**NGUYỄN THÀNH NAM MSSV: N14DCCN079 Tên Đề tài: XÂY DỰNG WEBSITE KINH DOANH SÁCH CÓ ỨNG DỤNG KHAI PHÁ DỮ LIỆU Lớp: D14CQIS01-N**

**Gáy**

**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 website kinh doanh sách có ứng dụng khai phá dữ liệu để gợi ý sách”**

**Người hướng dẫn : THS NGUYỄN NGỌC DUY**

**Sinh viên thực hiện : NGUYỄN THÀNH NAM**

**Mã số sinh viên : N14DCCN079**

**Lớp : D14CQIS01 - N**

**Khoá** **: 2014**

**Hệ** **: ĐẠI HỌC CHÍNH QUY**

**TP.HCM, tháng 08/2018**

TP

HCM

**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 website kinh doanh sách có ứng dụng khai phá dữ liệu để gợi ý sách”**

**Người hướng dẫn : THS NGUYỄN NGỌC DUY**

**Sinh viên thực hiện : NGUYỄN THÀNH NAM**

**Mã số sinh viên : N14DCCN079**

**Lớp : D14CQIS01 - N**

**Khoá** **: 2014**

**Hệ** **: ĐẠI HỌC CHÍNH QUY**

**TP.HCM, tháng 08/2018**

**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

**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 ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC**

1. **Tên đề tài: Xây dựng website kinh doanh sách có ứng dụng khai phá dữ liệu để gợi ý sách**
2. **Sinh viên: Nguyễn Thành Nam Lớp: D14CQIS01-N**
3. **Giáo viên hướng dẫn: Nguyễn Ngọc Duy**
4. **Nơi công tác: Học viện Công nghệ Bưu chính Viễn thông cơ sở 2, TP. Hồ Chí Minh.**

**NỘI DUNG NHẬN XÉT**

1. Đánh giá chung:

………………………………………………………………………………………………

………………………………………………………………………………………………

1. Đánh giá chi tiết:

………………………………………………………………………………………………

………………………………………………………………………………………………

………………………………………………………………………………………………

………………………………………………………………………………………………

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

………………………………………………………………………………………………

………………………………………………………………………………………………

1. Kết luận:

………………………………………………………………………………………………

………………………………………………………………………………………………

1. Điểm hướng dẫn :

**GIẢNG VIÊN HƯỚNG DẪN**

(Ký, ghi rõ họ tên)

MỤC LỤC

[**LỜI CẢM ƠN 1**](#_Toc521440259)

[**MỤC LỤC 1**](#_Toc521440260)

[**DANH MỤC BẢNG 3**](#_Toc521440261)

[**DANH MỤC HÌNH 4**](#_Toc521440262)

[**LỜI MỞ ĐẦU 5**](#_Toc521440263)

[**CHƯƠNG 1: GIỚI TIỆU ĐỀ TÀI 6**](#_Toc521440264)

[**1. Tổng quan 6**](#_Toc521440265)

[**2. Tìm hiểu khai phá ý kiến người mua sách 6**](#_Toc521440266)

[**3. Công nghệ ứng dụng 6**](#_Toc521440267)

[**4. Công cụ sử dụng 6**](#_Toc521440268)

[**5. Mục tiêu đề tài 6**](#_Toc521440269)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 7**](#_Toc521440270)

[**1. Giới thiệu các công nghệ 7**](#_Toc521440271)

[**1.1. ASP.NET MVC 7**](#_Toc521440272)

[**1.2. SQL Server 7**](#_Toc521440273)

[**1.3. Entity Framework 8**](#_Toc521440274)

[**2. Giới thiệu công cụ 8**](#_Toc521440275)

[**2.1. Visual Studio 8**](#_Toc521440276)

[**2.2. SQL Server Management Studio: 11**](#_Toc521440277)

[**3. Các kỹ thuật trong đề tài 12**](#_Toc521440278)

[**3.1. Kỹ thuật xây dựng một từ điển cảm xúc 12**](#_Toc521440279)

[**3.2. Giải thuật rút trích đặc trưng trong văn bản tiếng Việt 12**](#_Toc521440280)

[**3.3. Kỹ thuật phân loại ý kiến trong văn bản tiếng Việt 16**](#_Toc521440281)

[**3.4. Thuật toán KMean 18**](#_Toc521440282)

[**3.4.1. Khái niệm 18**](#_Toc521440283)

[**3.4.2. Ưu và nhược điểm 18**](#_Toc521440284)

[**CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH 19**](#_Toc521440285)

[**1. Cơ sở dữ liệu: 19**](#_Toc521440286)

[**1.1. Từ điển cảm xúc 19**](#_Toc521440287)

[**1.2. Cơ sở dữ liệu cho website sách 20**](#_Toc521440288)

[**2. Quá trình khai phá 22**](#_Toc521440289)

[**2.1. Tri thức ban đầu 22**](#_Toc521440290)

[**2.2. Chuẩn hóa dữ liệu 22**](#_Toc521440291)

[**2.3. Áp dụng KMean 22**](#_Toc521440292)

[**CHƯƠNG 4: KẾT QUẢ THỰC HIỆN 23**](#_Toc521440293)

[**1. Trang web 23**](#_Toc521440294)

[**1.1. Trang chủ 23**](#_Toc521440295)

[**1.2. Danh sách sản phẩm 25**](#_Toc521440296)

[**1.3. Trang chi tiết sản phẩm 26**](#_Toc521440297)

[**1.4. Trang quản lý 29**](#_Toc521440298)

[**2. Xây dựng module rút trích, phân trích và phân loại cảm xúc 31**](#_Toc521440299)

[**3. Chuẩn hỏa dữ liệu và học máy để đánh giá. 35**](#_Toc521440300)

[**3.1. Chuẩn hóa 35**](#_Toc521440301)

[**3.2. Học máy đánh giá ý kiến 36**](#_Toc521440302)

[**- Sad 36**](#_Toc521440303)

[**+ Sada 36**](#_Toc521440304)

[** dsadas 36**](#_Toc521440305)

[**CHƯƠNG 5: TỔNG KẾT 37**](#_Toc521440306)

[**DANH MỤC TÀI LIỆU THAM KHẢO 38**](#_Toc521440307)

# DANH MỤC BẢNG

[**Bảng đặc trưng ngữ nghĩa 20**](#_Toc521438044)

[**Bảng đặc trưng cảm xúc 20**](#_Toc521438045)

[**Bảng trọng số 21**](#_Toc521438046)

[**Bảng sách 22**](#_Toc521438047)

[**Bảng tác giả: 22**](#_Toc521438048)

[**Bảng thể loại 22**](#_Toc521438049)

[**Bảng nhà xuất bản 22**](#_Toc521438050)

[**Bảng ý kiến 22**](#_Toc521438051)

# DANH MỤC HÌNH

[**Hình 2.2.1: Hình ảnh minh họa công cụ 12**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438098)

[**Hình 2.3.1. Thuật toán đối sánh văn bản từ phải sang trái 16**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438099)

[**Hình 2.3.2. Công thức tính độ tương đồng của các vector G, P, N 17**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438099)

[**Hình 3.1.1 Từ điển cảm xúc diagram 20**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438100)

[**Hỉnh 3.1.2: Sách Database diagram 21**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438102)

[**Hình 4.1.1.1: Trang chủ với gợi ý là những sách mới thêm 24**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438103)

[**Hình 4.1.1.2: Trang chủ với danh sách gợi ý là sách đánh giá cao 25**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438104)

[**Hình 4.1.2: Trang danh sách sản phẩm theo sách mới 26**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438105)

[**Hình 4.1.3.1: Chi tiết sản phẩm 27**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438106)

[**Hình 4.1.3.2: Chức năng bình luận 28**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438107)

[**Hình 4.1.3.3: Xem ý kiến tích cực 29**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438108)

[**Hình 4.1.3.4: Xem ý kiến tiêu cực 29**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438109)

[**Hình 4.1.3.5: Xem ý kiến trung lập 29**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438110)

[**Hình 4.1.4.1: Trang quản lý sách 30**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438111)

[**Hình 4.1.4.2: Trang thêm sách 31**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438112)

[**Hình 4.1.4.3: Trang sửa thông tin sách 31**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438113)

[**Hình 4.2.1: Ví dụ rút trích 1 32**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438114)

[**Hình 4.2.2: Ví dụ rút trích 2 33**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438115)

[**Hình 4.2.3: Ví dụ rút trích 3 34**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438116)

[**Hình 4.2.4: Ví dụ rút trích 4 35**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438117)

[**Hình 4.3.2.1: File kho dữ liệu ý kiến 37**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438118)

[**Hình 4.3.2.2: Kết quả gom cụm với thuật toán KMean sử dụng Weka 38**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438118)

[**Hình 4.3.2.3: Rút trích dữ liệu và đánh giá văn bản cả học máy và tính toán vector 39**](file:///E:\NguyenNgocDuy_NguyenThanhNam_D14CQIS01_N14DCCN079_BCDK.docx#_Toc521438118)

**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 đến nhà sách xa xôi giữa thời tiết nắng nóng để chọn sách và mua sách, nhưng đôi khi ra đến nơi bạn lại chẳng biêt là cuốn nào hay cuốn nào dỡ, bạn chẳng thể nào đứng đó cả ngày mà đọc từng cuốn để ra quyết định để mua. Từ đó, rất nhiều trang web bán sách đã hình thành và cho phép người mua đánh giá về một cuốn sách để người dùng có thể dựa trên mà ra quyết định mua.

Hiện nay, các ứng dụng như vậy đa phần dựa theo các tiêu chí như lượt xem (view), lượt yêu thích (like), cho sao, phân loại theo tiêu đề … để đánh giá một cuốn sách.

