BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO
THỰC TẬP TỐT NGHIỆP
ĐẠI HỌC

***Đề tài*:**

"XÂY DỰNG HỆ THỐNG KHAI PHÁ Ý KIẾN NGƯỜI XEM PHIM TRỰC TUYẾN"

Người hướng dẫn : ThS. NGUYỄN NGỌC DUY
Sinh viên thực hiện : NGUYỄN THỊ THANH HẰNG
Mã số sinh viên : N14DCCN142
Lớp : D14CQIS01 - N
Khoá : 2014
Hệ : ĐẠI HỌC CHÍNH QUY

TP.HCM, tháng 08/2018

BỘ THÔNG TIN VÀ TRUYỀN THÔNG
HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÁO CÁO
THỰC TẬP TỐT NGHIỆP
ĐẠI HỌC

***Đề tài*:**

"XÂY DỰNG HỆ THỐNG KHAI PHÁ Ý KIẾN NGƯỜI XEM PHIM TRỰC TUYẾN"

Người hướng dẫn : ThS. NGUYỄN NGỌC DUY
Sinh viên thực hiện : NGUYỄN THỊ THANH HẰNG
Mã số sinh viên : N14DCCN142
Lớp : D14CQIS01 - N
Khoá : 2014
Hệ : ĐẠI HỌC CHÍNH QUY

TP.HCM, tháng 08/2018

HỌC VIỆN
CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
CƠ SỞ TẠI TP. HỒ CHÍ MINH
Khoa Công nghệ thông tin 2

CỘNG HOÀ XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. Hồ Chí Minh, ngày 5 tháng 5 năm 2018

PHIẾU ĐĂNG KÝ THỰC TẬP TỐT NGHIỆP

Kính gửi: Lãnh đạo Khoa Công nghệ thông tin 2

1. **Họ và tên sinh viên: Nguyễn Thị Thanh Hằng** **Mã SV: N14DCCN142**
   Lớp: D14CQIS01-N Ngành: Công nghệ Thông tin. Hình thức đào tạo: Chính Quy.
2. **Tên đề tài: Xây dựng hệ thống khai phá ý kiến người xem phim trực tuyến.**
3. **Nơi đăng ký thực tập: Trung tâm Đào tạo SaigonLab**
   Đơn vị chủ quản:
   Địa chỉ: 643, Điện Biên Phủ, Phường 1, Quận 3, tp.HCM
   Số ĐT: 0901 838 638 **Số Fax:**
4. **Nội dung thực hiện:**

- **Lý thuyết:**
  + Tìm hiểu sử dụng Java và MongoDB.
  + Tìm hiểu kỹ thuật xây dựng một từ điển cảm xúc.
  + Tìm hiểu các giải thuật rút trích đặc trưng trong văn bản tiếng Việt.
  + Tìm hiểu kỹ thuật phân loại ý kiến trong các văn bản tiếng Việt.
- **Thực hành:**
  + Phân tích, thiết kế cơ sở dữ liệu cho hệ thống.
  + Xây dựng một từ điển cảm xúc theo chủ đề phim đơn giản.
  + Xây dựng website đơn giản có các chức năng:
    - Có khả năng quản lý và chạy file video trực tuyến.
    - Ghi nhận và lưu trữ các ý kiến đánh giá của người xem.
    - Phân loại và xếp hạng các bộ phim dựa trên các ý kiến đánh giá.

5. **Giáo viên hướng dẫn: ThS. Nguyễn Ngọc Duy**
6. **Thời gian thực hiện:**
   Từ ngày tháng năm 201 đến ngày tháng năm 201

TRƯỞNG BỘ MÔN **GIÁO VIÊN HƯỚNG DẪN** **SINH VIÊN ĐĂNG KÝ**

LỜI CẢM ƠN

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Trong suốt thời gian từ khi bắt đầu học tập tại trường đến nay, em đã nhận được rất nhiều sự quan tâm, giúp đỡ của quý thầy cô, gia đình và bạn bè. Với lòng biết ơn sâu sắc nhất, em xin gửi đến quý thầy cô ở khoa Công Nghệ Thông Tin 2 – Học viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại TP.HCM đã tận tâm chỉ bảo, truyền đạt vốn kiến thức quý báu cho chúng em trong suốt thời gian học tập tại trường. Và đặc biệt trong học kỳ này, nếu không có những lời hướng dẫn, dạy bảo của các thầy cô thì em nghĩ bài báo cáo này của em rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn thầy cô. Bài báo cáo thực tập thực hiện trong khoảng thời gian 10 tuần. Bước đầu đi vào thực tế của em còn hạn chế và còn nhiều bỡ ngỡ. Do vậy, không tránh khỏi những thiếu sót là điều chắc chắn, em rất mong nhận được những ý kiến đóng góp quý báu của quý Thầy Cô và các bạn học cùng lớp để kiến thức của em trong lĩnh vực này được hoàn thiện hơn.

Em xin gửi lời cảm ơn chân thành và sự tri ân sâu sắc đối với các thầy cô ở khoa Công Nghệ Thông Tin 2 – Học viện Công Nghệ Bưu Chính Viễn Thông cơ sở tại TP.HCM đã tạo điều kiện cho em để em có thể hoàn thành tốt bài báo cáo thực tập này. Đặc biệt em xin gửi lời cảm ơn đến thầy ThS.Nguyễn Ngọc Duy với những buổi vừa học kết hợp với thực hành trên lớp cùng những buổi nói chuyện rất hữu ích đã giúp em định hướng và hoàn thành tốt đề tài này.

Trong quá trình làm bài báo cáo, cũng như là trong quá trình làm đề tài khó tránh khỏi sai sót, rất mong các Thầy, Cô bỏ qua. Đồng thời do trình độ lý luận cũng như kinh nghiệm thực tiễn còn hạn chế nên bài báo cáo không thể tránh khỏi những thiếu sót, em rất mong nhận được ý kiến đóng góp Thầy, Cô để em học thêm được nhiều kinh nghiệm làm hành trang vững chắc để em tự tin theo đuổi sự nghiệp của mình.

Em xin chân thành cảm ơn!

Thành phố Hồ Chí Minh, tháng 8 năm 2018

Nguyễn Thị Thanh Hằng

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM

Độc lập- Tự do- Hạnh phúc

TP. Hồ Chí Minh, ngày tháng năm 20……

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN
THỰC TẬP TỐT NGHIỆP ĐẠI HỌC

1. **Tên đề tài:**
2. **Sinh viên:** **Lớp:**
3. **Giáo viên hướng dẫn:**
4. **Nơi công tác:**

NỘI DUNG NHẬN XÉT

1. Đánh giá chung:

……
……

2. Đánh giá chi tiết:

……
……
……
……

3. Nhận xét về tinh thần, thái độ làm việc:

……
……

4. Kết luận:

……
……

5. Điểm hướng dẫn ():

GIẢNG VIÊN HƯỚNG DẪN

(Ký, ghi rõ họ tên)

MỤC LỤC

DANH MỤC HÌNH

LỜI MỞ ĐẦU

Lý do chọn đề tài

Ngày nay với tốc độ phát triển của khoa học kỹ thuật trên thế giới ngày càng mạnh mẽ. Cuộc cách mạng công nghệ thông tin đã và đang diễn ra trên hầu hết các nước tiên tiến trên thế giới. Bên cạnh việc bạn phải ra ngoài rạp phim xa xôi giữa thời tiết nắng nóng chỉ để chờ đợi xếp hàng mua vé cho một bộ phim bom tấn thì với công nghệ hiện đại ngày nay bạn hoàn toàn có thể ngồi nhà thưởng thức bộ phim với độ phân giải cao không kém gì ngoài rạp chỉ bằng những thao tác đơn giản click vào một trang web xem phim trực tuyến bất kỳ mà không tốn bất kỳ phí thu nào. Từ đó, rất nhiều ứng dụng đặc biệt là những ứng dụng được phát triển trên nền web được xây dựng nhằm đáp ứng nhu cầu giải trí đó của người dùng, các ứng dụng phát triển liên tục và nhanh chóng theo sự phát triển của xã hội về qui mô và chất lượng.

Hiện nay, các ứng dụng như vậy đa phần dựa theo các tiêu chí như lượt xem (view), lượt yêu thích (like) … để đánh giá một video.

Trong quá trình học tại trường, em có học và tìm hiểu môn học Khai phá dữ liệu do thầy Nguyễn Ngọc Duy giảng dạy. Em thấy có một nguồn dữ liệu rất đáng giá để đánh giá các video đó chính là các bình luận (comment) về video mà người dùng xem. Từ các bình luận đó, nếu thu thập đầy đủ, lưu trữ lại, ứng dụng các giải thuật rút trích hợp lý sẽ phân loại được những từ khóa (keyword) miêu tả cảm xúc của người dùng đối với các video. Những từ khóa được đánh những trọng số khác nhau biểu diễn mức độ cảm xúc của người dùng với video đó.

Từ ý tưởng đó nên em quyết định chọn đề tài đồ án thực tập tốt nghiệp là “Xây dựng hệ thống khai phá ý kiến người xem phim trực tuyến” như là một bước đầu để thể nghiệm ý tưởng đánh giá video dựa trên từ ngữ biểu thị cảm xúc.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1.1.Tổng quan

Đề tài "Xây dựng hệ thống khai phá ý kiến người xem phim trực tuyến" là một ứng dụng trên nền web. Người dùng truy cập và sẽ lựa chọn xem các video được phân loại dựa vào thể loại, quốc gia, … Ứng dụng có các chức năng như: tìm kiếm video theo thể loại hoặc theo quốc gia, xem các video, người dùng sử dụng tài khoản gmail để đăng nhập vào website thì có thể bình luận vào các video nếu muốn, cùng một số chức năng của quản trị viên (admin).

1.2. Tìm hiểu khai phá ý kiến người xem phim trực tuyến

- **Lý thuyết:**
  + Tìm hiểu sử dụng Java và MongoDB.
  + Tìm hiểu kỹ thuật xây dựng một từ điển cảm xúc.
  + Tìm hiểu các giải thuật rút trích đặc trưng trong văn bản tiếng Việt.
  + Tìm hiểu kỹ thuật phân loại ý kiến trong các văn bản tiếng Việt.
- **Thực hành:**
  + Phân tích, thiết kế cơ sở dữ liệu cho hệ thống.
  + Xây dựng một từ điển cảm xúc theo chủ đề phim đơn giản.
  + Xây dựng website đơn giản có các chức năng:
    - Có khả năng quản lý và chạy file video trực tuyến.
    - Ghi nhận và lưu trữ các ý kiến đánh giá của người xem.
    - Phân loại và xếp hạng các bộ phim dựa trên các ý kiến đánh giá.

1.3. Công nghệ ứng dụng

- MongoDB
- Node JS
- Java

1.4. Công cụ sử dụng

- Visual Studio Code 1.25.0
- Netbean IDE 8.3
- RoboMongo 1.2.1

1.5. Mục tiêu đề tài

- Hệ thống phải dễ sử dụng, có tính khả thi, đầy đủ thông tin, tránh dư thừa thông tin.
- Xem và bình luận được các video.
- Mới mẻ trong cách đánh giá video giải trí.

1.6. Lĩnh vực

Chuyên ngành : Hệ thống Thông tin

Chuyên môn : Lập trình ứng dụng web xem phim trực tuyến. Sử dụng ngôn ngữ Java cùng công nghệ Webservice bằng Node JS kết nối dữ liệu từ hệ cơ sở dữ liệu MongoDB.

Lĩnh vực liên quan : Giải trí.

1.7. Chức năng chính

- Cơ sở dữ liệu: Bình luận của người dùng được lưu vào database dưới dạng document trong MongoDB nhờ công cụ trực quan RoboMongo.
- Trang chủ : Chứa thông tin các bộ phim lẻ, phim chiếu rạp hot.
- Trang Menu :
    + Search bar : Tìm kiếm phim nhanh theo tên phim, thể loại, quốc gia.
    + Table view : Hiển thị các mục Thể loại, Quốc gia, Phim lẻ, TOP Phim.
- Trang Thông tin: Hiển thị thông tin phim (Tên phim, thể loại, quốc gia, đạo diễn, diễn viên, thời lượng,đánh giá…). Mức độ Khen, chê bộ phim (dựa vào phân loại cảm xúc bình luận người xem phim để đánh giá ý kiến)
- Trang Xem phim: Hiển thị video cho phép chạy phim trực tuyến cho người dùng xem.

