

`  

**ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**------------------------**

**BÁO CÁO ĐỒ ÁN**

**ỨNG DỤNG TÌM KIẾM PHIM**

**DỰA TRÊN MÔ TẢ**

*Môn học*: **Truy vấn thông tin đa phương tiện**

*Giảng viên*: TS. Nguyễn Vinh Tiệp

*Danh sách thành viên*:

Võ Huy Khôi 18520949

Nguyễn Thịnh Quyền 18521322

Hứa Văn Sơn 18521344

**PHÂN CÔNG THÀNH VIÊN**

|  |  |  |  |
| --- | --- | --- | --- |
| **Tên** | **Võ Huy Khôi** | **Hứa Văn Sơn** | **Nguyễn Thinh Quyền** |
| **Mã số sinh viên** | 18520949 | 18521344 | 18521322 |
| **Công việc thực hiện** | Lọc data, tiền xử lý ban đầu, xây dựng bảng kết quả. | Cài đặt thuật toán truy vấn | Cào data, xây dựng giao diện ứng dụng |

1. **Tổng quan đề tài**

Hiện nay có rất nhiều trang web có thể tìm kiếm phim dựa trên các từ khóa như tên phim, tên diễn viên,.. nhưng chưa có nhiều web có thể tìm kiếm phim dựa trên mô tả về nội dung hay tên nhân vật trong phim, đặc biệt là các trang web tìm kiếm bằng tiếng Việt. Vì vậy, đề tài của nhóm mục tiêu là tạo ra một ứng dụng sơ khai bằng python để có thể giúp người dùng có thể tìm kiếm phim theo nội dung mà mình muốn xem.

1. **Hướng tiếp cận**
2. **Mô tả data**

Bộ data được sử dụng cho đề tài được cào từ trang web phim dongphym.net (<https://dongphym.net/>) bằng thư viện beautiful soup của python. Bộ data bao gồm hơn 8000 phim khác nhau bao gồm cả phim bộ và phim lẻ với đầy đủ mọi thể loại.

Data phim được cào với các thông tin bao gồm: đường dẫn (URL), tên tiếng Việt, tên tiếng Anh, thể loại và nội dung mô tả phim. Sau đó, bộ data được lọc để loại bỏ các phim trùng lặp và phim bị thiếu thông tin về nội dung.

1. **Tiền xử lý cho việc truy vấn**

Khi người dùng nhập vào một câu query tìm kiếm, trong câu thường có những từ ngữ bị nhiễu, không quan trọng còn gọi là stopwords vì vậy nhóm sử dụng phương pháp loại bỏ các stopwords trong cả câu query tìm kiếm và cả phần nội dung phim.

***Danh sách stopwords:***

* 'bị', 'bởi', 'cả', 'các', 'cái', 'cần', 'càng', 'chỉ','chiếc', 'cho', 'chứ', 'chưa', 'chuyện',

'có', 'có\_thể', 'cứ', 'của', 'cùng', 'cũng', 'đã', 'đang', 'đây', 'để', 'đến\_nỗi', 'đều',

'điều', 'do', 'đó', 'được', 'dưới', 'gì', 'khi', 'không', 'là','lại', 'lên', 'lúc', 'mà', 'mỗi',

'một\_cách', 'này', 'nên', 'nếu', 'ngay', 'nhiều', 'như', 'nhưng', 'những','nơi', 'nữa',

'phải', 'qua', 'ra', 'rằng', 'rằng', 'rất', 'rất', 'rồi', 'sau', 'sẽ', 'so', 'sự', 'tại', 'theo',

           'thì', 'trên', 'trước', 'từ', 'từng', 'và', 'vẫn', 'vào', 'vậy', 'vì', 'việc', 'với', 'vừa'.

Để đảm bảo mô hình có thể truy vấn chính xác nhất và hiểu được đầy đủ ý nghĩa của câu truy vấn, nhóm sử dụng phương pháp word segmentation cho cả cây query và cả phần nội dung phim để gộp các từ ghép ví dụ từ :“khổng lồ” sẽ được xem như một từ: “khổng\_lồ”.

Ngoài ra, nhóm còn sử dụng phương pháp pos tagging để gán nhãn từ loại cho các từ trong câu query để giữ lại những từ khóa quan trọng trong câu. Các nhãn được giữ lại: 'N': Common Noun, 'Nc': Noun Classifier, 'Ny': Noun Abbreviation, 'Np': proper noun, 'Nu': Unit Noun, 'A': adjective, 'V': Verb.