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 cuốn sách đó chính là các bình luận (comment) về cuốn sách 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 cuốn sách. Những từ khóa được đánh những trọng số khác nhau biễu diễn mức độ cảm xúc của người dùng với cuốn sách đó.

Từ ý tưởng đó nên em quyết định chọn đề tài đồ án thực tập tốt nghiệp là “Xây dựng website kinh doanh sách có ứng dụng khai phá dữ liệu để gợi ý sách” như là một bước đầu để thể nghiệm ý tưởng đánh giá sách dựa trên từ ngữ biểu thị cảm xúc và hỗ trợ người dùng băng cách gợi ý những cuốn sách được mọi người đánh giá cao để ra quyết định trước khi đến cửa hàng để mua.

# CHƯƠNG 1: GIỚI TIỆU ĐỀ TÀI

1. **Tổng quan**

Đề tài “Xây dựng website kinh doanh sách có ứng dụng khai phá dữ liệu để gợi ý sách” là một ứng dụng trên nền web. Người dùng truy cập và sẽ được gợi ý những cuốn sách được phân loại dựa vào thể loại, tác giả, … Ứng dụng có các chức năng như: tìm kiếm sách theo thể loại hoặc theo tác giả, xem chi tiết sách và cho phép có thể bình luận đánh giá, cũng như xem các đánh giá theo loại và một số chức năng của quản trị viên (admin).

1. **Tìm hiểu khai phá ý kiến người mua sách**

* **Lý thuyết:**
* Tìm hiểu sử dụng ASP.NET và SqlServer.
* 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 sách đơn giản.
  + Xây dựng module phân loại ý kiến.
  + Xây dựng website đơn giản có các chức năng:
    - Quản lý mặt hàng sách.
    - Cho phép người xem nhận xét, bình luận về sản phẩm.
    - Ứng dụng học máy để xếp hạng các sách dựa theo nhận xét của khách hàng.

1. **Công nghệ ứng dụng**

* ASP.NET MVC
* SQL Server
* Entity Framework

1. **Công cụ sử dụng**

* Visual Studio 2017.
* SQL Server 2014 Management Studio.
* Weka 3.8

1. **Mục tiêu đề tài**
   * + Xem và bình luận được các sản phẩm.
     + Xếp hạng sách theo ý kiến của người dùng.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

1. **Giới thiệu các công nghệ**
   1. **ASP.NET MVC**
      1. **ASP.NET**
      * ASP.NET là một nền tảng ứng dụng web (web application framework) được phát triển và cung cấp bởi Microsoft, cho phép những người lập trình tạo ra những trang web động, những ứng dụng web và những dịch vụ web. Lần đầu tiên được đưa ra thị trường vào tháng 1 năm 2002 cùng với phiên bản 1.0 của.NET framework, là công nghệ nối tiếp của Microsoft's Active Server Pages(ASP). ASP.NET được biên dịch dưới dạng Common Language Runtime (CLR), cho phép những người lập trình viết mã ASP.NET với bất kỳ ngôn ngữ nào được hỗ trợ bởi.NET language.
      1. **Giới thiệu về MVC**

* MVC là cụm từ viết tắt của Model-View-Controller, nó phân chia pattern của ứng dụng thành 3 phần - model, controller và view :
  + Model giúp lưu trữ dữ liệu của ứng dụng và trạng thái của nó. Nó là một cơ sở dữ liệu hoặc cũng có thể chỉ là một tập tin XML.
  + View được coi là một giao diện người dùng được sử dụng bởi khách truy cập trang web của bạn để nhìn thấy các dữ liệu. Các trang ASPX thường được sử dụng để hiển thị view trong các ứng dụng ASP.NET MVC.
  + Controller chịu trách nhiệm xử lý các tương tác của người dùng với trang web. Nó được sử dụng để xác định loại view nào cần phải được hiển thị. Controller cũng được sử dụng cho mục đích giao tiếp với model.
* ASP.Net MVClà 1 framework lập trình web của microsoft, công nghệ này ứng dụng mô hình MVC vào trong ASP.Net.
  1. **SQL Server**
* SQL là viết tắt của Structured Query Language, là ngôn ngữ truy vấn mang tính cấu trúc.
* Nó được thiết kế để quản lý dữ liệu trong một hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS).
* SQL là ngôn ngữ cơ sở dữ liệu, được sử dụng để tạo, xóa trong cơ sở dữ liệu, lấy các hàng và sửa đổi các hàng, …
* Tất cả DBMS như MySQL, Oracle, MS Access, Sybase, Informix, Postgres và SQL Server sử dụng SQL như là ngôn ngữ cơ sở dữ liệu chuẩn.
  1. **Entity Framework**
* ADO.NET Entity Framework (EF) là một nền tảng được sử dụng để làm việc với database thông qua cơ chế ánh xạ Object/Relational Mapping (ORM). Do đó EF có thể truy vấn, thao tác với database gián tiếp thông qua các đối tượng lập trình.

1. **Giới thiệu công cụ**
   1. **Visual Studio**

* Ngày 19/11/2007, Microsoft chấm dứt tiến trình thử nghiệm kéo dài tới 18 tháng và tung ra phiên bản chính thức Visual Studio 2008. Đây là một sự kiện quan trọng trong ngành công nghệ tin học.Visual Studio 2008 cung cấp một bộ công cụ mạnh để xây dựng phần mềm cho Windows, Web, các thiết bị di động và Microsoft Office. Dù nền tảng phát triển là gì, Visual Studio vẫn nâng cao hiệu quả hoạt động của nhà phát triển bằng cách cung cấp công cụ để họ viết mã nhanh hơn và đơn giản hoá những tác vụ liên quan đến việc phát triển phần mềm khác.
* Microsoft Visual Studio là môi trường phát triển tích hợp chính (Integrated Development Environment (IDE) được phát triển từ Microsoft. Đây là một loại phần mềm máy tính có công dụng giúp đỡ các lập trình viên trong việc phát triển phần mềm.
  + Các môi trường phát triển hợp nhất thường bao gồm:
* Một trình soạn thảo mã (source code editor): dùng để viết mã.
* Trình biên dịch (compiler) và/hoặc trình thông dịch (interpreter).
* Công cụ xây dựng tự động: khi sử dụng sẽ biên dịch (hoặc thông dịch) mã nguồn, thực hiện liên kết (linking), và có thể chạy chương trình một cách tự động.
* Trình gỡ lỗi (debugger): hỗ trợ dò tìm lỗi.
* Ngoài ra, còn có thể bao gồm hệ thống quản lí phiên bản và các công cụ nhằm đơn giản hóa công việc xây dựng giao diện người dùng đồ họa (GUI).
* Nhiều môi trường phát triển hợp nhất hiện đại còn tích hợp trình duyệt lớp (class browser), trình quản lí đối tượng (object inspector), lược đồ phân cấp lớp (class hierarchy diagram),… để sử dụng trong việc phát triển phần mềm theo hướng đối tượng.
* Như vậy, Microsoft Visual Studio được dùng để phát triển console (thiết bị đầu cuối – bàn giao tiếp người máy) và GUI (giao diện người dùng đồ họa) cùng với các trình ứng dụng như Windows Forms, các web sites, cũng như ứng dụng, dịch vụ wed (web applications, and web services). Chúng được phát triển dựa trên một mã ngôn ngữ gốc (native code ) cũng như mã được quản lý (managed code) cho các nền tảng được được hỗ trợ Microsoft Windows, Windows Mobile, .NET Framework, .NET Compact Framework và Microsoft Silverlight.
* Visual Studio hỗ trợ rất nhiều ngôn ngữ lập trình, có thể kể tên như sau: C/C++ ( Visual C++), VB.NET (Visual Basic .NET), va C# (Visual C#)… cũng như hỗ trợ các ngôn ngữ khác như F#, Python, và Ruby; ngoài ra còn hỗ trợ cả XML/XSLT, HTML/XHTML, JavaScript và CSS…
* Microsoft Visual Studio có những chức năng cơ bản sau: soạn thảo mã ( code editor); Trình gỡ lỗi (debugger); và Thiết kế (Designer). ở đây tôi chỉ trình bày một số công cụ quan trọng của chức năng Designer – đây được xem là một trong những điểm nhấn của microsft visual studio.
  + WinForms Designer: đây là công cụ tạo giao diện đồ họa dùng WinForms. Điểm đặc biệt ở đây là giao diện với người dùng sinh động, dễ nắm bắt. Nó bao gồm các phím bấm, thanh tác vụ, hay các box đa dạng (textbox, list box, grid view…). Bạn có thể di chuyển, kéo ra, nhúng thả chúng một cách dễ dàng.
  + WPF Designer: WPF Designer còn có tên mã là Cider, được hỗ trợ trong Visual Studio 2008. Nó tạo các mã XAML cho giao diện người sử dụng (UI), mã này tích hợp với trình ứng dụng Microsoft Expression Design.
  + Web designer: Visual Studio cũng hỗ trợ cộng cụ thiết kế trang web, trong đó cho phép các công cụ thiết kế trang web được kéo, thả, rê, nhúng một cách dễ dàng… Công cụ này dùng để phát triển trình ứng dụng ASP.NET và hỗ trợ HTML, CSS and JavaScript.
  + Class designer: Đây là công cụ dùng để thực thi và chỉnh sửa lớp. Nó có thể dùng mã C# và VB.NET …
  + Data designer: Đây là công cụ dùng để chỉnh sửa một cách sinh động, linh hoạt các lược đồ dữ liệu, bao gồm nhiều loại lược đồ, liên kết trong và ngoài
  + Mapping designer: Đây là công cụ tạo các mối liên hệ giữa sơ đồ dữ liệu và các lớp để quản lý dữ liệu một cách hiệu quả hơn.
  + Ngoài ra còn có thể kể tên một số công cụ khác như:
    - Open Tabs Browser: Nó được dùng để liệt kể các tab đã mở và chuyển đổI giữa chúng. Bạn cũng có thể dùng phím nóng: CTRL + TAB.
    - Properties Editor: Chức năng dùng để chỉnh sửa các chức năng của các cửa sổ giao diện đồ họa ngườI dùng (GUI) trong Visual Studio. Nó có thể áp dụng cho các lớp, các mẫu định dạng hay trang web và các đốI tượng khác.
    - Object Browser: Đây là một thư viện tên miền và lớp trình duyệt cho Microsoft.NET.
    - Solution Explorer: theo ngôn ngữ của Visual Studio, solution là một bộ phận của mã file và mã nguồn khác được dùng để xây dựng các trình ứng dụng. Công cụ Solution Explorer được dùng để để quản lý và trình duyệt các file trong solution.
    - Team Explorer: Đây là công cụ dùng để hợp nhất các máy tính trong Team Foundation Server, và RCS (revision control system - hệ thống điều khiển xét duyệt) vào trong IDE
    - Data Explorer: Data Explorer dùng để quản lý các dữ liệu trên các phiên bân của Microsoft SQL Server. Nó cho phép tạo lập và chỉnh sửa các bảng dữ liệu được tạo T- SQL commands hay dùng Data designer
    - Server Explorer: Đây là công cụ dùng để quản lý dữ liệu trên máy tính được kết nối.
  + Microsoft Visual Studio đã phát hành các dòng sản phẩm sau:
  + Visual Studio Express: trong đó bao gồm: Visual Basic Express; Visual C++ Express; Visual C# Express ; Visual Web Developer Express.
    - Visual Studio Express: trong đó bao gồm: Visual Basic Express; Visual C++ Express; Visual C# Express ; Visual Web Developer Express.
    - Visual Studio Standard
    - Visual Studio Professional
    - Visual Studio Tools for Office
    - Visual Studio Team System: trong đó bao gồm: Team Explorer (basic TFS client); Architecture Edition; Database Edition; Development Edition; Test Edition

