HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Đỗ Ngọc Sơn

KHÁM PHÁ QUAN TÂM CỦA NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI PHỤC VỤ CÔNG TÁC TRUYỀN THÔNG TUYỂN SINH CỦA TRƯỜNG ĐẠI HỌC

Chuyên ngành: Hệ thống thông tin

Mã số: 8.48.01.04

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI - 2020

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **PGS.TS TRẦN ĐÌNH QUẾ**

Phản biện 1: **PGS.TS. NGUYỄN LONG GIANG**

Phản biện 2: **TS. NGUYỄN DUY PHƯƠNG**

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 9 giờ 00 ngày 20 tháng 6 năm 2020

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Thành tố quan trọng nhất trong thời đại bùng nổ công nghệ thông tin hiện nay là mạng Internet. Nói tới Internet, là nói tới các kết nối trực tuyến và sự tiện lợi. Internet thúc đẩy ứng dụng công nghệ thông tin trong phát triển kinh tế, văn hóa, xã hội và đây còn được xem là nhân tố giúp cho các quốc gia đang phát triển dần bắt kịp với các quốc gia hàng đầu trên thế giới. Internet thực sự là một trong những phát minh có tầm ảnh hưởng lớn nhất trong lịch sử loài người. Khi các dịch vụ Internet phát triển, đặc biệt là sự xuất hiện các mạng xã hội và các thiết bị di động thông minh, con người tương tác đa chiều hơn, phản ánh sinh động hơn, tức thời hơn mọi mặt đời sống. Từ quá trình này, con người thể hiện đa dạng đời sống và các quan hệ xã hội trên Internet, biến Internet thành không gian xã hội, hay không gian mạng, nơi có thể giao tiếp, lao động, sáng tạo, học tập, sản xuất, tiêu dùng, vui chơi, giải trí….

Với yếu tố phổ biến, bám sát vào gần như mọi mặt của đời sống cũng như tâm tư tình cảm và dường như là một phần không thể thiếu được của con người hiện nay đó chính là mạng xã hội mà điển hình là Facebook, Twitter, Youtube, Instagram hay Zalo…. Mạng xã hội là nơi mà người dùng cập nhật những thông tin, sở thích, mối quan tâm của bản thân, chia sẻ và nói lên những quan điểm, đánh giá về mọi lĩnh vực trong xã hội như kinh tế, văn hóa, giáo dục, chính trị…. Từ đó, mạng xã hội ngày càng tạo ra một lượng dữ liệu khổng lồ. Với lượng thông tin khổng lồ mà người dùng tạo ra từ mạng xã hội đó là thách thức nhưng cũng là điều kiện thuận lợi để các nhà khoa học, doanh nghiệp hay các chính phủ nghiên cứu và phát hiện những quan tâm, nhu cầu cũng như viện định hướng cho quan điểm của người dùng.

Với mạng xã hội người dùng sẽ thể hiện mối quan tâm, quan điểm của mình bằng cách thích (like), chia sẻ (share) các bài viết của người dùng khác hay bằng chính các bài viết (status) và bình luận (comments) của họ. Người dùng mạng xã hội sẽ thể hiện rõ ràng đầy đủ nhất những mối quan tâm, những gì muốn truyền đạt thông qua các bài viết của họ. Thông qua những bài viết này ta có thể khám phá ra những lĩnh vực, vấn đề mà người dùng mạng xã hội quan tâm.

Hiện nay giáo dục cũng chính là một loại hình "dịch vụ", cũng như các loại hình dịch vụ khác giáo dục cũng cần có các nỗ lực tiếp thị và thúc đẩy hình ảnh. Bên cạch yếu tố chất lượng được đặt lên hàng đầu thì yêu tố tiếp thị hình ảnh đang ngày càng được coi trọng. Ở Việt Nam những năm gần đây, từ những trường đại học lớn tới các trường đại học nhỏ việc

thu hút sinh viên giỏi, xây dựng thương hiệu và tên tuổi đang là nhiệm vụ sống còn, trong bối cảnh các trường đang phát triển theo lộ trình tự tuyển sinh, cũng như tự chủ về tài chính.

Vì vậy, tác giả chọn đề tài "Khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học". Luận văn sẽ dựa trên việc phân tích các bài viết (status) của người dùng trên mạng xã hội để khám phá mối quan tâm của người dùng liên quan đến những ngành học cụ thể nào của một trường đại học và từ đó sẽ đưa ra các phương án truyền thông phù hợp.