1. **Phương pháp truy vấn**

**3.1. Đo độ tương phản của văn bản**

* Term frequency (TF): mô hình hóa văn bản về các vector, sử dụng mô hình bag of words trong đó thứ tự của các từ sẽ bị loại bỏ, giá trị của các thành phần trong vector này được tính bằng cách tính tỉ lệ xuất hiện của các từ trong văn bản đó.
* TF-IDF: sử dụng để làm nổi bật các từ quan trọng: là những từ xuất hiện nhiều trong một văn bản nhưng ít xuất hiện trong các văn bản khác. IDF được tính bằng công thức:



Số từ sử dụng trong các văn bản càng nhiều thì log dần tiến về 0 tương đương với từ này kém giá trị, ngược lại số từ sử dụng trong các văn bản càng ít thì log sẽ tiến về giá trị lớn hơn.

**3.2. Truy vấn thông tin**

Thêm câu query truy vấn vào cuối bảng data, xây dựng ma trận TF-IDF biểu diễn vector của tập data với số chiều là 15750 x 10000.

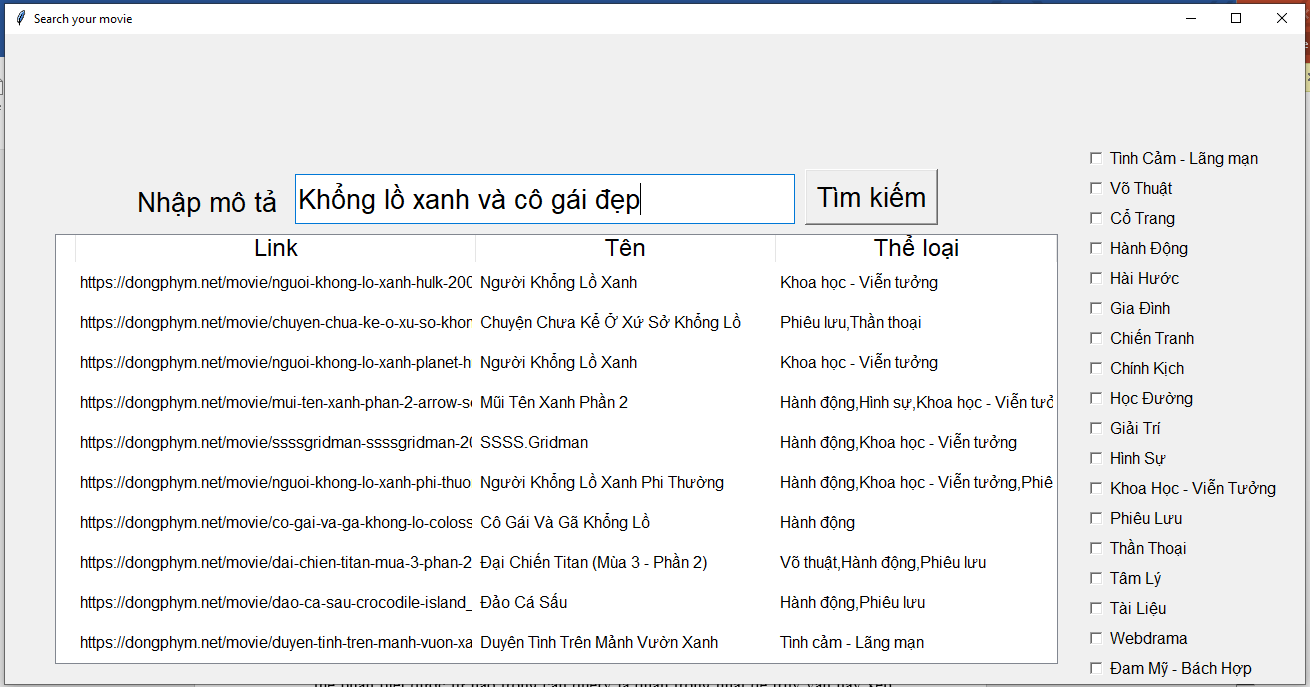
Sử dụng độ đo cosin để tính độ tương đồng giữa 2 vector, càng giống nhau thì giá trị càng gần 1, ngược lại sẽ càng gần 0.



Sử dụng K nearest neighbor để tìm K = 10 văn bản có độ tương đồng cao nhất với câu query trong ngữ liệu thông qua ma trận vector TF-IDF.

1. **Ứng dụng**

Ứng dụng được xây dựng bằng thư viện tkinter của python, bao gồm thanh tìm kiếm, thể loại phim để người dùng chọn và kết quả trả về là bảng gồm 5 phim liên quan nhất đến câu query.



1. **Kết quả**

Mô hình đưa ra các kết quả dự đoán khá tốt, mặc dù vẫn còn hạn chế như chưa thể phân biệt được từ nào trong câu query là quan trọng nhất để truy vấn hay xếp hạng kết quả phim chưa có độ chính xác cao.

**TÀI LIỆU THAM KHẢO**

[1] Động phim, <https://dongphym.net/>

[2] Ông Xuân Hồng – Truy vấn văn bản – Document Retrieval, <https://ongxuanhong.wordpress.com/2017/01/16/truy-van-van-ban-document-retrieval/\>