 Về Visual Studio Team System: Năm 2006, Microsoft đã mở rộng dòng sản phẩm Visual Studio của họ thêm một số nhóm sản phẩm có tên gọi là Visual Studio Team System. Chúng có một sản phẩm mới đó là Team Foundation Server cho việc điều khiển mã nguồn, quản lý dự án, kiểm tra và mô hình hóa cũng như một số phiên bản của môi trường phát triển đã được tích hợp Visual Studio 2005 (IDE) hỗ trợ các tính năng của Team Foundation Server.

Visual Studio Team System là một dòng sản phẩm được thiết kế để hỗ trợ sự cộng tác và truyền thông giữa các chuyên gia phát triển phần mềm, những người đang sử dụng Visual Studio IDE. Team System hỗ trợ kiểm soát mã nguồn, quản lý dự án, quản lý xây dựng phần mềm, kiểm tra và các nhóm nhiệm vụ phát triển khác. Nó gồm có Visual Studio Team Foundation Server và một tập các phiên bản đặc biệt của Visual 2005 Studio IDE hỗ trọ các role phát triển cụ thể như các kiến trúc sư, các chuyên gia phát triển phần mềm hay các kiểm tra viên. Các khả năng của Team System gồm có:

* + Kiểm soát mã nguồn: Team System cung cấp một hệ thống kiểm soát mã nguồn mới, hệ thống này cho phép các chuyên gia phát triển có thể thực hiện hài hòa các thay đổi đối với các file mã nguồn cho một dự án.
  + Quản lý dự án: Team System cung cấp một cơ sở dữ liệu quản lý dự án phần mềm có khả năng tùy chỉnh cho việc kiểm tra lỗi, các yêu cầu trong tương lai, các trường hợp thử và lĩnh vực khía cạnh công việc của các chuyên gia phát triển phần mềm khác thông qua toàn bộ một chu trình thiết kế, viết mã và kiểm tra.

Về Visual Studio Tools for Office: Visual Studio Tools for Office cho phép các chuyên gia phát triển phần mềm có thể tạo các ứng dụng tùy thích hoặc ứng dụng để mở rộng Word, Excel và Outlook với sự logic riêng (như một giao diện cuối đối với một hệ thống thanh toán) đang chạy dưới .NET Framework

Phiện bản em đang sử dụng là Microsoft Visual Studio 2013.

* 1. **SQL Server Management Studio:**

SQL Server Management Studio là một công cụ trực quan để quản lý SQL Server. Với SQL Server Management Studio chúng ta có thể thực hiện các tương tác với database trên giao diện người dùng hoặc bằng câu lệnh. SQL Server Management Studio được thiết kế thân thiện và dễ sử dụng.

* 1. **Weka 3.8**
* WEKA là một công cụ phần mềm viết bằng Java, phục vụ lĩnh vực học máy và khai phá dữ liệu.

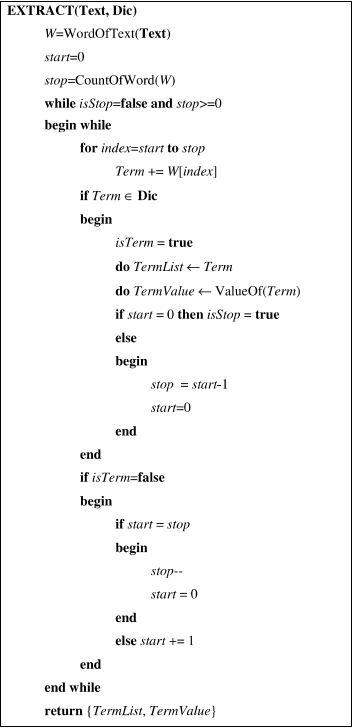
 

Hình 2.2.1 Hình ảnh minh họa công cụ WEKA- Nguồn: <http://www.kdd.org>

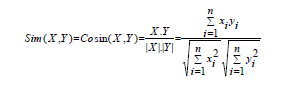
* Các tính năng chính:
  + - Một tập các công cụ tiền xử lý dữ liệu, các giải thuật học máy, khai phá dữ liệu, và các phương pháp thí nghiệm đánh giá.
    - Giao diện đồ họa (gồm cả tính năng hiển thị hóa dữ liệu).
    - Môi trường cho phép so sánh các giải thuật học máy và khai phá dữ liệu,.

1. **Các kỹ thuật trong đề tài**
   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 mỗi cuốn sách, phần bình luận có một bình luận như sau: “hay, cuốn sách 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.
  1. **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 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 bài toán tóm tắt văn bản, nhất là với đối tượng văn bản không tiêu chuẩn như những ý kiến phát biểu.
* Phương pháp này dựa trên một từ điển cảm xúc tiếng Việt để đối sánh.
* Có hai phương pháp đối sánh thực thể dài nhất:
  + - Đối sánh từ phải sang trái.
    - Đối sánh từ trái sang phải.
* Ở đề tài này em sẽ dựa vào thuật toán của dụng phương pháp đối sánh từ phải sang trái để rút trích các ý kiến đánh giá sản phẩm.



Hình 2.3.1: Thuật toán đối xánh văn bản từ phải sang trái

* 1. **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 cuốn sách 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 chủ sách hoặc về các phương diện chính như nôi dung,tác giả giả, chất lượng…
* 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 :

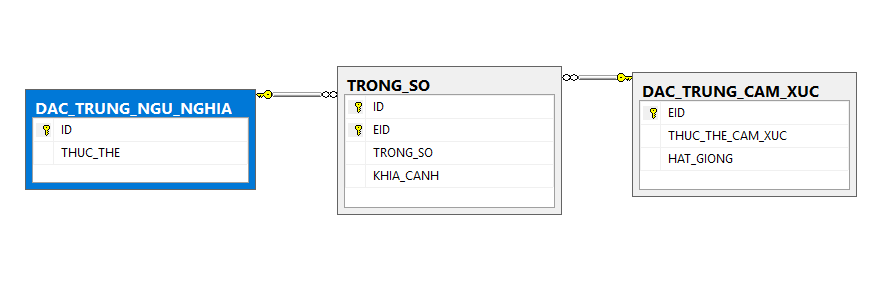
Hình 2.3.2. Công thức tính độ tương đồng của các vector G, P, N

* 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 : “Sách đẹp, nội dung khá hay nhưng viết hơi khó hiểu.”
  + 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 1.
    - “khó hiểu” từ hạt giống có trọng số là -1.
  + Ta có vector G = {1; 1; -1} ; vector N = {0; 0; -1} ; vector P = {1;1; 0}
  + Áp dụng công thức tính cosin và sin ở trên ta có :