Tổng quan về vấn đề nghiên cứu

Tại Việt Nam, mạng xã hội đặc biệt nở rộ và phát triển hết sức mạnh mẽ. Nó khiến nhiều người chú ý và quan tâm, đã có nhiều nghiên cứu cũng như bài viết về việc khai thác nguồn dữ liệu khổng lồ của mạng xã hội để phục vụ cho các mục đích khác nhau. Hiện nay nhiều cơ quan nhà nước cũng như doanh nghiệp, cá nhân cũng đã tận dụng dữ liệu từ mạng xã hội để tìm hiểu những vấn đề người dùng quan tâm nhằm đưa ra những chính sách và chiến lược kinh doanh phù hợp. Có thể kể đến nghiên cứu "Ước lượng quan tâm người dùng trên mạng xã hội dựa trên tương tự bài viết" [1] của PGS.TS. Trần Đình Quế và cộng sự. Nghiên cứu đề xuất một mô hình dựa trên việc phân tích các bài viết của người dùng trên các mạng xã hội để phát hiện và so sánh tương quan về quan tâm của họ. Kết quả thực nghiệm cho thấy rằng nếu hai người dùng có nhiều bài viết giống nhau thì sẽ có quan tâm tương tự nhau và ngược lại, nếu hai người dùng có quan tâm giống nhau thì cũng có nhiều bài viết tương tự nhau.

Trên thế giới đã có nhiều công trình nghiên cứu về vấn đề phát hiện quan tâm người dùng trên mạng xã hội. Điển hình như Schwartz và các cộng sự của ông [7] đã đề xuất mô hình phân tích dựa trên đồ thị để phát hiện quan tâm người dùng có cùng sở thích. Tuy nhiên các tiếp cận bài toán phát hiện quan tâm người dùng trong mạng xã hội bằng mô hình này chỉ tập trung vào việc tìm kiếm, phát hiện quan tâm người dùng trong một tổ chức cộng đồng đã có sự quan tâm nhất định về một chủ đề nào đó.

Với một số mạng xã hội khác như hệ thống mạng xã hội Del.icio.us, Xin Li và các cộng sự [9] đã chỉ ra rằng tần suất xuất hiện của một số tag và hash-tag trong mạng xã hội này có xu hướng ổn định. Tức là một người có xu hướng quan tâm một vấn đề trong thời gian nhất định. Dựa vào tag và hash-tag, có thể phát hiện quan tâm người dùng trong khi họ không

thuộc một tổ chức cộng đồng nào.

Với lĩnh vực công tác thuộc về truyền thông của một trường đại học, đề tài mà tác giả quan tâm là phân tích cơ sở dữ liệu của mạng xã hội đem lại để có thể định hướng, tiếp thị hình ảnh thương hiệu cũng như giới thiệu các ngành học của trường đại học một cách chủ động, đến những đối tượng người học tại những khu vực cụ thể. Từ đó nâng cao chất lượng quảng bá thương hiệu và phục vụ đắc lực cho công tác truyền thông thu hút sinh viên.

Luận văn này sẽ tập trung vào việc xử lý bài toán khám phá quan tâm của người dùng mạng xã hội dựa vào các bài viết (status) để phục vụ công tác truyền thông tuyển sinh của trường đại học.

Mục tiêu nghiên cứu

Mục tiêu nghiên cứu của luận văn là nghiên cứu bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học dựa trên bài viết của người dùng và thử nghiệm đánh giá kết quả bài toán.

Cụ thể như sau:

- Tìm hiểu về bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

- Nghiên cứu sử dụng Mô hình túi từ - Bag of Words (Bow) TF-IDF trong khai phá dữ liệu văn bản.

- Khảo sát các cách phân loại người dùng dựa trên bài viết trên mạng xã hội.

Cấu trúc của luận văn

Nội dung của luận văn ngoài phần mở đầu và phần kết luận được chia làm 03 chương với bố cục như sau:

Mở đầu: Khái quát về đề tài, tổng quan về vấn đề nghiên cứu và cấu trúc của luận văn.

Chương 1: Tổng quan về nghiên cứu quan tâm người dùng trên mạng xã hội: Giới thiệu về mạng xã hội và bài toán khám phá quan tâm người dùng trên mạng xã hội. Trình bày các vấn đề liên quan đến bài toán này như khai phá dữ liệu, biểu diễn dữ liệu văn bản. Ý nghĩa và những khó khăn thách thức trong việc giải quyết bài toán

Chương 2: Khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác

truyền thông tuyển sinh của trường đại học: Trình bày Mô hình túi từ - Bag of Words (Bow) TF-IDF dùng để xử lý ngôn ngữ tự nhiên. Đưa ra một mô hình xử lý bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học dựa vào bài viết trên mạng xã hội.