CHƯƠNG 2: GIỚI THIỆU CÁC CÔNG NGHỆ VÀ CÔNG CỤ SỬ DỤNG

2.1. Giới thiệu các công nghệ

2.1.1. MongoDB

2.1.1.1. Đặt vấn đề

Với sự phát triển không ngừng của ngành công nghệ thông tin. Khối dữ liệu cần xử lý trong các ứng dụng là rất lớn. Đặc biệt là sự bùng nổ công nghệ Web 2.0, nơi các mạng dịch vụ dữ liệu cộng đồng cho phép người dùng tự do tạo nội dung trên web, dẫn đến dữ liệu tăng lên rất nhanh, vượt qua giới hạn xử lý của các Hệ quản trị cơ sở dữ liệu quan hệ truyền thống. Để đáp ứng được nhu cầu phát triển của xã hội, đòi hỏi một cơ sở dữ liệu (CSDL) có thể lưu trữ, xử lý được một lượng dữ liệu lớn một cách nhanh chóng và hiệu quả. NoSQL đã ra đời, thay thế hệ quản trị CSDL quan hệ, giải quyết bài toán trên.

2.1.1.2. Giới thiệu về NOSQL

Với hầu hết các thời kỳ web, Hệ quản trị cơ sở dữ liệu quan hệ dựa trên SQL đã thống trị hầu hết các hệ Quản trị Cơ sở dữ liệu. Tuy nhiên, thời gian gần đây, một cách tiếp cận mới đã bắt đầu biết đến là NoSQL, tạo ra sự thay thế cho các hệ quản trị cơ sở dữ liệu quan hệ truyền thống.

NoSQL còn có nghĩa là Non-Relational - không ràng buộc. Tuy nhiên, thuật ngữ đó ít phổ dụng hơn và ngày nay người ta thường dịch NoSQL thành Not Only SQL - Không chỉ SQL. NoSQL ám chỉ đến những cơ sở dữ liệu không dùng mô hình dữ liệu quan hệ để quản lý dữ liệu trong lĩnh vực phần mềm. Thuật ngữ NoSQL được giới thiệu lần đầu vào năm 1998 sử dụng làm tên gọi chung cho các cơ sở dữ liệu quan hệ nguồn mở nhỏ nhưng không sử dụng SQL cho truy vấn.

Database là một cơ sở dữ liệu, gồm các bảng, hàng, cột chứa dữ liệu cho hệ thống. Ta dễ dàng theo tác truy vấn dữ liệu qua các hệ quản trị cơ sở dữ liệu (Database Management System – DBMS) như MongoDB Compass, MongoBooster, Oracle, Microsoft SQL Server, MySQL,…

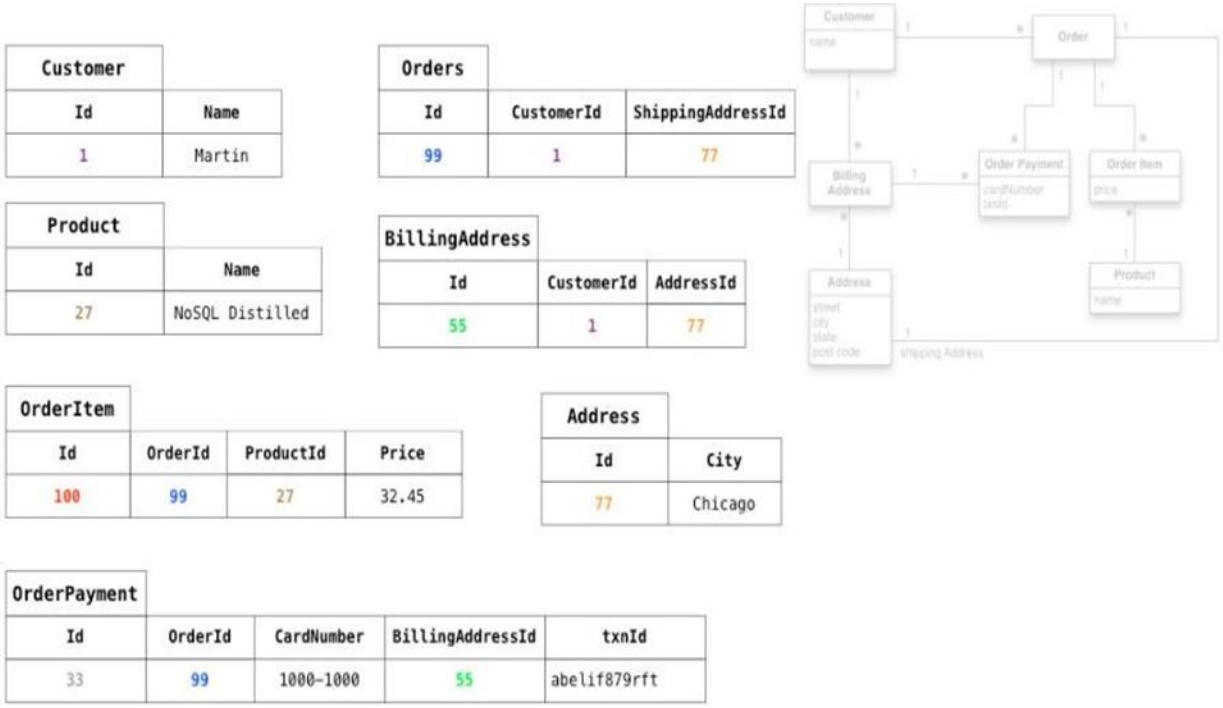
Structured Query Language – SQL là ngôn ngữ truy vấn cấu trúc trên nền một hệ quản trị cơ sở dữ liệu quan hệ (Relational Database Management System – RDBMS) như Oracle, MySQL,… được sử dụng rất rộng rãi vì một số lý do:

- Tính ACID (Atomicity, Consistency, Isolation, Durability) của một transaction được đảm bảo.
- Với database chuẩn 3, dữ liệu được đảm bảo tính đồng nhất và toàn vẹn (consistency).
- Có rất nhiều driver cho mọi ngôn ngữ: Java, C#, PHP.

- Số lượng lập trình viên biết và dùng SQL rất nhiều.

Tuy nhiên, RDBMS vẫn còn một số khuyết điểm:

- Việc mapping giữa các bảng trong database với các object trong code khá rắc rối và phức tạp. (*Mặc dù 1 số ORM như Entity Framework, Hibernate đã đơn giản hóa chuyện này).*
- Performance sẽ bị chậm khi phải join nhiều bảng để lấy dữ liệu (Đó là lý do ta sử dụng "giảm chuẩn" để tăng hiệu suất cho RDBMS).
- Việc thay đổi cấu trúc dữ liệu (Thêm/xóa bảng hoặc thêm/xóa một field) rất mệt mỏi, kéo theo vô số thay đổi trên code.
- Không làm việc được với dữ liệu không có cấu trúc (un-structured).
- RDBMS được thiết kế để chạy trên một máy chủ. Khi muốn mở rộng, nó khó chạy trên nhiều máy (clustering).



Customer

Id	Name
1	Martin

Orders

Id	CustomerId	ShippingAddressId
99	1	77

Product

Id	Name
27	NoSQL Distilled

BillingAddress

Id	CustomerId	AddressId
55	1	77

OrderItem

Id	OrderId	ProductId	Price
100	99	27	32.45

Address

Id	City
77	Chicago

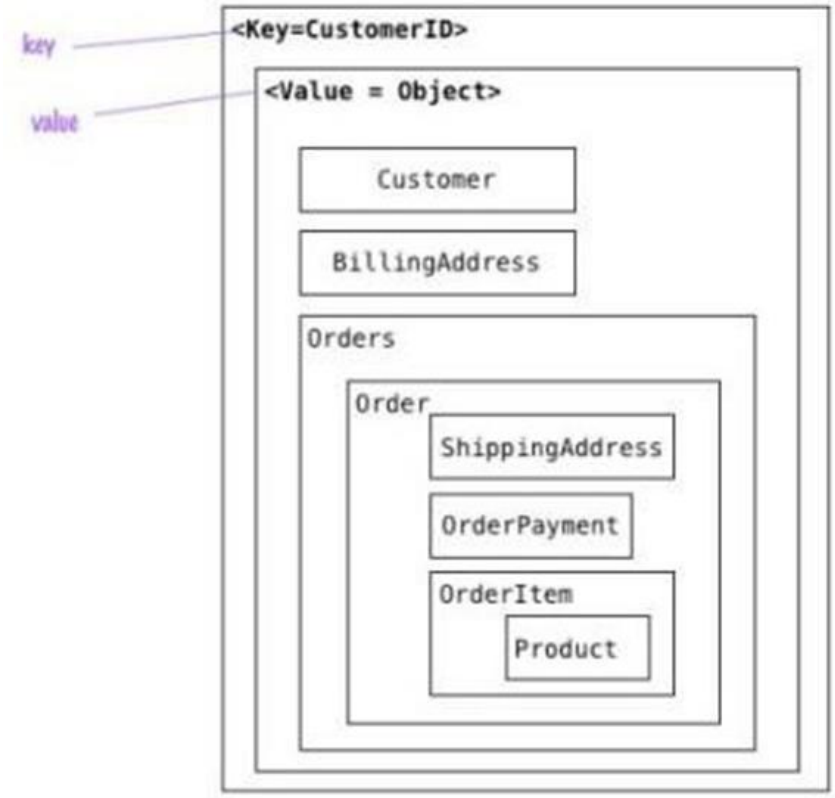
OrderPayment

Id	OrderId	CardNumber	BillingAddressId	txnId
33	99	1000-1000	55	abelif879rft

Hình 2.1. Dữ liệu được lưu thành nhiều bảng trong RDBMS, khi query ta phải join lại rất khó khăn.

NoSQL Database ra đời, giải quyết được những khuyết điểm của RDBMS:

- Dữ liệu trong NoSQL DB được lưu dưới dạng document, object. Truy vấn dễ dàng và nhanh hơn RDBMS nhiều.
- NoSQL có thể làm việc hoàn toàn tốt với dữ liệu dạng không có cấu trúc.
- Việc đổi cấu trúc dữ liệu (Thêm, xóa trường hoặc bảng) rất dễ dàng và nhanh gọn trong NoSQL.
- Vì không đặt nặng tính ACID của transactions và tính nhất quán của dữ liệu, NoSQL DB có thể mở rộng, chạy trên nhiều máy một cách dễ dàng.



Hình 2.2. Mô hình khối một document lưu thông tin customer.

Hình 2.3. Dữ liệu được lưu dưới dạng object. Mặc dù bị trùng nhưng truy vấn rất nhanh và đơn giản.

NoSQL bỏ qua tính toàn vẹn của dữ liệu và transaction để đổi lấy hiệu suất nhanh và khả năng mở rộng (scalability). Với những ưu điểm trên, NoSQL đang được sử dụng nhiều trong các dự án Big Data, các dự án Real-time, số lượng dữ liệu nhiều.

2.1.1.3. Hệ cơ sở dữ liệu MongoDB

Trong những gương mặt góp phần làm suy tàn đế chế SQL thì MongoDB nổi lên là một CSDL đáng tin cậy và dễ dùng nhất. Mongo viết bằng C++. Nó thích hợp cho các ứng dụng tầm trung trở lên. Nếu tỉ lệ lượng dữ liệu ghi vào CSDL của ứng dụng lớn hơn lượng đọc thì đây càng là lựa chọn hợp lý.

MongoDB là một CSDL có khả năng mở rộng, hiệu suất cao, mã nguồn mở và hướng văn bản.