Cosin(N,G) = = ; Cosin(P,G) = =

* + So sánh ta thấy : Cosin(P,G) > Cosin(N,G)
  + Vậy suy ra vector G gần vector P hơn nên câu bình luận trên mang ý nghĩa tiêu cực. Tích cực.
  + Tương tự nếu lấy câu bình luận : “ Sách gì tệ hết sức, đươc mỗi cái bìa đẹp.”
  + Ta có các từ được rút trích ra:
    - “tệ hết sức” có trọng số là -2
    - “đẹp” có trọng số là 1
* Vector G = {-2 ; 1} ; vector N = {-2; 0} ; vector P = {0 ; 1}
  + Áp dụng công thức như trên ta có Cosin(N, G) > Cosin(P,G) nên vetor G gần vector N hơn. Câu bình luận trên mang ý nghĩa tiêu cực.
  1. **Thuật toán KMean**
     1. **Khái niệm**
     + K-Means là thuật toán rất quan trọng và được sử dụng phổ biến trong kỹ thuật phân cụm. Tư tưởng chính của thuật toán K-Means là tìm cách phân nhóm các đối tượng (objects) đã cho vào K cụm (K là số các cụm được xác đinh trước, K nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm  
       (centroid) là nhỏ nhất.
     + Thuật toán K-Means thực hiện qua các bước chính sau:
       - Bước 1: Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
       - Bước 2: Tính khoảng cách giữa các đối tượng (objects) đến K tâm. (thường dùng khoảng cách Euclidean).
       - Bước 3: Nhóm các đối tượng vào nhóm gần nhất.
       - Bước 4: Xác định lại tâm mới cho các nhóm.
       - Bước 5: Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.
     1. **Ưu và nhược điểm**
     + *Ưu điểm:*
       - Đơn giản, dễ hiểu và tương đối hiệu quả.
       - Các đối tượng tự động gán vào các nhóm.
       - Thường đạt tối ưu cục bộ.
     + *Nhược điểm:*
       - Số cụm k phải được xác định trước.
       - Tất cả các đối tượng phải gán vào các cụm. Trong trường hợp nhiễu hoặc cá biệt cũng phải gán vào nhóm nào đó gây ra sự không chính xác.
       - Phụ thuộc vào việc chọn nhóm đầu tiên ảnh hưởng tới quá trình gom cụm kéo dài.
       - Nhạy cảm với dữ liệu nhiễu, cá biệt. Do tất cà các đối tượng đều được gán vào cụm nào đó và dữ liệu nhiễu cũng được gán vào các cụm gây ra sự hỗn loạn trong cụm đó.
       - Gặp vấn đề khi nhóm có kích thước, mật độ khác nhau hoặc hình dáng không phải hình cầu thì thuật toán sẽ đưa ra kết quả không chính.

# CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH

1. **Cơ sở dữ liệu:**
   1. **Từ điển cảm xúc**
      * Tự điển cảm xúc là nơi lưu trữ các đặc trưng cảm xúc và các đặc trưng ngữ nghĩa, từ điển cung cấp cho chương trình lấy ra những trọng số tương đương với một cụm ngữ nghĩa và cảm xúc trong một câu sau quá trình rút trích để tính toán đưa ra kết luận rằng đánh giá là tích cực, tiêu cực hay trung lập.
      * Từ điển cảm xúc gồm 3 bảng như *hình 2* bên dưới:

Hình 3.1.1 Từ điển cảm xúc diagram

Hình 2: Từ điển cảm xúc database diagram

* + - Bảng đặc trưng ngũ nghĩa là nơi lưu trừ những từ, cụm từ có vai trò quyết định xem câu đang nói về đối tượng nào và cũng là chủ ngữ của câu.
    - Bảng đặc trưng cảm xúc liệt kê các từ mang ý nghĩa về mặt cảm xúc như khen hay chê trong một câu.
    - Vì đôi khi cùng một từ cảm xúc nhưng với ngữ nghĩa khác nhau thì sẽ mang ý nghĩa khác nhau, nên ta cần phải gán cho mỗi cặp một trọng số cùng khía cạnh của nó.

Bảng đặc trưng ngữ nghĩa

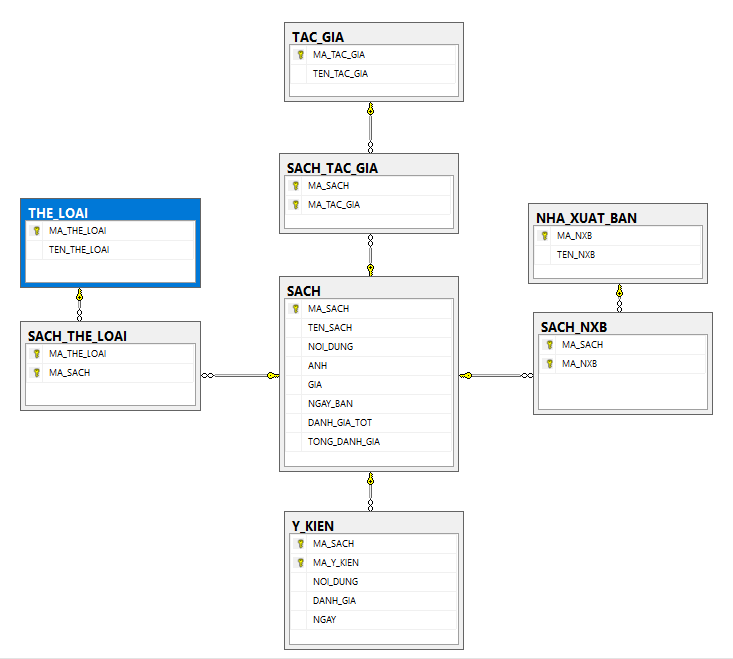
|  |  |  |
| --- | --- | --- |
| **ID** | **Char(5)** | **Primary key(PK)** |
| **THUC\_THE** | **Nvarchar(30)** | **Unique** |

Bảng đặc trưng cảm xúc

|  |  |  |
| --- | --- | --- |
| **EID** | **bigint** | **Primary key(PK)** |
| **THUC\_THE\_CAM\_XUC** | **Nvarchar(30)** | **Unique** |
| **HAT\_GIONG** | **Bit** |  |

Bảng trọng số

|  |  |  |
| --- | --- | --- |
| **ID** | **Char(5)** | **PK,** **Foreign key (FK)** |
| **EID** | **bigint** | **PK,FK** |
| **TRONG\_SO** | **int** |  |
| **KHIA\_CANH** | **Nvarchar(20)** |  |

* 1. **Cơ sở dữ liệu cho website sách**
     + Do mục tiêu đề tài là xây dựng website sử dụng rút trích để đánh giá các đánh giá để chỉ ra sản phẩm nào tốt nên dữ liệu của website tương đối đơn giản.

Hỉnh 3.1.2: Sách Database diagram

Bảng sách

|  |  |  |
| --- | --- | --- |
| MA\_SACH | int |  |
| TEN\_SACH | ntext |  |
| NOI\_DUNG | ntext |  |
| ANH | ntext |  |
| GIA | Char(20) |  |
| NGAY\_BAN | datetime |  |
| Đánh giá tốt | int |  |
| Tổng đánh giá | int |  |

Bảng tác giả:

|  |  |  |
| --- | --- | --- |
| **MA\_TAC\_GIA** | **Int** | **PK** |
| **TEN\_TAC\_GIA** | **nvarchar(30)** |  |

Bảng thể loại

|  |  |  |
| --- | --- | --- |
| **MA\_THE\_LOAI** | **char(10)** | **PK** |
| **TEN\_THE\_LOAI** | **Nvarchar(20)** |  |

Bảng nhà xuất bản

|  |  |  |
| --- | --- | --- |
| **MA\_NXB** | **Char(10)** | **PK** |
| **TEN\_NXB** | **Nvarchar(100)** |  |

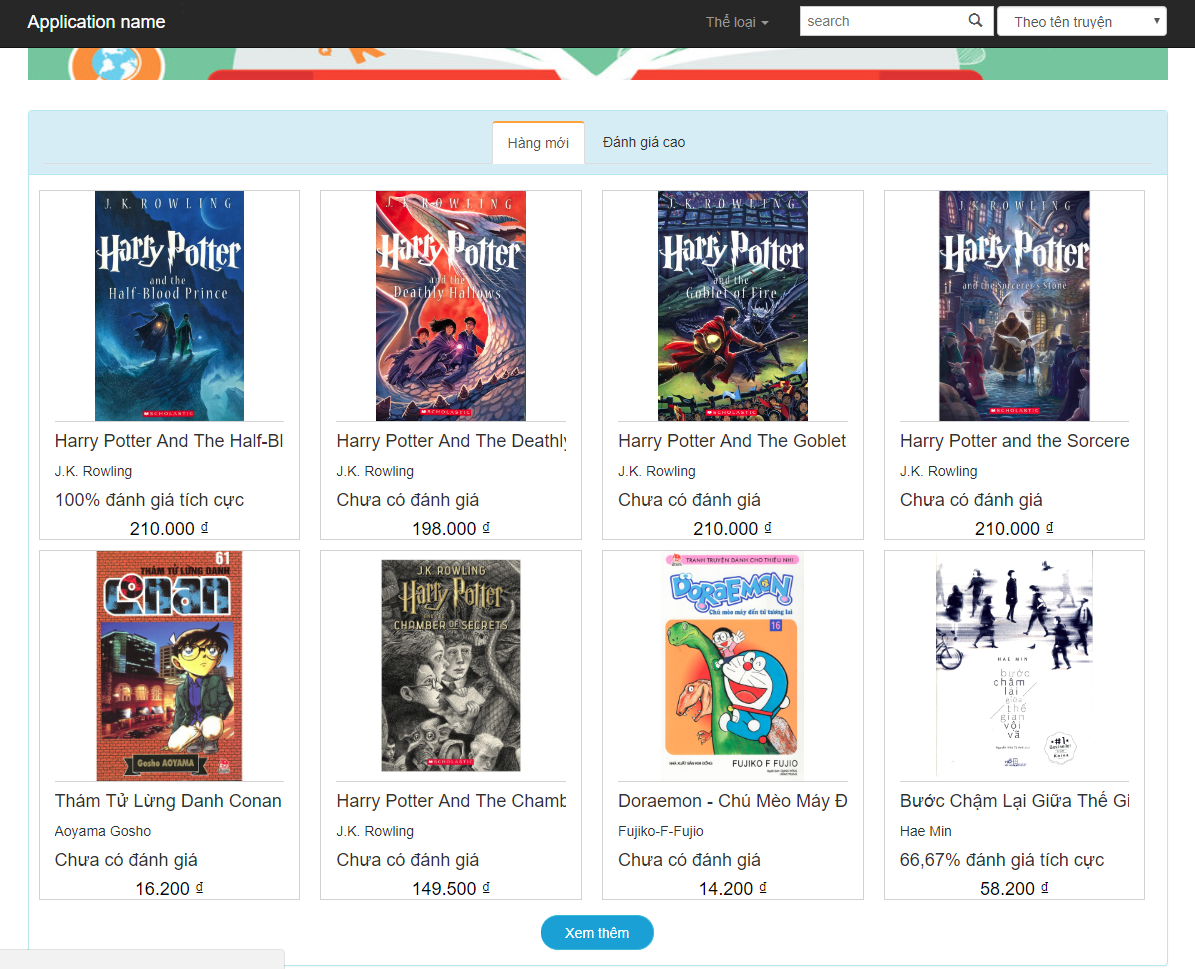
Bảng ý kiến

|  |  |  |
| --- | --- | --- |
| **MA\_Y\_KIEN** | **bigint** | **PK** |
| **MA\_SACH** | **int** | **FK** |
| **NOI\_DUNG** | **ntext** |  |
| **DANH\_GIA** | **Nvarchar(10)** |  |
| **NGAY** | **datetime** |  |

