúng em upload lên mongodb, do số lượng reviews có hạn nên không phải phim nào cũng có review

1. **Kế hoạch**



1. **Tổng kết**

Trang web được thiết kế đơn giản nhưng mang lại rất nhiều hiệu quả. Khắc phục được lỗi truy xuất chậm, ý tưởng thuật toán hiệu quả. Với công cụ search đa năng, các phim review chất lượng với ưu tiên xếp hạng các phim có điểm imdb cao