Một số khái niệm cơ bản của MongoDB:

- Văn bản (Document) là đơn vị cơ bản của dữ liệu trong MongoDB, nó tương đương với một dòng trong CSDL quan hệ
- Bộ sưu tập (Collection) có thể được coi như tương đương với một bảng.
- MongoDB có thể lưu trữ nhiều CSDL độc lập, mỗi CSDL này có các bộ sưu tập và điều khoản riêng của mình
- MongoDB đi kèm với một trình tiện ích JavaScript đơn giản nhưng mạnh mẽ, nó hữu ích trong quản trị và thao tác dữ liệu.
- Mỗi văn bản có một khóa đặc biệt, đó là "_id", nó là duy nhất trong bộ sưu tập của văn bản.

Văn bản

Văn bản là một khái niệm quan trọng trong MongoDB. Văn bản bao gồm tập hợp các khóa với các giá trị tương ứng.

Ví dụ: {"greeting" : "Hello, world!"}

Văn bản trên gồm một khóa là "greeting", với giá trị là "Hello, world!". Các văn bản có thể chứa nhiều cặp khóa/giá trị.

Ví dụ: {"greeting" : "Hello, world!", "foo" : 3}.

Một số lưu ý:

- Các cặp khóa/ giá trị trong văn bản được sắp xếp. Văn bản trên sẽ khác với văn bản sau

{"foo" : 3, "greeting" : "Hello, world!"}

- Khóa trong văn bản là một chuỗi
- MongoDB phân biệt chữ hoa chữ thường
- Văn bản trong MongoDB không được chứa những khóa giống nhau.

Ví dụ văn bản sau là không hợp lệ:

{"greeting" : "Hello, world!", "greeting" : "Hello, MongoDB!"}

Bộ sưu tập

Bộ sưu tập là một nhóm các văn bản. Nếu văn bản tương đương với dòng trong CSDL quan hệ thì bộ sưu tập tương đương với bảng.

Bộ sưu tập là một Schema-Free, nghĩa là các văn bản có hình dạng khác nhau có thể cùng được lưu trữ trong 1 bộ sưu tập.

Ví dụ các văn bản sau có thể cùng được lưu trong một bộ sưu tập:

{"greeting" : "Hello, world!"}

{"foo" : 5}

Bộ sưu tập được xác định bởi tên của nó là một chuỗi UTF-8

Các đặc trưng của MongoDB:

- Lưu trữ hướng văn bản: Văn bản theo phong cách JSON với những lược đồ động đơn giản
- Hỗ trợ chỉ mục đầy đủ: chỉ mục trên bất kỳ các thuộc tính
- Tính sao lặp và tính sẵn sàng cao: mở rộng
- Auto-sharding: mở rộng theo chiều ngang mà không ảnh hưởng đến chức năng
- Truy vấn: đa dạng, truy vấn dựa trên văn bản
- Cập nhật nhanh:
- Map/Reduce
- GridFS: lưu trữ file với bất kỳ kích cỡ nào mà không làm phức tạp ngăn xếp
- Hỗ trợ thương mại: hỗ trợ doanh nghiệp, đào tào, tư vấn.

2.1.2. Node JS

Node.js là một phần mềm mã nguồn mở được viết dựa trên ngôn ngữ JavaScript cho phép lập trình viên có thể xây dựng các ứng dụng chạy trên máy chủ. Ban đầu, Node.js được phát triển bởi Ryan Dahl. Phiên bản đầu tiên của Node.js được cho ra mắt vào năm 2009.

Node.js có thể chạy được trên nhiều nền tảng khác nhau như Windows, Linux hay Mac OS. Node.js được phát triển sử dụng V8 Engine là bộ thư viện JavaScript được Google phát triển để viết trình duyệt web Chrome. Bản thân Node.js không phải là một ngôn ngữ lập trình mới, thay vào đó Node.js là một nền tảng mã nguồn mở (hay phần mềm mã nguồn mở) được viết dựa trên ngôn ngữ JavaScript.

Node.js có thể được dùng để tạo các ứng dụng chạy trên môi trường máy chủ như các ứng dụng web. Tuy nhiên Node.js không chỉ giới hạn ở việc tạo các website mà nó còn có thể được dùng để phát triển các công cụ chạy trên máy tính cá nhân.

Trong khi JavaScript thường được dùng trên trình duyệt thì Node.js lại được sử dụng để phát triển ứng dụng chạy trên máy chủ server. Node.js cũng có thể được chạy như một ứng dụng độc lập trên máy tính cá nhân (mà không cần phải thông qua môi trường của trình duyệt). Nói chính xác hơn thì chúng ta không thể chạy Node.js sử dụng môi trường trình duyệt.

2.1.3. JAVA

2.1.3.1. Tổng quan

Java là một ngôn ngữ lập trình và nền tảng điện toán trình độ cao chạy trên hơn 850 triệu máy tính cá nhân và hàng tỉ thiết bị trên toàn cầu, kể cả các công cụ di động và TV. Đây là công nghệ hình thành nên những chương trình tiên tiến như các tiện ích, trò chơi, và ứng dụng doanh nghiệp.

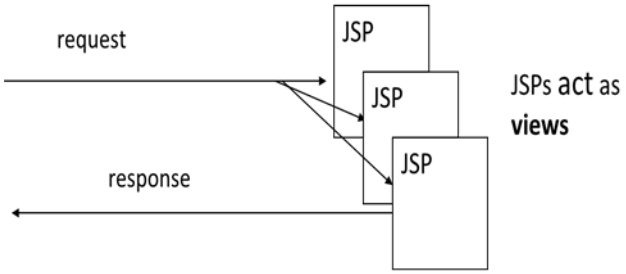
Công nghệ Java là một môi trường lập trình hướng đối tượng, độc lập với nền tảng, đa tuyến.Cho dù là phát triển các ứng dụng cho người tiêu dùng hoặc các nhiệm vụ quan trọng, bạn có thể tin tưởng vào một môi trường lập trình đồng nhất.

Lợi ích của java:

- Đơn giản - Java đã lược bỏ nhiều phần phức tạp của C++ và C, cho ra đời một ngôn ngữ đơn giản hơn (không con trỏ, không tổ hợp không thừa kế).
- Hướng đối tượng - Java là một ngôn ngữ hướng về vật thể gốc đơn, thừa kế đơn.
- Lập trình đa tuyến - Java có sự hỗ trợ được tích hợp sẵn cho lập trình đa tuyến.
- Phân tán - Khi Sử dụng Java RMI (gọi hàm từ xa) bạn có thể truy cập các vật thể trên các máy khác như thểtrên các máy nội bộ.
- Di động - Chương trình được viết trong ngôn ngữ Java độc lập với nền tảng

2.1.3.2. Mô hình một lớp V (View)

+ Mô hình:



Hình 2.4. Mô hình 1 lớp View

+ Ý nghĩa:

Lớp view:

file html hoặc jsp.

Tiếp nhận yêu cầu, xử lý yêu cầu, truy xuất CSDL.

Hiện giao diện/hiện thông báo lỗi/hiện kết quả.

+ Ưu/khuyết điểm:

Ưu điểm:

Lập trình đơn giản, thực thi nhanh.

Phù hợp với ứng dụng nhỏ, đơn giản.

Khuyết điểm:

Khi viết ứng dụng lớn sẽ rất khó sửa lỗi, khó mở rộng.

Không che dấu mã nguồn.

Một người phải biết tất cả:

- html, html5, css, javascript, frontpage, dreamweaver, ...: để thiết kế giao diện

- java, jsp,servlet, jme, hibernate, spring, strut,...: để lập trình web động.

- mysql, sql server, oracle ... và ngôn ngữ SQL: để sử dụng csdl và lập trình database.

- Biết nghiệp vụ của công ty để viết phần xử lý nghiệp vụ…

2.2. Giới thiệu công cụ

2.2.1. Visual Studio Code

Visual Studio Code là công cụ soạn thảo mã nguồn (code) được phát triển bởi Microsoft và có thể chạy trên Windows, Linux và MacOS. Nó hỗ trợ debug, nhúng Git control, highlight cú pháp, đoán code thông minh, snippets và code refactor.

Phiện bản ổn định hiện tai là 1.25 phát hành ngày 5/06/2018.

2.2.2. Netbean IDE

Netbeans là một dự án mã nguồn thành công với quy mô sử dụng rộng lớn, một cộng đồng đang phát triển mạnh và có gần 100 (và vẫn còn tăng) đối tác trên toàn thế giới. Công ty Sun Microsystems đã khởi đầu cho dự án mã nguồn mở này vào tháng 6 năm 2000 và vẫn tiếp tục là người tài trợ chính cho dự án. Vào thời điểm hiện tại đang có 2 sản phẩm: Netbeans IDE và Netbeans Platforms.

NetBean IDE là một công cụ dành cho lập trình viên để viết, biên dịch, gỡ lỗi (debug) và triển khai (deploy) chương trình. Chương trình được viết bằng Java nhưng có thể hỗ trợ bất kỳ ngôn ngữ lập trình nào. Có một số lượng rất lớn các module cho phép mở rộng Netbeans IDE. Netbeans IDE là một sản phẩm miễn phí và không có giới hạn nào trong việc sử dụng nó.

NetBean IDE là một "môi trường phát triển tích hợp" (Integrated Development Environment) kiểu như Visual Studio của Microsoft và được xem là một một bộ ứng dụng "must-download" dành cho các nhà phát triển phần mềm.

2.2.3. RoboMongo

RoboMongo là một công cụ trực quan giúp quản lý Database MongoDB. Nó là một phần mềm mã nguồn mở miễn phí, hỗ trợ cả 3 loại hệ điều hành Windows, Linux, Mac OS.

CHƯƠNG 3: CƠ SỞ LÝ THUYẾT

3.1. Kỹ thuật xây dựng một từ điển cảm xúc

Từ điển thông thường bao gồm các từ và ý nghĩa của từ, giúp cho người dùng tra cứu nghĩa của từ. Với sự phát triển của thời đại công nghệ thông tin, từ điển cũng phát triển với rất nhiều dạng, chẳng hạn như từ điển hành vi trong nghiên cứu hành vi con người, từ điển biểu cảm khuôn mặt trong nhận dạng sinh trắc,… Trong đề tài này, em muốn hướng tới việc xây dựng một từ điển từ khóa cảm xúc.

Từ điển từ khóa cảm xúc dùng để tra cứu cảm xúc/ý kiến của người dùng trong một văn bản. Từ điển chứa nhiều từ khóa, mỗi từ khóa mang một trong số biểu diễn mức độ cảm xúc của người dùng.

VD: Hiện tại từ điển có 2 từ khóa: “hay” trọng số là 1 , “rất hay” trọng số là 2

Trong một video, phần bình luận có một bình luận như sau: “hay, phim này rất hay”. Với kỹ thuật rút trích đặc trưng sẽ đề cập tới trong phần kế tiếp thì em đã tách được từ khóa “hay” với trọng số 1 và từ khóa “rất hay” trọng số 2. Sau đó sử dụng phương pháp tính trọng số cảm xúc ở phần kế tiếp để biết được bình luận này có ý kiến khen, chê hay là không quan tâm.

Để tiếp cận sơ lược thì từ điển này rất dễ để xây dựng, yếu tố cần thiết gồm có: từ điển tương ứng với ngôn ngữ cần xử lý và kỹ thuật phân loại tính từ với trọng số phù hợp. Nhưng thực tế, để áp dụng tốt nhất từ điển từ khóa cảm xúc, cần đến yếu tố thứ ba là con người nhằm đánh giá khách quan và chính xác nhất mức độ của các từ khóa. Cho nên, trong giới hạn thời gian và con người, em tạm thời xây dựng một từ điển khá đơn giản với ý kiến đánh trọng số chủ quan. Mong các thầy cô thông cảm để em có thể hoàn thành tốt hơn từ điển trong thời gian sắp tới.