Chương 3: Thử nghiệm và đánh giá: Phát biểu bài toán bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học dựa vào bài viết trên mạng xã hội. Giới thiệu bộ dữ liệu về các bài viết thu thập được trên mạng xã hội, các công cụ và phần mềm để xây dựng chương trình thử nghiệm. Một số kết quả và đánh giá kết quả cho bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

Kết luận: Tóm lược các kết quả đạt được của luận văn và định hướng nghiên cứu trong tương lai.

Chương 1: TỔNG QUAN VỀ NGHIÊN CỨU QUAN TÂM NGƯỜI DÙNG TRÊN MẠNG XÃ HỘI

1.1 Tổng quan về Mạng xã hội

1.1.1 Giới thiệu về Mạng xã hội

Các mạng xã hội phổ biến nhất hiện nay ở Việt Nam có thể kể đến như Facebook, Youtube, Twitter, Instagram, Zalo…. Zalo là mạng xã hội Việt phổ biến nhất hiện nay. Zalo được phát triển từ một ứng dụng chat đa phương tiện (OTT) và dần mở rộng tính năng chia sẻ thông tin trên tường theo dòng thời gian (timeline) tương tự các mạng xã hội chính thức khác. Zalo hiện đã thu hút được hơn 100 triệu tài khoản người dùng tại Việt Nam.

1.1.2 Đặc điểm của Mạng xã hội

Nhìn chung có nhiều mô hình mạng xã hội khác nhau, nhưng hầu hết mạng xã hội có những đặc điểm chung như:

+ Mạng xã hội là ứng dụng trên nền tảng Internet

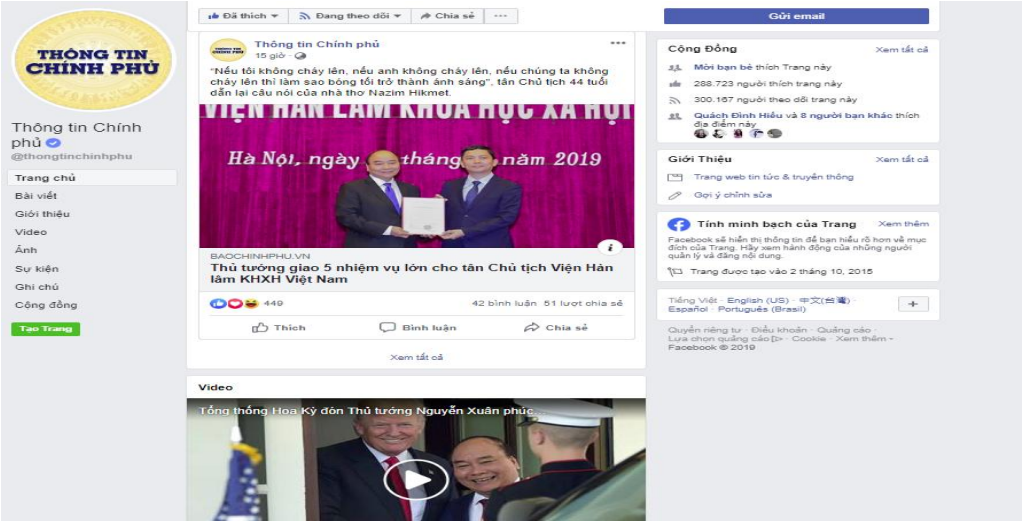
+ Nội dung trên mạng xã hội là do người dùng tự sáng tạo, chia sẻ

+ Người dùng tạo ra hồ sơ cá nhân phù hợp cho trang hoặc ứng dụng được duy trì trên nền tảng mạng xã hội

+ Mạng xã hội tạo điều kiện cho sự phát triển của cộng đồng xã hội trên mạng bằng cách kết nối tài khoản của người dùng với tài khoản của các cá nhân, tổ chức khác.

1.1.3 Ứng dụng của Mạng xã hội

Sự tiếp cận đến từng cá nhân người dùng với tốc độ nhanh tạo ra nhiều cơ hội và lợi ích về truyền tải, tiếp nhận, chia sẻ, thông tin, tri thức; phục vụ các nhu cầu đa dạng của cộng đồng như: kết bạn, giải trí, kinh doanh, bày tỏ quan điểm, phản biện xã hội, lan tỏa những điều tốt đẹp…



Hình 1.1. Trang Thông tin Chính phủ của Việt Nam trên mạng xã hội Facebook

Hình 1.2. Sử dụng mạng xã hội để kinh doanh, quảng cáo đang trở nên rất phổ biến và nở rộ

Hình 1.3. Những dòng trạng thái của Tổng thống Donald Trump luôn nhận được lượng tương tác rất lớn trên mạng xã hội Twitter