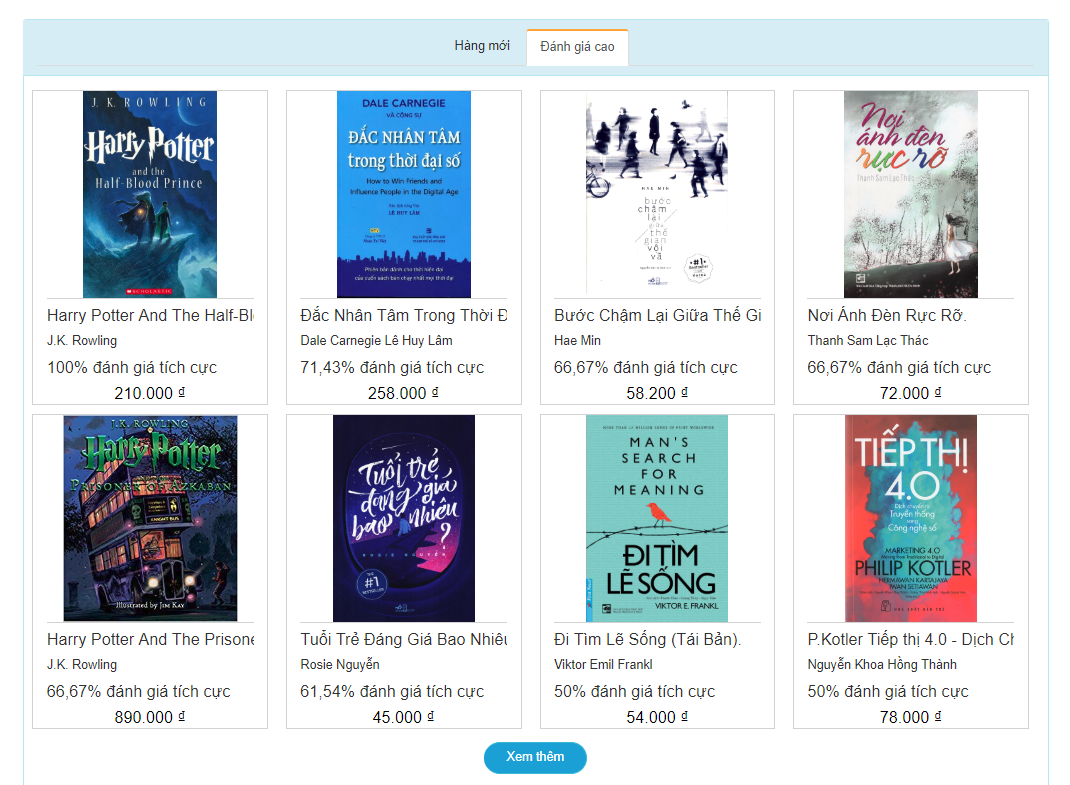
1. **Quá trình khai phá**
   1. **Tri thức ban đầu**
      * Để xây dựng một hệ thống học máy cần phải có một lượng dữ liệu ban đầu để làm tri thức ban đầu cho máy tính.
      * Do đó để xây dựng một hệ thống học máy đánh giá ý kiến để gợi ý sản phẩm cho khách hang, ta cần phải có những đánh giá chuẩn để làm cơ sở cho máy tính để tiền hành đánh giá ý kiến.
      * Để có được những tri thức ấy, trong đề tài này ta chọn lọc những đánh giá có mức độ chuẩn cao thông rút trích dữ liệu và tình toán các vector Tiêu cực (N),  
        vector Tích cực (P) và so sánh với vector tổng (G) để quyết định ra những ý kiến nào là tích cực và tiêu cực.
      * Từ những ý kiến chuẩn ấy ta sẽ tiến hành chuẩn hóa dữ liệu về dạng vector theo các đặc trưng ngữ nghĩa với các từ cảm xúc là trọng số phân chia mức độ.
   2. **Chuẩn hóa dữ liệu**
      * Như đã nêu trên quá trình chuẩn hóa dữ liệu là quá trình tạo ra nhưng tri thức ban đầu cho máy tính để tiến hành học máy để đánh giá các bình luận một cách tự động.
      * Trong đề tài bán sách ta sẽ chuẩn hóa dữ liệu về dạng vector như sau:
        + Khía cạnh tổng quát
        + Khía cạnh nội dung
        + Khía cạnh tác giả
        + Khía cạnh chất lượng sách
        + Khía cạnh trình bày

* Ta sẽ được một bộ dữ liệu sẽ có dạng giống như sau (0,0,0,0,0) với mỗi giá trị tương ứng mỗi khía cạnh và trọng số của từ cảm xúc trong bình luận.
  + - Ta sẽ chuẩn hóa tầm 100 dữ liệu chuẩn để làm tri thức cơ sở trong để tài này.
    - Và tiền hành sử dụng thuật toán KMean để gom cụm các ý kiến.
  1. **Áp dụng KMean**
     + Với những dữ liệu gốc ban đầu ta sẽ bắt đầu gom cụm chúng lại bằng thuật toán KMean xác định chúng vào 3 cụm gồm: tích cực, tiêu cực và trung lập
     + Với mỗi một cụm ta sẽ tính toán và lưu lại trọng tâm của mỗi cụm.
     + Và khi phân loại ý kiến ta sẽ chuẩn hóa ý kiến đó về dạng vector và tính khoảng cách với tâm mỗi cụm, chọn ra cụm gần nhất để đưa ra quyết định cho đánh giá đó thuộc nhóm ý kiến nào.

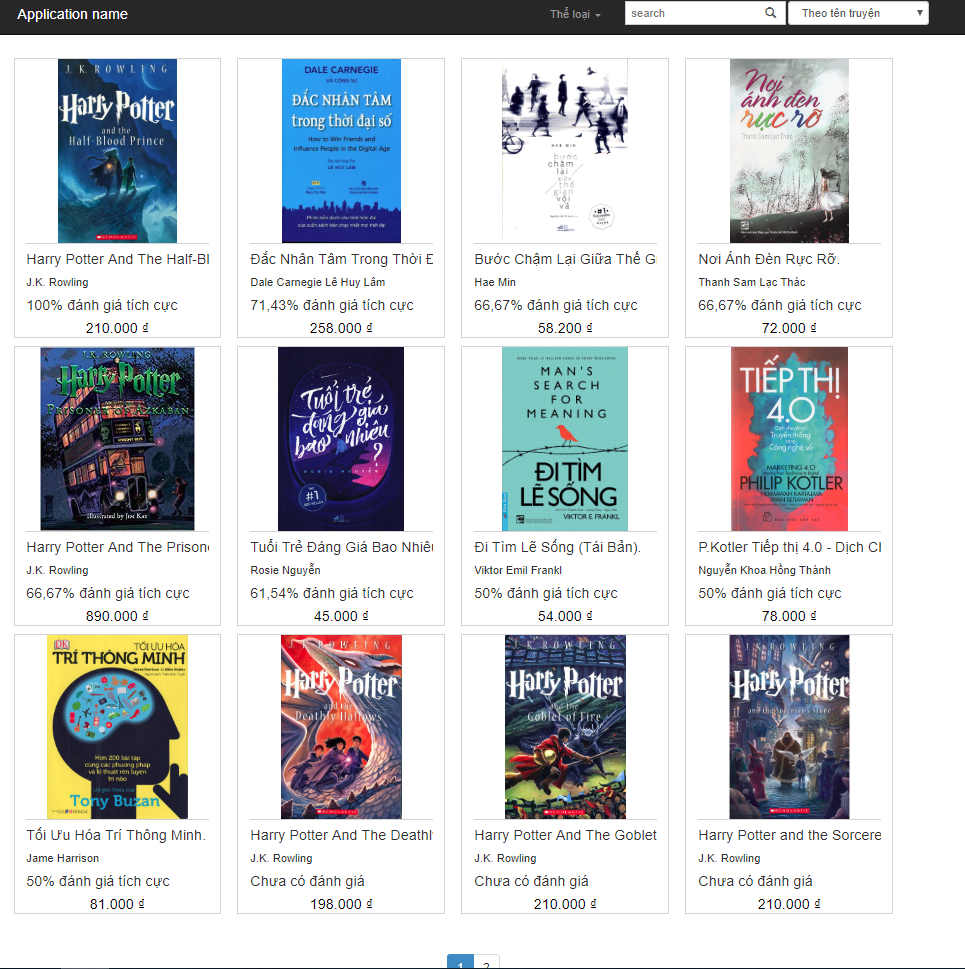
# CHƯƠNG 4: KẾT QUẢ THỰC HIỆN

1. **Trang web**
   1. **Trang chủ**
      * Giao diện:
        + Trang chủ của website bán sách tương đối đơn giản, nó cung cấp cho người dùng danh sách các sản phẩm mới và những sản phẩm có tỉ lệ đánh giá cao.