3.2. Giải thuật rút trích đặc trưng trong văn bản tiếng Việt

Đối với các ngôn ngữ như tiếng Anh, tiếng Pháp, tiếng Đức việc tách từ được thực hiện khá đơn giản dựa vào các ký tự phân cách như: khoảng trắng, ký tự tab, các dấu câu, dấu ngoặc,... Ngược lại, đối với tiếng Việt (và các ngôn ngữ châu Á khác như tiếng Trung Quốc, tiếng Nhật Bản, tiếng Hàn) khoảng trắng ngoài việc ngăn cách các từ với nhau, còn được dùng để ngăn cách các âm tiết (syllable) của một từ ghép, ví dụ: câu “Học sinh đi học” phải được tách thành “Học_sinh/đi_học”. Khoảng trắng thứ nhất và thứ ba dùng để ngăn cách các âm tiết của một từ và khoảng trắng thứ hai dùng để ngăn cách hai từ với nhau.

Điều này gây khó khăn cho quá trình tách từ. Các phương pháp tách từ tiếng Việt (và các ngôn ngữ châu Á khác) đều dựa trên thông tin về sự xuất hiện cạnh nhau của các âm tiết (colocation).

Ba hướng tiếp cận chính để giải quyết bài toán tách từ:

- Tiếp cận dựa vào từ điển.
- Tiếp cận dựa vào thống kê.
- Kết hợp cả 2 hướng trên.

Trong hướng tiếp cận dựa vào từ điển, một chuỗi các âm tiết sẽ được xem là một từ ghép nếu chuỗi các âm tiết này có trong từ điển. Tiếp cận thống kê dựa trên sự xuất hiện cạnh nhau của các âm tiết, nếu sự xuất hiện cạnh nhau này xảy ra thường xuyên thì các âm tiết này rất có thể thuộc về một từ ghép nào đó (để thống kê sự xuất hiện cạnh nhau có thể nhờ đến sự trợ giúp của các công cụ tìm kiếm trên internet – search engine của Google, Yahoo, Bing). Tuy nhiên, dù tiếp cận theo hướng nào, nhập nhằng trong việc tách từ/ rút trích từ chắc chắn vẫn có thể xảy ra ngoài mong muốn dù người phát triển có xây dựng tốt đến cỡ nào vì ngôn ngữ của con người là biến đổi liên tục theo thời gian và không gian.

Hiện nay, có rất nhiều phương pháp dược sử dụng theo các hướng tiếp cận ở trên:

- Đối sánh thực thể dài nhất (Longest Matching).
- Đối sánh cực đại (Maximum Matching).
- Mô hình Markov ẩn (Hidden Markov Models- HMM).
- Học dựa trên sự cải biến (Transformation-based Learning – TBL).
- Chuyển đổi trạng thái trọng số hữu hạn (Weighted Finite State Transducer – WFST).
- Độ hỗn loạn cực đại (Maximum Entropy – ME).
- Máy học sử dụng vectơ hỗ trợ (Support Vector Machines).
- Trường xác xuất có điều kiện (CRFs).
- Đồ thị chuyển trạng thái (Transducing Graph).
- Tách dựa cú pháp.

- …

Trong đề tài này, em tiếp cận theo hướng sử dụng từ điển từ khóa cảm xúc và áp dụng phương pháp đối sánh thực thể dài nhất (Longest Matching). Đây là một phương pháp dễ cài đặt, tốc độ nhanh, độ chính xác chấp nhận được đối với bài toán tóm tắt văn bản, nhất là với đối tượng văn bản không tiêu chuẩn như những ý kiến phát biểu trên mạng xã hội. Phương pháp này dựa trên một từ điển tiếng Việt, gồm những từ và cụm từ

sau đây gọi chung là thực thể. Có hai phương pháp Đối sánh thực thể dài nhất là đối sánh từ trái qua và đối sánh từ phải qua.

Ví dụ: Rút trích thực thể của câu "Hôm nay nắng đẹp" bằng giải thuật từ trái qua. Giả sử trong từ điển của chúng ta có các thực thể: "hôm nay", "nắng", "nay", "hôm", "đẹp".

- Kiểm tra xem có thực thể "hôm nay nắng đẹp" không.
- Nếu có thì dừng lại và kết thúc quá trình.
- Nếu không có thì tách bớt âm tiết cuối ra, kiểm tra có thực thể "hôm nay nắng" trong kho ngữ liệu hay không.
- Nếu có thì dừng lại và kiểm tra phần còn lại của câu (cụ thể ở đây là "đẹp").
- Nếu không có thì tách bớt âm tiết cuối ra, kiểm tra có thực thể "hôm nay" trong kho ngữ liệu hay không.
- Nếu có thì dừng lại và kiểm tra phần còn lại của câu (cụ thể ở đây là "nắng đẹp").

Với giải thuật trên ta có thể nhận được tập thực thể (tương ứng với một từ điển cụ thể): "hôm nay", "nắng", "đẹp".

Thuật toán đối sánh từ phải qua ngược với thuật toán trên là lấy chuỗi dài nhất từ cuối câu. Khi cắt chuỗi hay âm tiết thì cắt phần bên trái nhất đi, giữ lại phần bên phải. Khi kết thúc thuật toán ta phải đảo ngược thứ tự các thực thể để có được trật tự các thực thể như trong câu ban đầu.

Đối với tiếng Việt, độ chính xác của thuật toán đối sánh từ phải qua cao hơn thuật toán đối sánh từ bên trái qua.

Ví dụ: Xét câu "Ban công tác hoàn thành nhiệm vụ".

Giả sử trong từ điển cảm xúc có các thực thể: "ban", "ban công", "công tác", "hoàn thành", "nhiệm vụ".

Kết quả phân tích của giải thuật đối sánh từ trái qua là: "ban công", "hoàn thành", "nhiệm vụ".

Kết quả phân tích của giải thuật đối sánh từ phải qua là: "nhiệm vụ", "hoàn thành", "công tác", "ban".

Thay đổi thứ tự các thực thể ta được “ban”, “công tác”, “hoàn thành”, “nhiệm vụ”.

Chúng ta nhận thấy kết quả từ giải thuật đối sánh từ phải tốt hơn rất nhiều giải thuật từ trái qua.

```
EXTRACT(Text, Dic)
    W=WordOfText(Text)
    start= CountOfWord(W)
    stop=0
    while isStop=false and start>=0
    begin while
        for index=stop to start
            Term += W[index]
        if Term ∈ Dic
        begin
            isTerm = true
            do TermList ← Term
            do TermValue ← ValueOf(Term)
            if start = CountOfWord(W) then isStop = true
            else
            begin
                stop  = start+1
                start = CountOfWord(W)
            end
        end
        if isTerm = false
        begin
            if start = stop
            begin
                stop++
                start = CountOfWord(W)
            end
            else start -= 1
        end
    end while
    return {TermList, TermValue}
```

Hình 3.1. Giải thuật tách thực thể từ trái qua.

Tuy nhiên, với những câu phức tạp như “Học sinh học sinh học” thì cả giải thuật đối sánh từ trái qua cũng như từ phải qua đều không thể có được kết quả chính xác.

```
EXTRACT(Text, Dic)
    W=WordOfText(Text)
    start=0
    stop=CountOfWord(W)
    while isStop=false and stop>=0
    begin while
        for index=start to stop
            Term += W[index]
        if Term ∈ Dic
        begin
            isTerm = true
            do TermList ← Term
            do TermValue ← ValueOf(Term)
            if start = 0 then isStop = true
            else
            begin
                stop  = start-1
                start=0
            end
        end
        if isTerm=false
        begin
            if start = stop
            begin
                stop--
                start = 0
            end
            else start += 1
        end
    end while
    return {TermList, TermValue}
```

Hình 3.2. Giải thuật tách thực thể từ phải qua.

3.3. Kỹ thuật phân loại ý kiến trong văn bản tiếng Việt

Giới hạn trong đề tài này, em sử dụng những bình luận của người dùng về một video nào đó như là đoạn văn bản cần xử lý. Đặt trường hợp người dùng bình luận với dấu câu như dấu chấm, dấu phẩy, dấu chấm than, dấu chấm hỏi… một cách hợp lý và đầy đủ văn phong chuẩn tiếng Việt, không dùng ngôn ngữ mạng, chữ viết tắt, tiếng anh lồng ghép, không spam. Vì từ điển cảm xúc cần có thời gian lâu dài xây dựng mà thời gian và phạm vi có hạn nên em sẽ tạm áp dụng cho những bình luận nằm trong phạm vi nhận xét về tổng thể bộ phim hoặc về các phương diện chính như diễn viên, nhạc phim, tình tiết, kỹ sảo…

- Đầu tiên, mỗi câu trong ý kiến sẽ được phân loại cảm xúc bằng phương pháp Naïve Bayes.
- Sau đó, mô hình hóa tập đặc trưng cảm xúc của mỗi câu thành các vector.
- Tiếp theo, chuẩn hóa các vector về chiều, và tổng hợp thành vector đặc trưng cho mỗi lớp cảm xúc bằng cách tính tổng các vector trong đó.
- Cuối cùng là xây dựng vector đặc trưng cảm xúc cho cả văn bản.

Quá trình sẽ chuẩn hóa 3 vector:

- Vector tổng (G): là vector chứa tất cả các đặc trưng cảm xúc của ý kiến. Các phần tử cảm xúc của G có thứ tự như trong văn bản gốc.
- Vector lớp tích cực P (positive): là vector tập hợp tất cả các đặc trưng cảm xúc có thứ tự như trong văn bản gốc, trong đó các phần tử của các vector lớp tiêu cực N (negative) suy biến bằng 0.
- Vector lớp tiêu cực N (negative): là vector tập hợp tất cả các đặc trưng cảm xúc có thứ tự như trong văn bản gốc, trong đó các phần tử của các vector lớp positive suy biến bằng 0.

Để phân cực cảm xúc cho văn bản, em sẽ tính độ tương đồng của G, P và N theo từng cặp: Sin(G, P) và Sin(G, N) theo công thức hình dưới :

$$Sim(X,Y)=Cosin(X,Y)=\frac{X.Y}{|X||Y|}=\frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

Hình 3.3. Công thức tính độ tương đồng của các vector G, P, N

Trong đó X, Y là 2 vector với:

X = {x1, x2, …, xn}, Y = {y1, y2, …, yn}.

So sánh hai giá trị độ tương đồng của các cặp vector trên để xác định G tương đồng với P

hoặc N. Vector G sẽ tương đồng với vector thành phần có giá trị độ tương đồng lớn hơn. Nếu hai giá trị là bằng nhau thì xếp ý kiến vào lớp trung hòa (neutral).

Ví dụ :

Ta xét câu bình luận : "Diễn viên thì đẹp, nội dung khá hay nhưng phim xem quá dở."

Xử dụng giải thuật rút trích từ phải qua trái ta được các từ sau :

- "đẹp" là từ hạt giống có trọng số là 1.
- "khá hay" là cụm từ có trọng số là 0,5.
- "quá dở" là cụm từ có trọng số là -2.

Ta có vector G = {1; 0,5; -2} ; vector N = {0; 0; -2} ; vector P = {1; 0,5; 0}

Áp dụng công thức tính cosin và sin ở trên ta có :

Cosin(N,G) = $\frac{|0+0+4|}{\sqrt{5,25}.\sqrt{4}} = \frac{4}{\sqrt{21}}$; Cosin(P,G) = $\frac{|1+0,25+0|}{\sqrt{1,25}.\sqrt{5,25}} = \frac{1,25}{\sqrt{6,5625}}$

So sánh ta thấy : Cosin(N,G) > Cosin(P,G)

Vậy suy ra vector G gần vector N hơn nên câu bình luận trên mang ý nghĩa tiêu cực. Ko tích cực.

Tương tự nếu lấy câu bình luận : " Phim này hay ghê, biên kịch quá đỉnh."

Ta có các từ được rút trích ra:

- "hay ghê" có trọng số là 0,5
- "quá đỉnh" có trọng số là 2

Vector G = {0,5 ; 2} ; vector N = {0; 0} ; vector P = {0,5 ; 2}

Áp dụng công thức như trên ta có Cosin(N, G) < Cosin(P,G) nên vetor G gần vector P hơn. Câu bình luận trên mang ý nghĩa tích cực.

3.4. Các yêu cầu chức năng

Yêu cầu lưu trữ :

- Lưu bình luận người xem phim tương ứng với phim.
- Lưu thông tin người xem phim(password, username) khi đăng ký tài khoản.

Yêu cầu nghiệp vụ :

- Hiện ra các video phim được admin tải lên.
- Trình phát video.
- Cung cấp mục bình luận cho người xem phim.
- Mục Top Phim hot (5 phim hot nhất được xem nhiều nhất).
- Danh sách phim theo thể loại (Khoa học – viễn tưởng; Phiêu lưu – Hành động; Kinh dị - Ma; Tâm lý – Tình cảm) và theo quốc gia (Mỹ, Trung Quốc, Hàn Quốc, Thái Lan).
- Tìm kiếm phim theo tên phim, tên đạo diễn, tên diễn viên.
- Cho phép người xem phim đánh giá chất lượng phim theo thang điểm 10 sao.
- …

Yêu cầu báo biểu :

- Hiện ra thông tin lượt xem, điểm đánh giá (5/10 sao) của 1 bộ phim.
- Từ việc phân loại cảm xúc đánh giá bình luận khen hay chê mà hiện ra bảng xếp hạng phim dựa vào số bình luận khen và chê của 1 bộ phim giúp người xem phim có thêm tiêu chí để đánh giá chất lượng phim hay hoặc dở đáng xem hay không.

3.5. Các yêu cầu phi chức năng

- Giao diện trang web phải dễ sử dụng, trực quan, thân thiện với mọi người dùng.
- Tốc độ xử lý của hệ thống phải nhanh chóng và chính xác.
- Khi xem phim không bị giật, video chạy mượt độ nét chuẩn.
- Phim luôn cập nhật mới nhất và đầy đủ nhất, phim có chọn lọc.

CHƯƠNG 4: XÂY DỰNG HỆ THỐNG

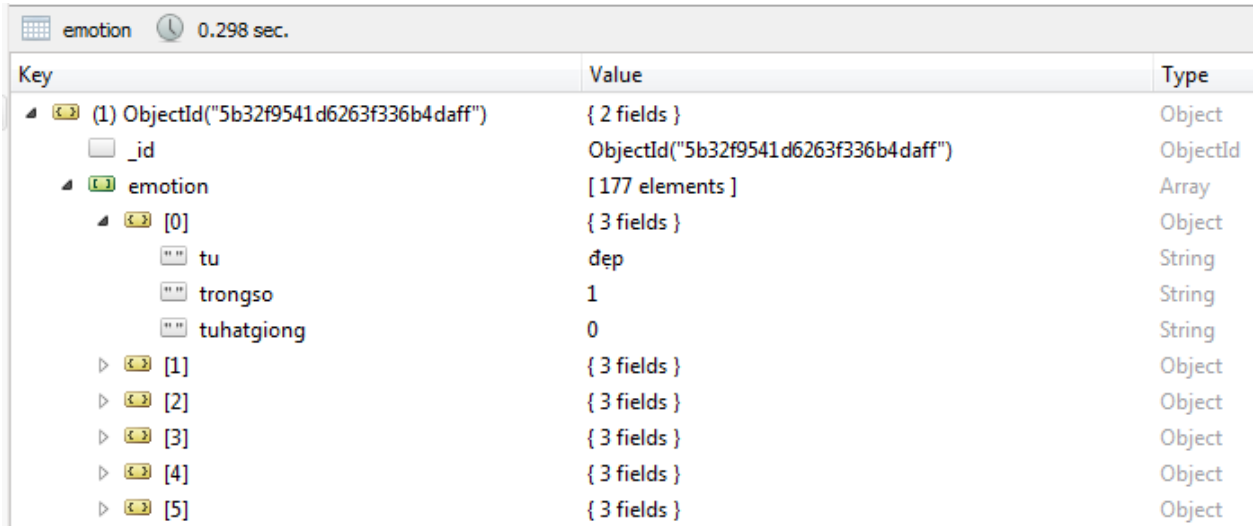
4.1. Phân tích – Thiết kế cơ sở dữ liệu cho hệ thống

4.1.1. Xác định các thực thể - collection trong đề tài

- emotion(_id, tu, trongso, tuhatgiong)
  - tu: từ khóa cảm xúc
  - trongso: trọng số cảm xúc của từ khóa đó
  - tuhatgiong: true or false? Xác định xem từ khóa đó có phải từ hạt giống hay không?
- movie(_id, phim, hinhanh, poster, linkphim, binhluan, khen, che)
  - phim(theloai, quocgia, tenphim, daodien, dienvien, thoiluong, luotxem, danhgia) : chứa thông tin bộ phim : tên phim, thể loại , quốc gia, đạo diễn, diễn viên, thời lượng, lượt xem, đánh giá bình luận.
  - khen, che: số bình luận khen chê của 1 bộ phim.
- user(_id, userlist)
  - userlist(username, password, role) : tên user, mật khẩu user đó và quyền của user ví dụ quyền admin hay người xem phim.

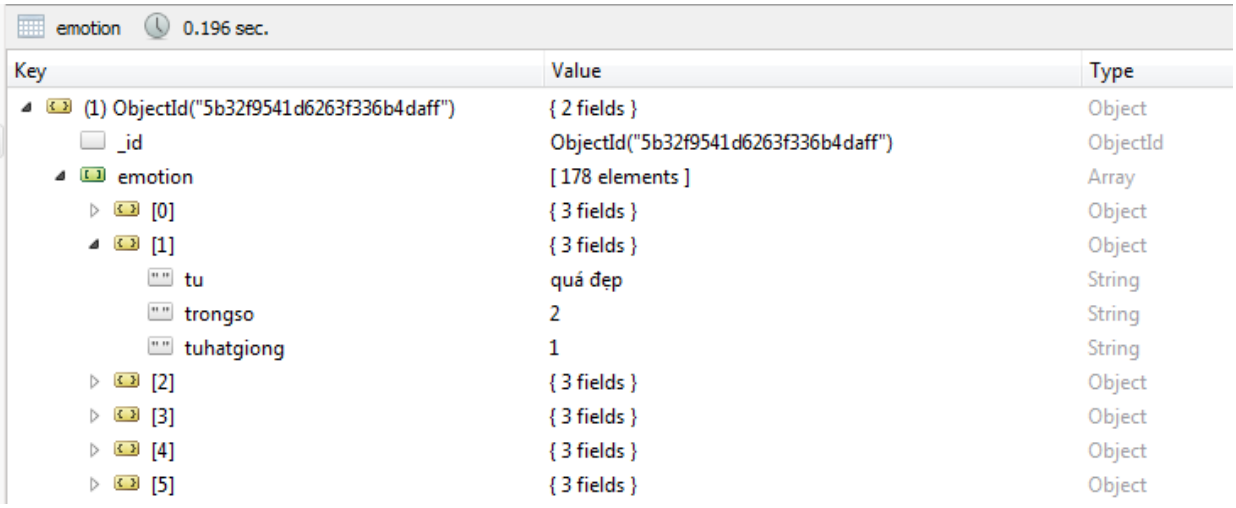
4.1.2. Từ điển dữ liệu

- **Cấu trúc 1 document trong Collection "emotion"**
  - Là từ hạt giống : Value = 0



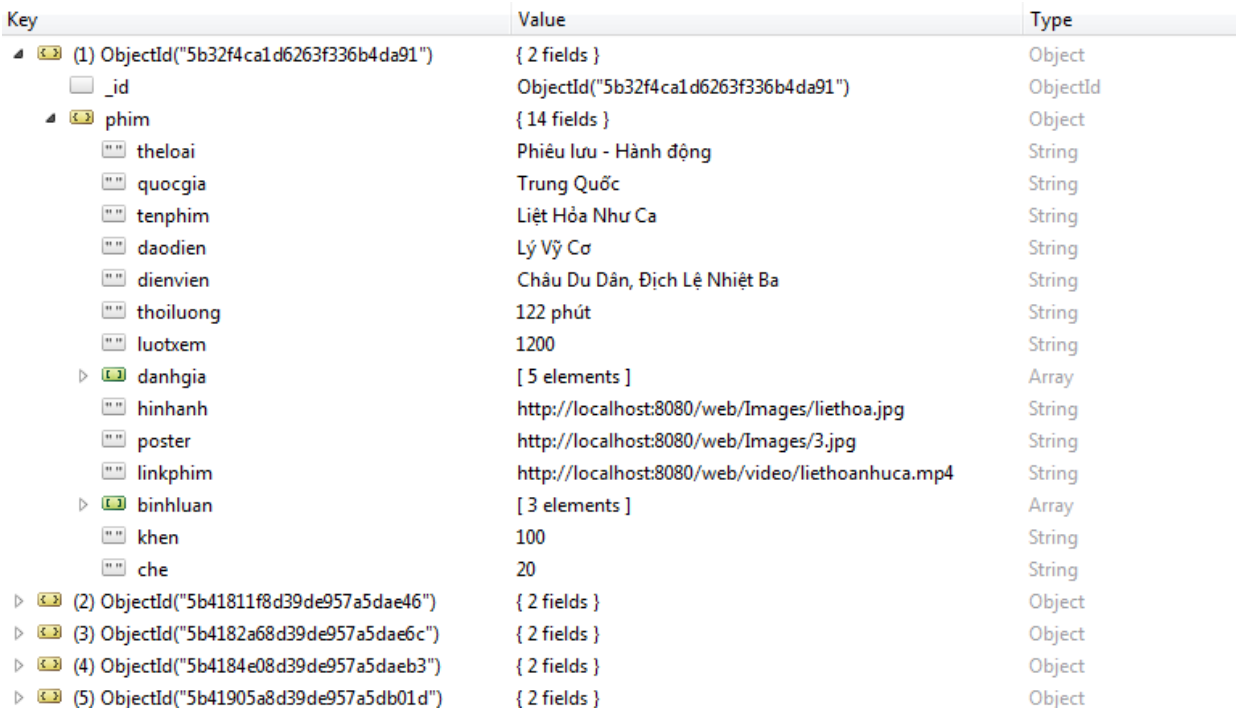
Hình 4.1. Một document chứa thông tin của một từ hạt giống trong từ điển bao gồm trọng số

- Không phải là từ hạt giống : Value = 1.

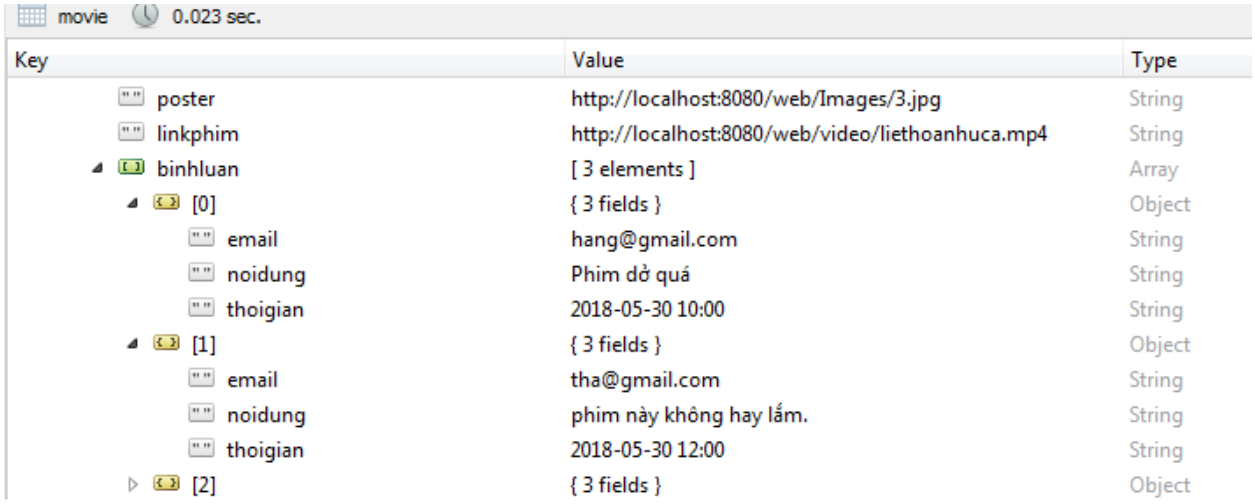


Hình 4.2. Một document chứa từ không phải hạt giống trong từ điển bao gồm trọng số.