Hình 4.1.1.1: Trang chủ với gợi ý là những sách mới thêm



Hình 4.1.1.2: Trang chủ với danh sách gợi ý là sách đánh giá cao

* 1. **Danh sách sản phẩm**
     + Giao diện:
       - Trang danh sách sản phẩm cung cấp danh sách sản phẩm khi người dung xem thêm sản phẩm mới, sản phẩm có đánh giá cao, hay tìm kiếm sản phẩm.

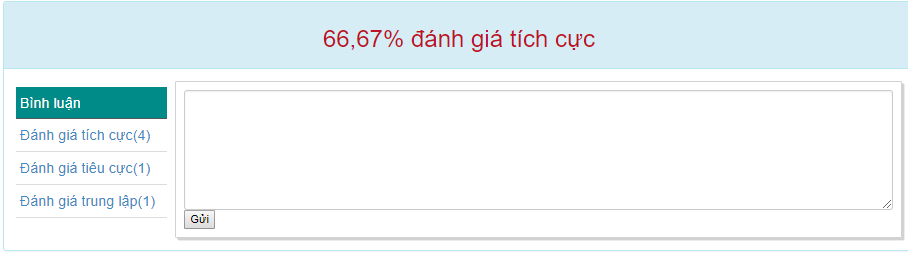
Hình 4.1.2: Trang danh sách sản phẩm theo sách mới

* + - Chức năng:
      * Sắp xếp danh sách sản phẩm theo tỉ lệ đánh giá sản phẩm giảm dần khi người dung tìm kiếm theo tác giá, thể loại hay xem danh sách sản phẩm đánh giá tốt
  1. **Trang chi tiết sản phẩm**
     + Giao diện
       - Trang chi tiết sản phẩm cung cấp cho người dung vè thông tin về sách

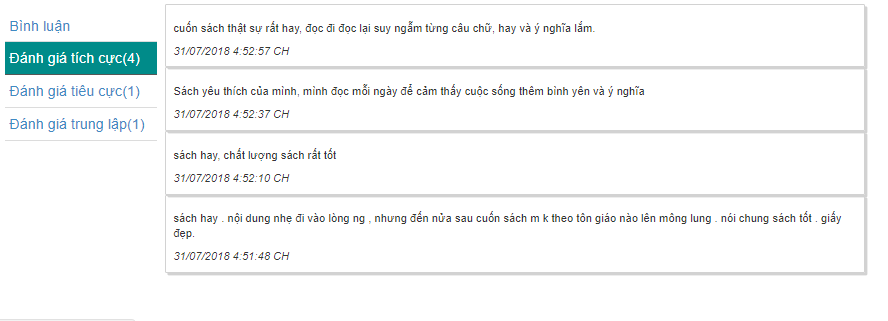
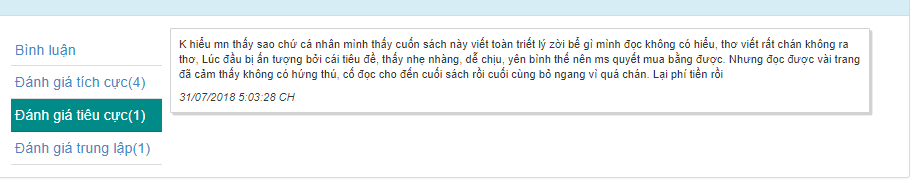


Hình 4.1.3.1: Chi tiết sản phẩm

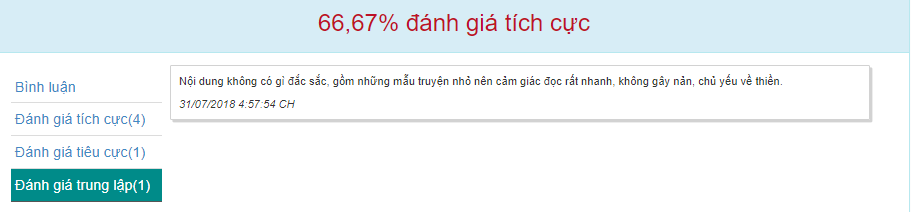
* + - **Chức năng**
      * Bình luận : Đưa ra ý kiến về sản phẩm
        + Sau khi bình luận, ý kiến sẽ qua module rút trích và được phân loại vào các các loại tương ứng.



Hình 4.1.3.2: Chức năng bình luận

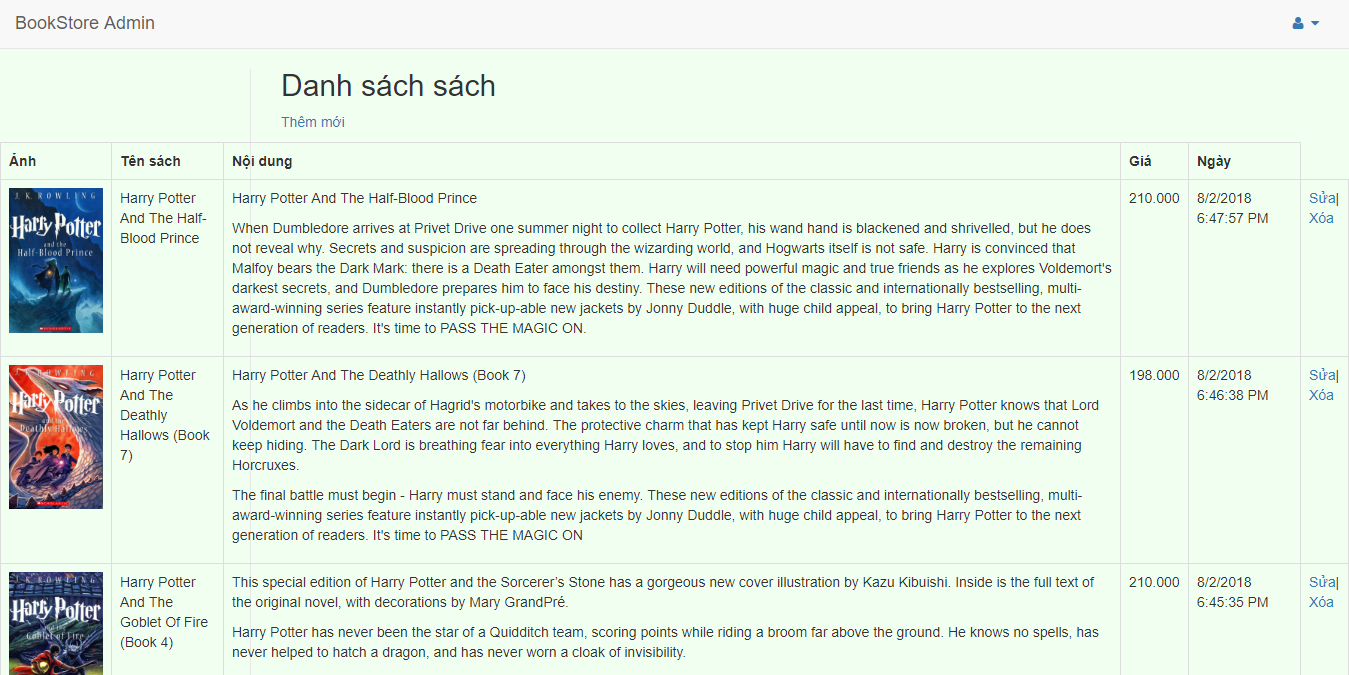
* + - * Xem bình luận : Cho phép người dung lựa chọn xem đánh giá theo từng loại để tham khảo.

Hình 4.1.3.3: Xem ý kiến tích cực

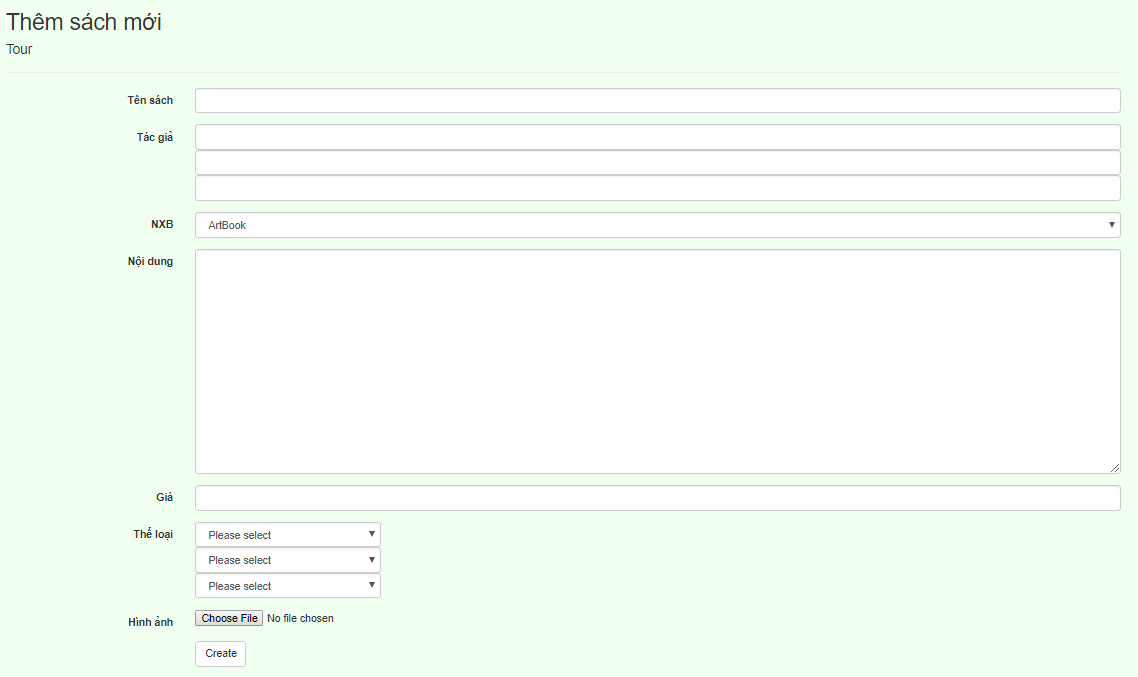


Hình 4.1.3.4: Xem ý kiến tiêu cực

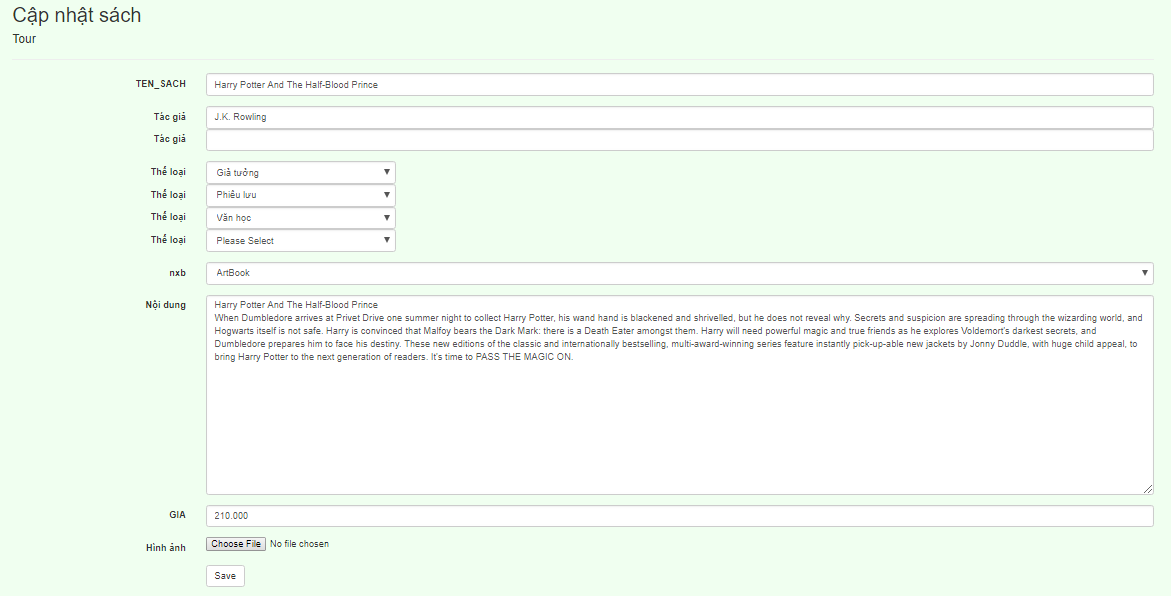
Hình 4.1.3.5: Xem ý kiến trung lập

* 1. **Trang quản lý**
     + Giao diện:
       - Do không có thời gian chau chuốt nên trang quản lý tương đối đơn giản

Hình 4.1.4.1: Trang quản lý sách

* + - Chức năng:
      * Thêm sách : giúp người quản lý cập nhất sách vào database.

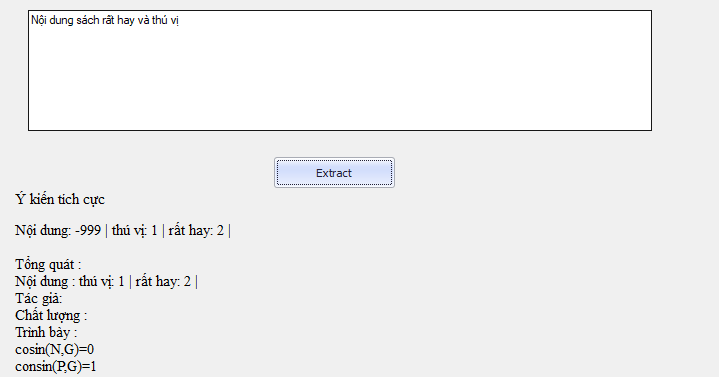
Hình 4.1.4.2: Trang thêm sách

* + - Chức năng sửa

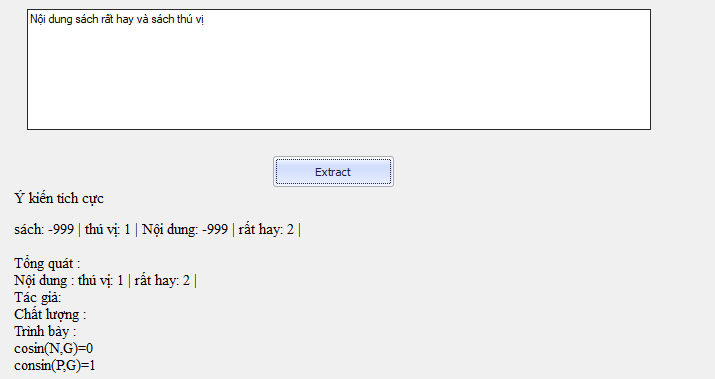
Hình 4.1.4.3: Trang sửa thông tin sách

1. **Xây dựng module rút trích, phân trích và phân loại cảm xúc**
   * + Sử dụng nguồn dư liệu về các ngữ nghĩa và cảm xúc về chủ đề Sách để rút trích các cụm từ trong một đánh giá về mặt hàng Sách.
     + Do ngôn ngữ tiếng Việt rất đa dạng và vốn từ phong phú và đa nghĩa với vốn dữ liệu trong từ điển cảm xúc chưa thể bao quát hết nên chắc chắn khi rút trích sẽ có sai sót trong quá trình rút trích.
       - Ví dụ: từ “hay” khi nói về một sách thì là khen nhưng trong một câu người ta có thể sử dụng từ “hay” như từ “hoặc” hoàn toán không có ý nghĩa cảm xúc nào, do đó gây nên sự sai sót khi tính toán và ra quyết định cho đánh giá.
     + Thuật toán rút trích sẽ rút trích một văn bản thánh nhiều câu và một câu thành từng vế để rút ra trọng số của cặp ngữ nghĩa và cảm xúc để tính toán và đánh giá cả một văn bản đánh giá.
     + Dưới đây là module rút trích ở một win-form để demo rõ hơn về module này.
     + Trong đề tài này thuật toán sẽ rút trích được những loại câu như sau:
       - Câu không có chủ ngữ (ngữ nghĩa)
         * Thì từ cảm xúc thì sẽ kết hợp với với từ mặc định là “sách”

Hình 4.2.1: Ví dụ rút trích 1

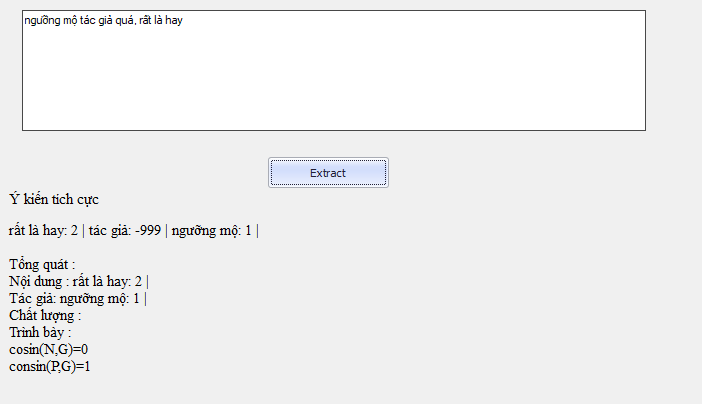
* + - * + Do mỗi từ cảm xúc kết hợp với ngữ nghĩa khác nhau thì sẽ mang ngữ nghĩa khác nhau, nên khi nói “sách rất hay” thì nó sẽ phân loại là “nội dung”, nhưng nếu khi nói “sách rất tốt” thì sẽ là “tổng quát” vì từ tốt có thể nói về cả “nội dung” và “chất lượng” sách.
        + Nhưng nếu cặp ngữ nghĩa và cảm xúc không có trong từ điển thì thuật toán sẽ bỏ qua, làm cho việc đánh giá câu sẽ có sai sót.
      * Nếu 1 cấu trúc câu có nhiều từ ngữ nghĩa thì sẽ chọn từ ngư nghĩa ở đầu câu để kết hợp với những từ cảm xúc có trong cùng cấu trúc câu để rút ra trọng số và khía cạnh.
        + Trong câu trên gồm có 2 ngữ nghĩa đó là Nội dung và sách được lưu trong từ điển, thì thuật toán rút trích sẽ chọn Nội dung vì nội dung là ngữ nghĩa năm ở đâu câu và kết hợp cùng những từ cảm xúc có trong câu để rút ra trọng số để tính toán.