❖ Cấu trúc một document trong collection “movie”

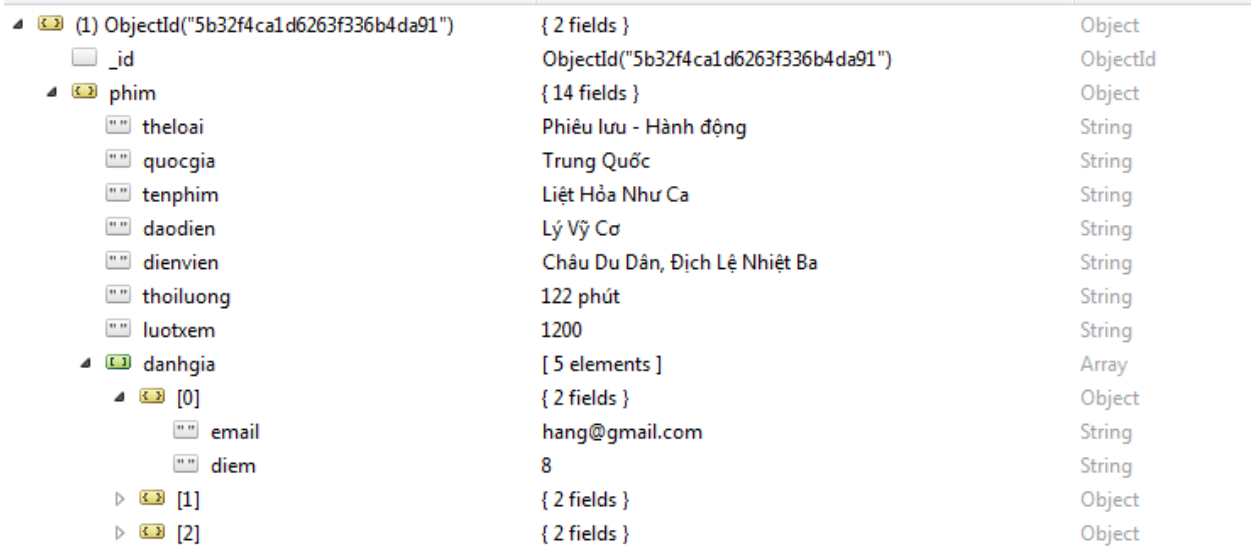


Hình 4.3. Một document chứa đối tượng phim.


movie 0.023 sec.

Key	Value	Type
poster	http://localhost:8080/web/Images/3.jpg	String
linkphim	http://localhost:8080/web/video/liethoanhuca.mp4	String
binhluan	[3 elements]	Array
[0]	{ 3 fields }	Object
email	hang@gmail.com	String
noidung	Phim dở quá	String
thoigian	2018-05-30 10:00	String
[1]	{ 3 fields }	Object
email	tha@gmail.com	String
noidung	phim này không hay lắm.	String
thoigian	2018-05-30 12:00	String
[2]	{ 3 fields }	Object

Hình 4.4. Một document chứa bình luận của người xem phim.



(1) ObjectId("5b32f4ca1d6263f336b4da91")	{ 2 fields }	Object
_id	ObjectId("5b32f4ca1d6263f336b4da91")	ObjectId
phim	{ 14 fields }	Object
theloai	Phiêu lưu - Hành động	String
quocgia	Trung Quốc	String
tenphim	Liệt Hỏa Như Ca	String
daodien	Lý Vỹ Cơ	String
dienvien	Châu Du Dân, Địch Lệ Nhiệt Ba	String
thoiluong	122 phút	String
luotxem	1200	String
danhgia	[5 elements]	Array
[0]	{ 2 fields }	Object
email	hang@gmail.com	String
diem	8	String
[1]	{ 2 fields }	Object
[2]	{ 2 fields }	Object

Hình 4.5. Một document chứa đánh giá của người xem phim.

❖ **Cấu trúc một document trong collection "user"**

user 0.007 sec.

Key	Value	Type
(1) ObjectId("5b32fbd21d6263f336b4db4e")	{ 2 fields }	Object
_id	ObjectId("5b32fbd21d6263f336b4db4e")	ObjectId
userlist	[2 elements]	Array
[0]	{ 3 fields }	Object
[1]	{ 3 fields }	Object

Hình 4.6. Một document chứa danh sách các user

user 0.007 sec.

Key	Value	Type
(1) ObjectId("5b32fbd21d6263f336b4db4e")	{ 2 fields }	Object
_id	ObjectId("5b32fbd21d6263f336b4db4e")	ObjectId
userlist	[2 elements]	Array
[0]	{ 3 fields }	Object
username	harry	String
password	55555	String
role	admin	String
[1]	{ 3 fields }	Object

Hình 4.7. Một document chứa user với quyền là admin.

user 0.007 sec.

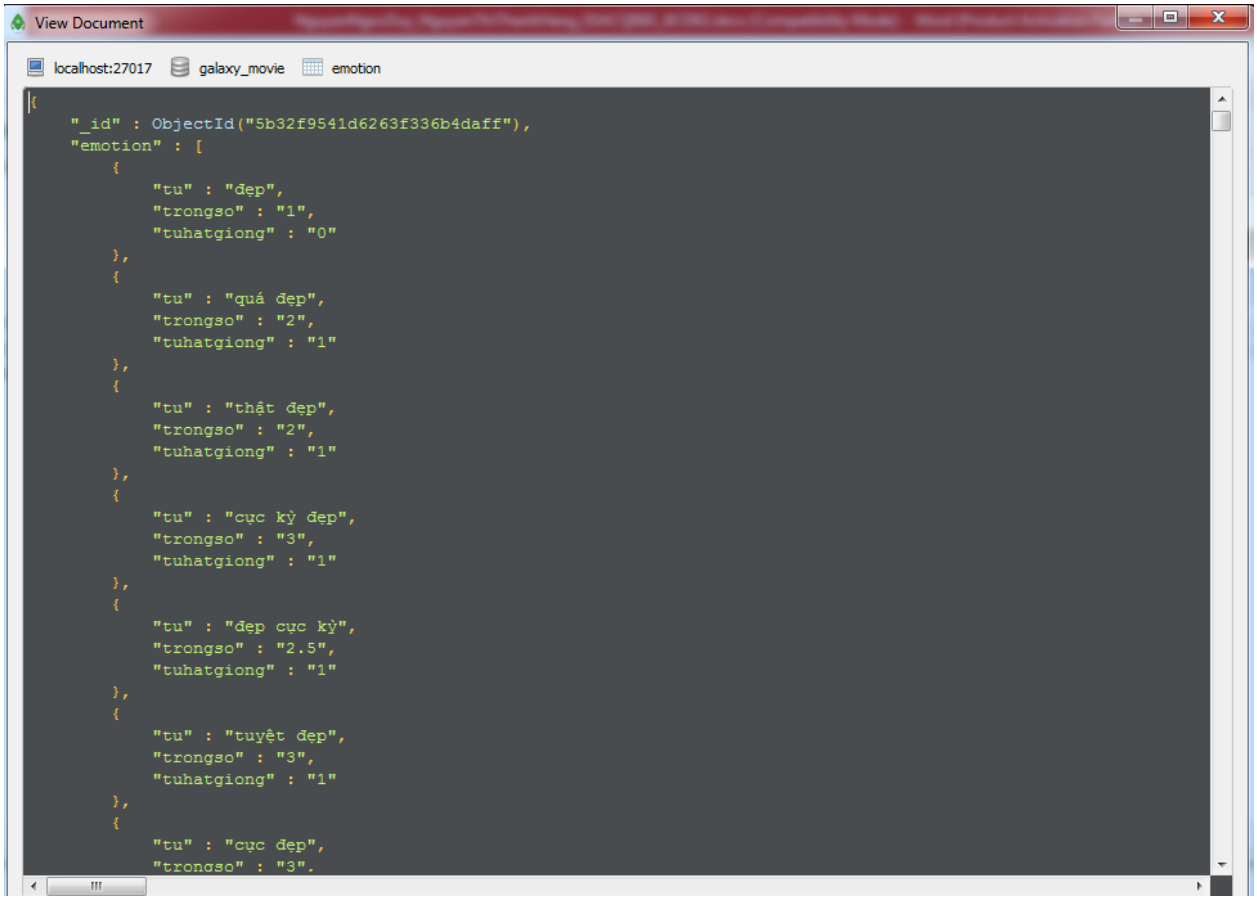
Key	Value	Type
(1) ObjectId("5b32fbd21d6263f336b4db4e")	{ 2 fields }	Object
_id	ObjectId("5b32fbd21d6263f336b4db4e")	ObjectId
userlist	[4 elements]	Array
[0]	{ 3 fields }	Object
[1]	{ 3 fields }	Object
username	Hermione	String
password	korea	String
role	normal	String
[2]	{ 3 fields }	Object
[3]	{ 3 fields }	Object

Hình 4.8. Một document chứa user với quyền là nomal (không phải admin).

4.2. Xây dựng một từ điển cảm xúc theo chủ đề phim đơn giản

Trong phạm vi đề tài này, vì thời gian có hạn nên việc xây dựng một hệ thống từ điển quy mô lớn với độ chính xác cao là khó có thể cho nên em đã cố gắng xây dựng một từ điển cảm xúc đơn giản theo chủ đề thường được mọi người quan tâm bình luận nhất là về bộ phim cũng như diễn viên và gán vào trọng số cảm xúc theo ý kiến chủ quan của bản thân. Trong từ điển sẽ có những từ biểu thị cảm xúc mà người xem phim thường bình luận, với tiêu chí là các bình luận phải đúng chuẩn mực tiếng Việt, có dấu chấm câu đầy đủ, không viết tắt, không dùng tiếng lóng và không bình luận bằng tiếng nước ngoài.

Tương ứng với mỗi từ cảm xúc sẽ được gán với một con số biểu thị trọng số cảm xúc với ý nghĩa tích cực (khen) là số dương, ý nghĩa tiêu cực (chê) là số âm, ý nghĩa trung lập (không khen cũng không chê) là bằng 0. Và từ đó sẽ được xét xem có phải là từ hạt giống hay không, vì từ hạt giống sẽ là từ luôn được gán với trọng số là 1 (đối với từ có ý nghĩa khen) và trọng số -1 (đối với từ có ý nghĩa chê), từ hạt giống sẽ là từ được làm dấu mốc để từ đó khi ta thêm vào những phó từ bổ nghĩa đằng sau hay đằng trước từ hạt giống để biểu thị rõ hơn mức độ cảm xúc thì ta sẽ căn cứ vào đó để điều chỉnh giá trị trọng số cảm xúc cho phù hợp với từng cụm từ.