Hình 4.2.2: Ví dụ rút trích 2

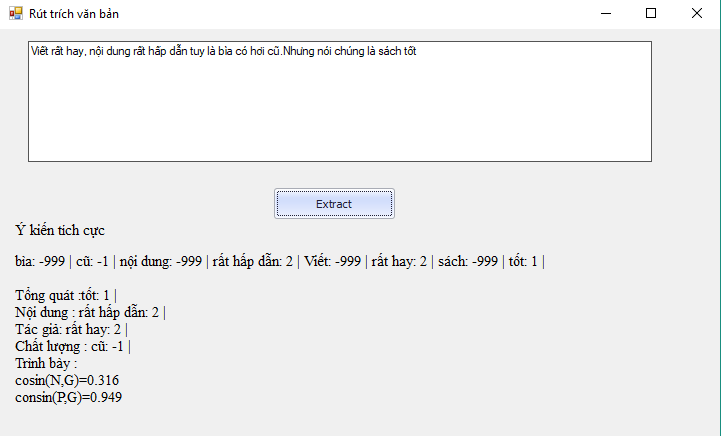


Hình 4.2.3: Ví dụ rút trích 3

* + - * + Trong ví dụ này vì sau “và” là 1 câu trúc câu khác nên “thú vị” sẽ kết hợp với “sách” và “Nội dung” sẽ kết hợp với “rất hay”.
        + Vì thuật toán nhận các từ và kí tự sau như ‘ , ’ ; ‘ ; ’ ; ‘ và ‘ ; ‘ nhưng ‘ ; … các từ phân vế trong tiếng việt để tách câu thành các vế. Và khi kết thúc một vế câu có cả ngữ nghĩa và cảm xúc thì sẽ bắt cặp chúng để lấy ra trọng số.
        + Trong đề tài ta chỉ dùng 1 số từ và kí hiệu để tách câu, nên chắc chắn khi rút trích sẽ có sai sót.
        + Nhưng nếu trong vế không có chủ ngữ thì từ cảm xúc sẽ lưu lại và dùng để kết hợp cùng với chủ ngữ của vế tiếp theo. Trong trường hợp cả 2 câu đều không có chủ ngữ thì sẽ mặc định kết hợp với ngữ nghĩa là “sách”.

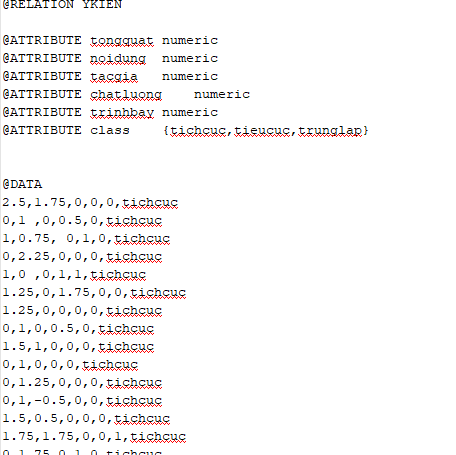
* + - * Các câu có chủ ngữ và vị ngữ đảo ngược.
        + Trong câu dưới từ “ngưỡng mộ” sẽ được kết hợp với “tác giả”, tuy nhiên  
          từ “rất là hay” sẽ kết hợp với từ mặc định là “sách”.
        + Vì theo ý kiến chủ quan thì đa phần những từ cảm xúc không chủ ngữ sau những cấu trúc ngược chưa chắc nói về chủ ngữ của vế câu trước nên sẽ để chúng ghép với ngữ nghĩa mặc đinh là “sách”.

Hình 4.2.4: Ví dụ rút trích 4

1. **Chuẩn hỏa dữ liệu và học máy để đánh giá.**
   1. **Chuẩn hóa**
      * Từ điển cảm xúc ta có một trường là khía cạnh lưu khía cạnh của cặp từ ngữ nghĩa và cảm xúc.
      * Do vậy sau khi rút trích ta sẽ được một danh sách trọng số của mỗi khía cạnh, sau đó ta sẽ chọn ra một trọng số đại diện cho các khía cạnh. Nếu trong trường hợp khía cạnh không có bất cứ trọng số nào thì mặc định khía cạnh đó sẽ là 0.
      * Sau khi chuẩn hóa ta sẽ được một vector gồm 5 trọng số của các thuôc tính như đã liệt kê ở [chương 3](#_CHƯƠNG_3:_XÂY) mục 2.

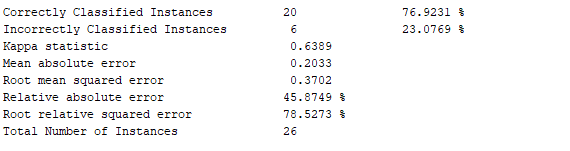
Hình 4.3.1: Ví dụ chuân hóa dữ liệu

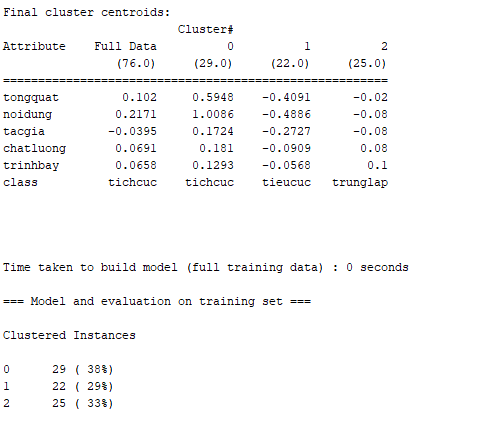
* + - Như ví dụ trên ta sẽ có vector như sau (1,2,2,-1,0).
  1. **Học máy đánh giá ý kiến**
     + Kho dữ liệu
       - Thực hiện bằng thủ công trên 76 bình luận để chuẩn hóa và lưu trữ vào một file có định dạng arff.



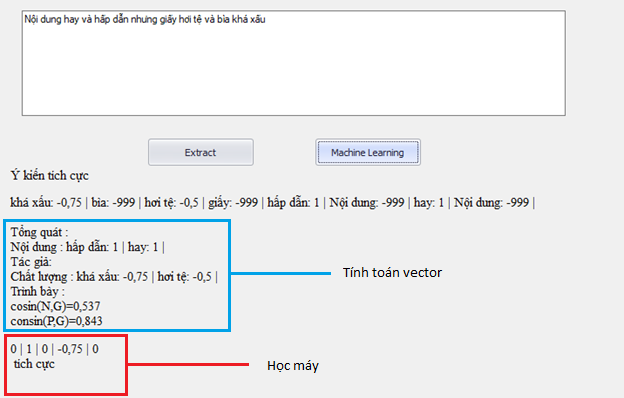
Hình 4.3.2.1: File kho dữ liệu ý kiến

* + - * Sử dụng weka để xem xét độ tin cậy của kho dữ liệu



* + - * + Thử nghiệm dung 66% kho để huấn luyện và 33% còn lại để đánh giả.
* ta được 20 mẫu chính xác (76.9231%) và 6 mẫu sai (23.0769%). Tỉ lệ cũng tương đối cao, nên kho dữ liệu có thể sử dụng được  
    
  + - * Tiến hành gom cụm với kho dữ liệu thành 3 cụm  
        

Hình 4.3.2.2: Kết quả gom cụm với thuật toán KMean sử dụng weka

* + - * Ta được 3 cụm dữ liệu và tâm của mỗi cụm:
        + Cụm tích cực: 0.5948, 1.0086 , 0.1724, 0.181, 0.1293
        + Cụm tiêu cực: -0.4091, -0.4886, -0.2727, -0.0909, -0.0568
        + Cụm trung lập: -0.02, -0.08, -0.08, 0.08, 0.1
      * Ta sẽ lưu lại tâm của 3 cụm lại như một biến hằng để sau khi rút trích sẽ tính khoảng cách giữa mỗi bình luận đã được chuẩn hóa và từng tâm để tìm ra xem là đánh giá ấy gần cụm nào nhất từ đó sẽ chỉ ra loại của ý kiến.
    - Do chưa tích hợp vào website nên là chỉ demo ở chương trình win-form để thể hiện.

Hình 4.3.2.3: Rút trích dữ liệu và đánh giá văn bản về học máy và tính toán vector

* + - * Với ví dụ trên với từng thuộc tính khi chuẩn hóa ta sẽ chọn giá trị max (dương) hoặc min (âm) của từng thuộc tính để làm giá trị đại diện.
      * Sau đó tính khoảng cách với từng tâm để tìm tâm gần nhất, trong ví dụ trên ta sẽ có lần lượt khoảng cách như sau:
        + Với tâm tích cực: khoảng cách là 1.1256730964183164
        + Với tâm tiêu cực: khoảng cách là 1.7015572602765974
        + Với tâm trung lập: khoảng cách là 1.368247053715081
* Như vậy ta thấy được là vector chuẩn hóa của câu ý kiến trên gần với tâm tích cực nhất, nên ý kiến sẽ được gom vào cụm tích cực cũng như sẽ là loại ý kiến tích cực.

# CHƯƠNG 5: TỔNG KẾT

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 gợi ý các cho người dùng những cuốn sách hay và đáng mua. Trích xuất ý kiến người dùng để dùng cho nhiều mục đích hữu ích như xếp hạng sach, 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, hướng đối tượng.
* Khai phá nguồn dữ liệu văn bản tiếng Việt.
* Hoàn thiện cách thức thiết kế cơ sở dữ liệu.
* **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 3 lớp (MVC) 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 và mở rộng từ điển cảm xúc:
  + Bỏ bớt những từ dư thừa
  + Cấu trúc lại database từ điển
  + Phân loại lại mức độ (trọng số) và khía cạnh của các cặp từ.
* Tối ưu hóa chức năng rút trích và ứng dụng học máy để đạt độ chính xác cao hơn và nhanh hơn.
* Một số giao diện chưa thân thiện với người dùng.
* Chưa làm kịp chức năng học máy để đánh giá ý kiến.
* Và một số tính năng cơ bản.

# DANH MỤC TÀI LIỆU THAM KHẢO

**Tiếng Việt :**

1. Nguyễn Tài Cần, “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://www.w3schools.com>
2. <https://stackoverflow.com>
3. <https://tiki.vn/nha-sach-tiki/c8322?src=mega-menu>
4. <https://www.vinabook.com/>