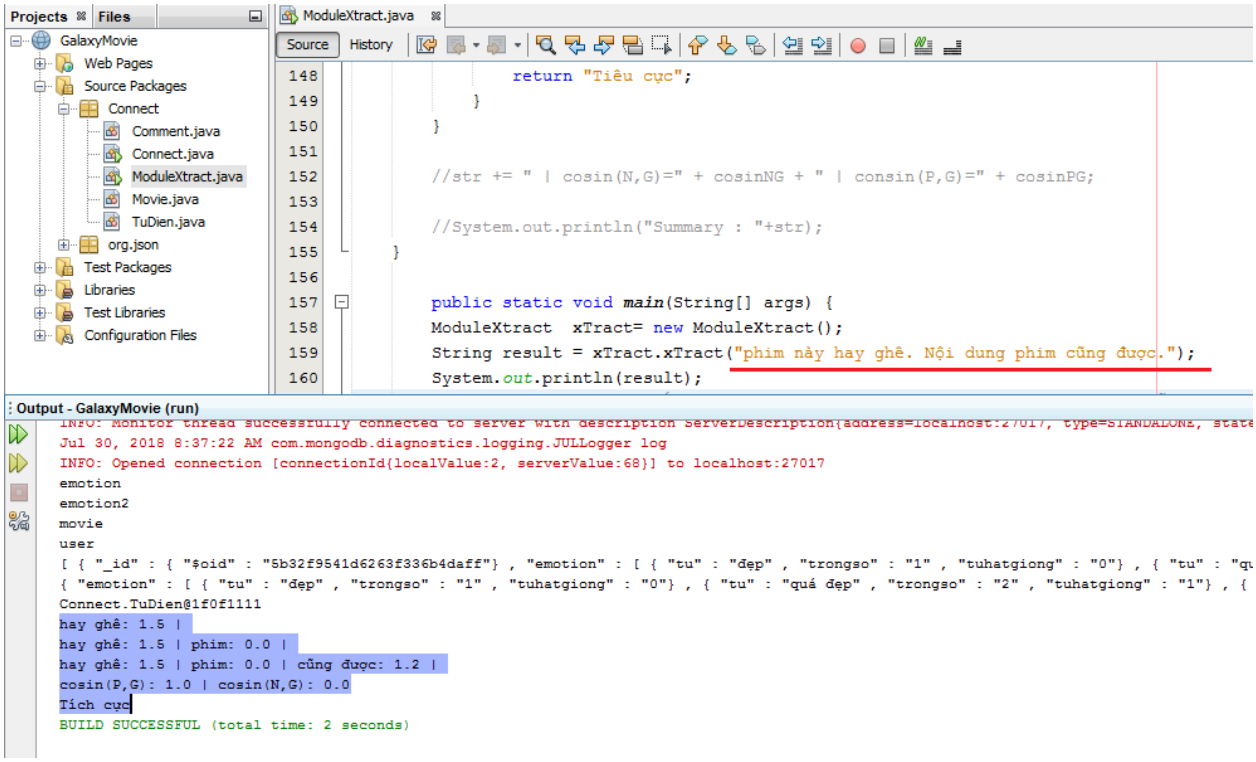
Hình 4.9. Từ điển cảm xúc theo chủ đề phim đơn giản.

4.3. Xây dựng module rút trích, phân loại cảm xúc

4.3.1. Module rút trích, phân loại cảm xúc văn bản tiếng Việt

- Module rút trích, phân loại 1 câu bình luận đơn giản. Giả sử ta rút trích và phân loại cảm xúc của 1 câu bình luận : "Phim này hay ghê. Nội dung phim cũng được." .

- Với các từ có trong từ điển cảm xúc : "hay ghê" có trọng số là 1.5 ; "phim" có trọng số là 0 (phim là thực thể không mang ý nghĩa cảm xúc, là 1 khía cạnh) ; "cũng được" có trọng số là 1.2
- Sau khi dùng thuật toán rút trích và phân loại ý kiến thì ta sẽ được kết quả là bình luận này mang ý nghĩa khen tức tích cực.



Hình 4.10. Rút trích, phân loại cảm xúc 1 câu bình luận.

4.3.2. Module rút trích, phân loại cảm xúc các bình luận của 1 bộ phim

- Sau khi đã rút trích, phân loại cảm xúc của 1 câu văn bản, ta sẽ áp dụng vào việc lấy toàn bộ bình luận của người xem phim trong 1 bộ phim ra để phân loại cảm xúc và phân các bình luận đó vào 2 mục là khen hay là chê. Đối với các bình luận mang tính trung lập tức không khen cũng không chê thì ta bỏ qua không lấy để đánh giá.

- Dưới đây là ví dụ về việc rút trích, phân loại tất cả các bình luận của bộ phim "Nắm đấm thép". Trước khi rút trích, phân loại thì số bình luận khen và chê của bộ phim này tương ứng là 6 và 4; sau khi rút trích, phân loại các bình luận mới thì số bình luận khen và chê sẽ tăng lên tương ứng là 7 và 5 (thêm 1 bình luận khen và 1 bình luận chê).

- Từ đó ta sẽ dựa vào số bình luận khen và chê để đánh giá và xếp hạng 1 bộ phim.

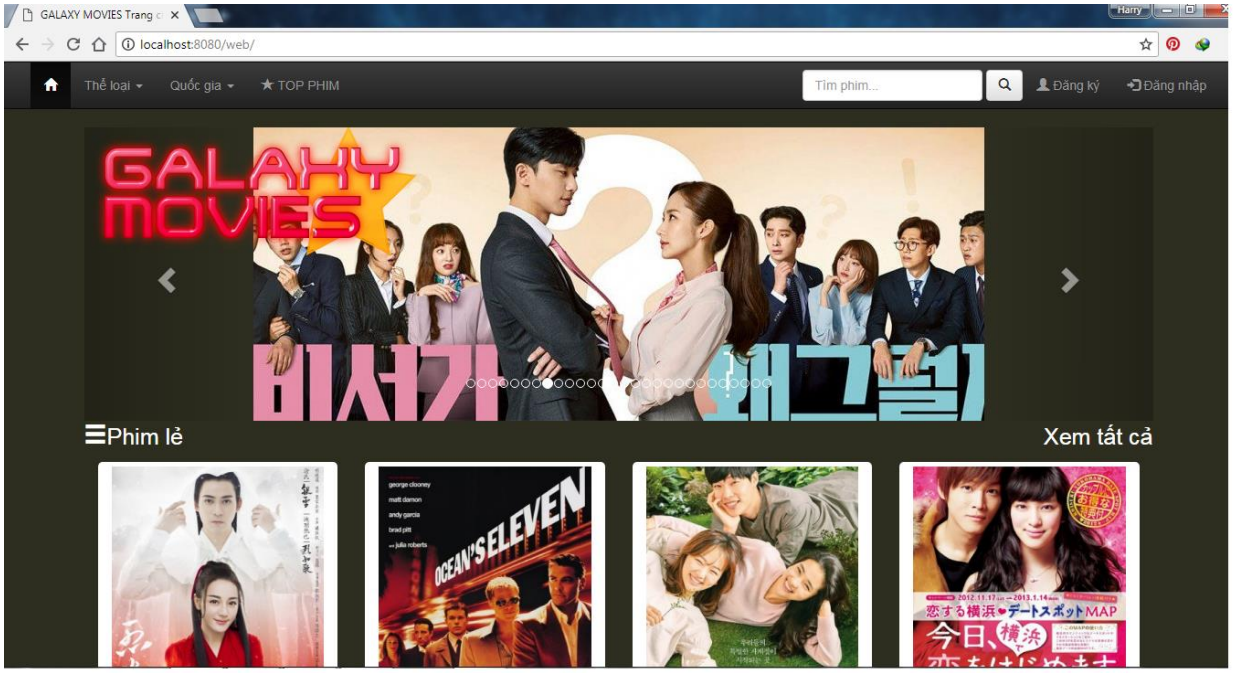
```
------------------------------
Nắm đấm thép
Trước
khen: 6
chê: 4
comment 1: Phim dở quá
dở quá: -1.5 |
cosin(P,G): 0.0 | cosin(N,G): 1.0
comment 2: phim này không hay lắm.
không hay lắm: 0.0 |
không hay lắm: 0.0 | phim: 0.0 |
cosin(P,G): 0.0 | cosin(N,G): 0.0
comment 3: diễn viên đẹp. Phim cũng được.
đẹp: 1.0 |
đẹp: 1.0 | cũng được: 1.0 |
cosin(P,G): 1.0 | cosin(N,G): 0.0
Sau
khen: 7
chê: 5
```

Hình 4.11. Rút trích, phân loại cảm xúc tất cả bình luận của 1 bộ phim.

4.4. Xây dựng trang web xem phim GALAXY MOVIES

4.4.1. Thiết kế giao diện

- Trang chủ trang web xem phim trực tuyến **Galaxy Movie:** cung cấp cho người dùng một trang web xem phim chuyên về phim chiếu rạp và phim lẻ hot nhất.



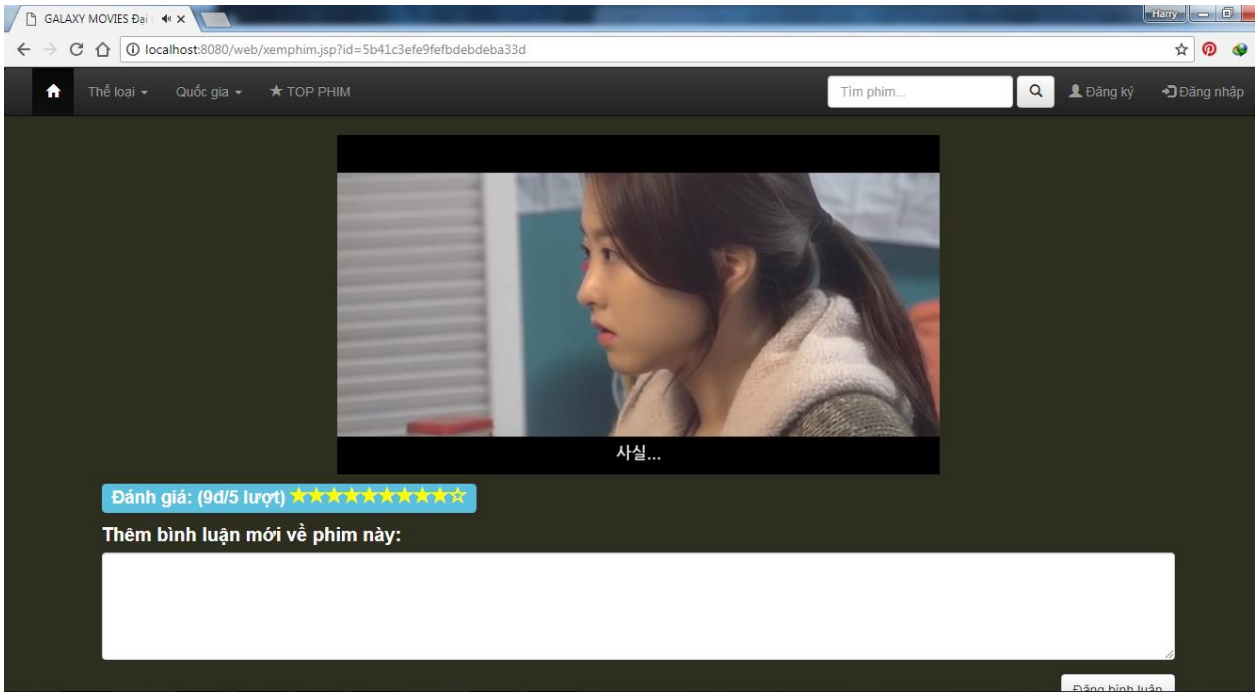
Hình 4.12. Trang chủ trang web xem phim **Galaxy Movie.**

- Trang thông tin phim: cung cấp các thông tin cơ bản như tên phim, đạo diễn, diễn viên, thời lượng, quốc gia, thể loại…

Hình 4.13. Trang thông tin phim của trang web.

4.4.2. Chức năng quản lý và chạy file video trực tuyến

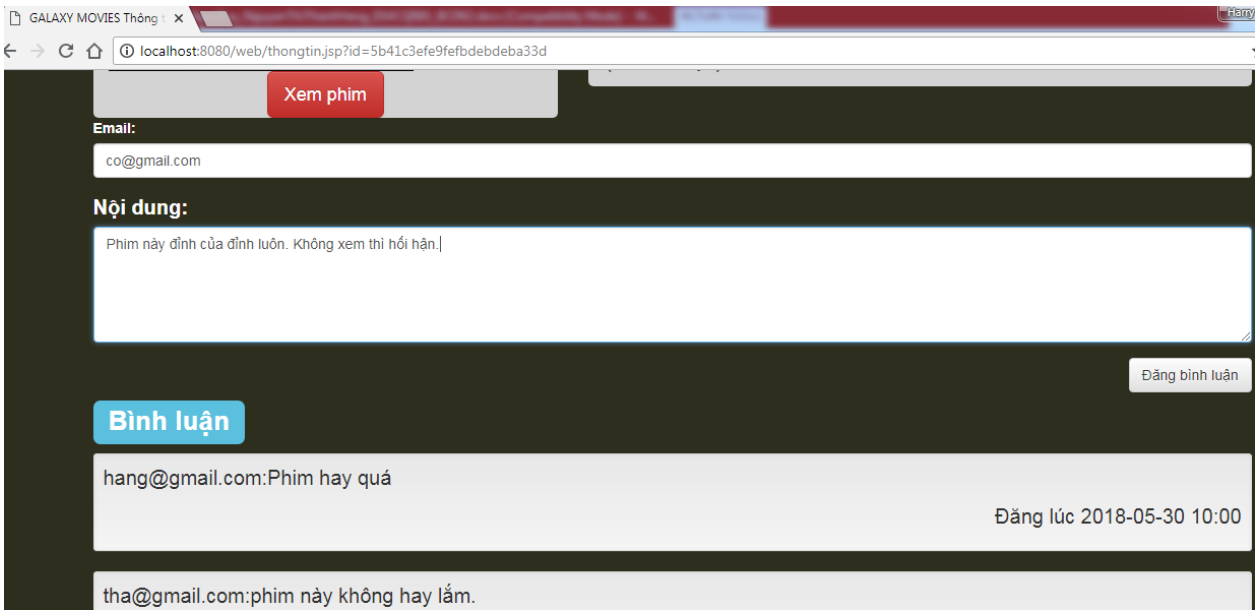
- Trang xem phim: cho phép chạy file video trực tuyến.



Hình 4.14. Trang xem phim cho phép chạy file video trực tuyến.

4.4.3. Chức năng ghi nhận và lưu trữ các ý kiến đánh giá của người xem

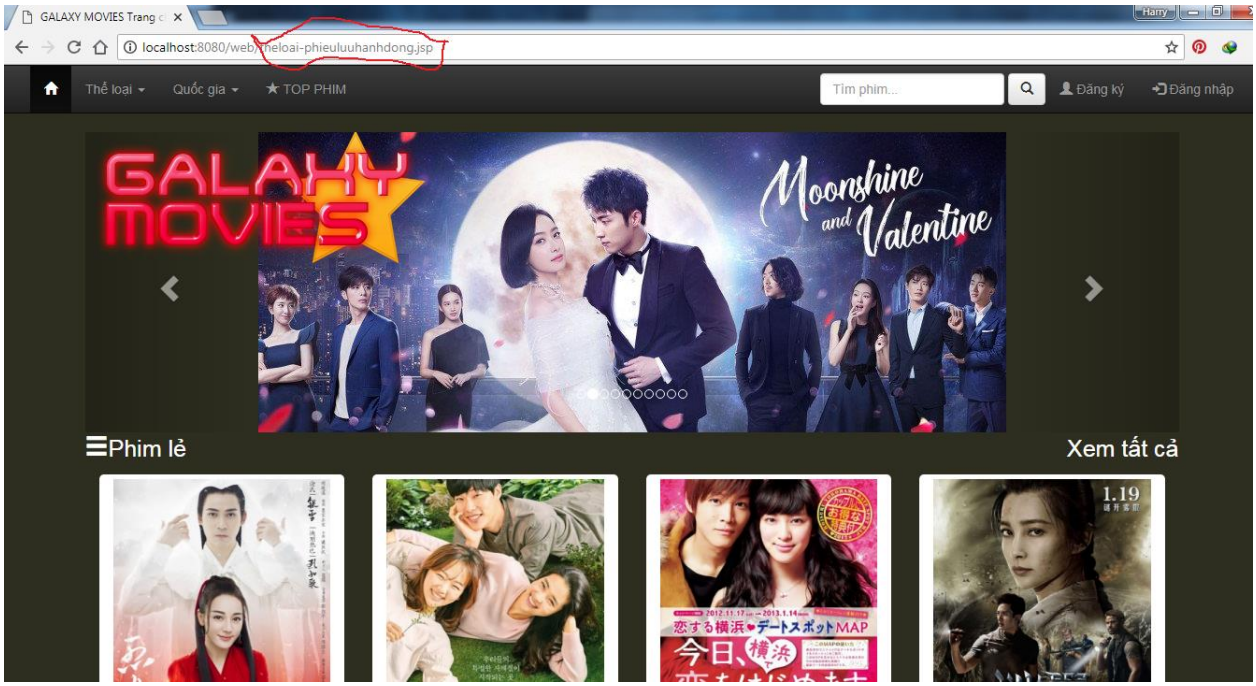
- Trang xem phim và trang thông tin phim cho phép người dùng đăng nhập và bình luận về phim, hiển thị bình luận phim và những bình luận này sẽ được lưu trữ vào cơ sở dữ liệu để tiến hành rút trích và phân loại cảm xúc.



Hình 4.15. Đăng nhập và ghi nhận bình luận người xem.

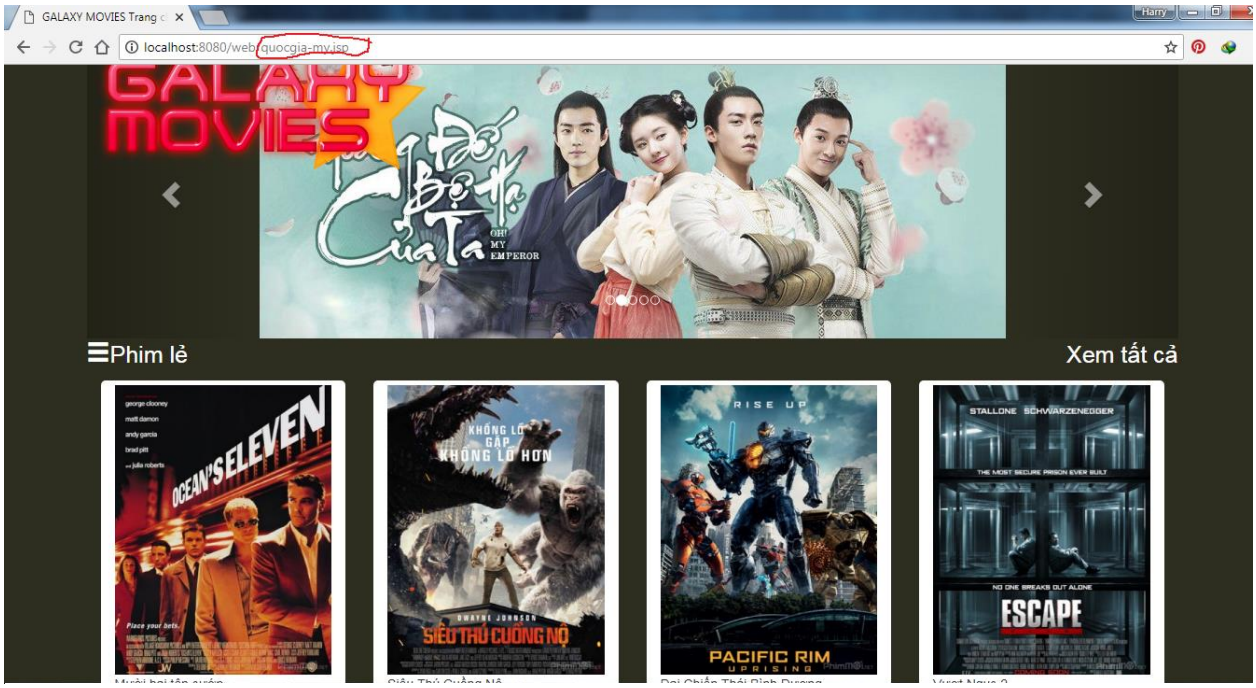
4.4.4. Phân loại và xếp hạng các bộ phim

- Phân loại các bộ phim theo từng thể loại phim: cho phép người xem phim chọn lọc phim theo 4 thể loại chính là Khoa học – Viễn tưởng; Phiêu lưu – Hành động; Tâm lý – Tình cảm; Kinh dị - Ma.



Hình 4.16. Phân loại theo thể loại phim.

- Phân loại phim theo quốc gia: cho phép người xem phim chọn lọc phim theo quốc gia như Mỹ, Trung Quốc, Hàn Quốc và Thái Lan.



Hình 4.17. Phân loại phim theo quốc gia.

KẾT LUẬN

Trải qua hơn 4 năm học tại trường, em đã học được rất nhiều kiến thức hữu ích từ các thầy cô và vận dụng các kiến thức đã học vào đề tài thực tập tốt nghiệp. Chẳng hạn như việc cung cấp các tính năng đáp ứng nhu cầu giải trí và chia sẻ video. Trích xuất ý kiến người dùng để dùng cho nhiều mục đích hữu ích như xếp hạng video, xây dựng các chức năng của hệ thống cũng như việc lưu trữ dữ liệu một cách chuyên nghiệp hơn…

❖ **Nội dung lý thuyết được củng cố :**

- Phân tích, thiết kế hệ thống thông tin phi cấu trúc, hướng đối tượng.
- Khai phá nguồn dữ liệu văn bản tiếng Việt.
- Hoàn thiện cách thức thiết kế cơ sở dữ liệu.
- Làm việc với các công nghệ và công cụ mới hiện nay như Web Service, NoSQL với MongoDB, các ngôn ngữ như Java, Node JS…

❖ **Các kỹ năng đã học hỏi được :**

- Kỹ năng thu thập thông tin trong giai đoạn tạo dựng từ điển cảm xúc.
- Kỹ năng viết báo cáo.
- Áp dụng được mô hình 1 lớp (View) trong quá trình xây dựng trang web.

❖ **Những kinh nghiệm thực tiễn học hỏi được :**

- Tập trung nhiều thời gian hơn cho việc tìm hiểu thu thập thông tin nghiệp vụ. Học cách sắp xếp thời gian hợp lý giữa thực tập và làm đồ án.
- Chọn đúng hướng phân tích, hiểu được những gì mình phải làm.
- Thường xuyên giữ liên lạc với giáo viên hướng dẫn, cán bộ hướng dẫn để báo cáo tiến độ thực hiện và điều chỉnh kịp thời những sai sót.

❖ **Hướng phát triển :**

- Tiếp tục thực hiện và phát triển thêm các chức năng mở rộng còn thiếu.
- Tối ưu hóa từ điển cảm xúc bằng cách xây dựng theo các khía cạnh.
- Tối ưu hóa chức năng đánh giá người xem phim…

❖ **Các phần chưa làm được :**

- Một số giao diện chưa thân thiện với người dùng.
- Chưa làm kịp chức năng quản lý phim, xếp hạng phim dựa vào đánh giá người xem phim.

DANH MỤC TÀI LIỆU THAM KHẢO

Tiếng Việt :

1. Nguyễn Tài Cẩn, "Ngữ pháp tiếng Việt", Đại học Quốc gia Hà Nội, Hà Nội, 1996.
2. Nguyễn Ngọc Duy, Phan Thị Tươi, "Tóm tắt văn bản trên cơ sở phân loại ý kiến độc giả của báo mạng tiếng Việt", Tạp chí Phát triển Khoa học và Công nghệ, Đại học Quốc gia Thành phố Hồ Chí Minh, K5, 19, pp. 53-61, 2016.

Tiếng Anh :

1. Erik Cambria, Daniel Olsher, Dheeraj Rajagopal, "SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis", Proceedings of the 28th AAAI Conference on Artifical Intelligence, pp. 1515-1521, 2014.
2. Zhen Zhang and Lirong Qiu, "A sentiment Calculation Method Based on Tibetan Semantic Relations", International Journal of Database Theory and Application Vol 9, No 9, pp. 149-156, 2016.

Danh mục các Website tham khảo :

1. https://docs.mongodb.com/tutorials/
2. https://o7planning.org/
3. https://code.visualstudio.com/docs/nodejs/nodejs-tutorial
4. http://www.phimmoi.net/
5. https://fptplay.vn/