1.2 Bài toán nghiên cứu quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học

1.2.1 Bài toàn nghiên cứu quan tâm người dùng trên mạng xã hội

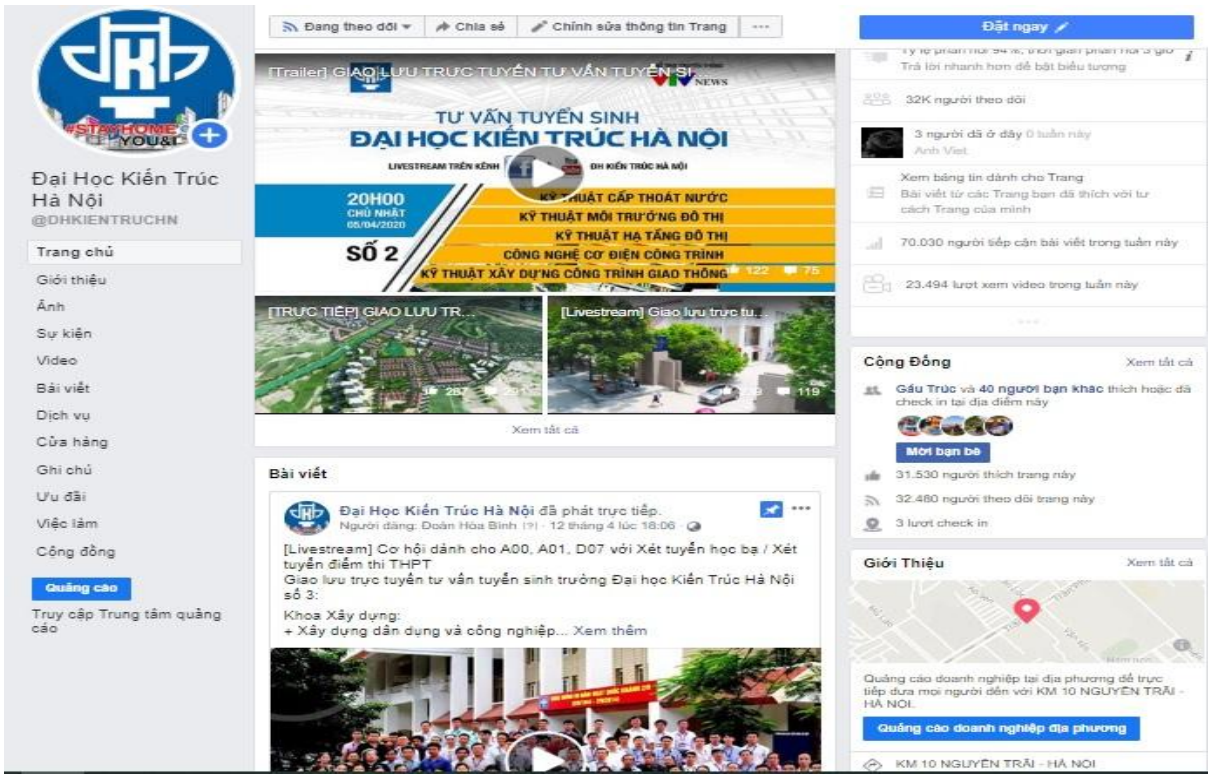
Mạng xã hội đã, đang và sẽ tiếp tục là một công cụ làm việc, ứng dụng giải trí, nguồn thông tin quan trọng trong cuộc sống của người Việt Nam. Mỗi ngày, trung bình một người trưởng thành (trên 16 tuổi) dành khoảng 2.12 tiếng để truy cập mạng xã hội theo "*Báo Cáo Nghiên Cứu Thói Quen Sử Dụng Mạng Xã Hội Của Người Việt Nam 2018*" [11]

Hình 1.4. Người dùng tương tác với một bài viết về thông tin tư vấn tuyển sinh trên mạng xã hội Facebook

Luận văn sẽ dựa trên phân tích về nội dung đưa lên của người dùng mạng xã hội để trình bày bài toán nghiên cứu quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

1.2.2 Bài toán nghiên cứu quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học

Như đã nói ở trên khám phá quan tâm của người dùng mạng xã hội đang thu hút được rất nhiều sự quan tâm nghiên cứu. Với việc khai phá dữ liệu có được từ mạng xã hội, các công ty có thể dựa trên những phân tích đánh giá xu hướng và tiếp cận tới khách hàng nhiều nhất có thể hay chính quyền cũng có thể đánh giá được sự hài lòng của người dẫn về các chính sách quản lý của mình.... Với thực tiễn hiện nay giáo dục cũng chính là một loại hình "dịch vụ" và cũng như các dịch vụ khác giáo dục cũng cần các nỗ lực tiếp thị hình ảnh.



Hình 1.5. Fanpage trường Đại học Kiến trúc Hà Nội trên mạng xã hội Facebook

Hình 1.6. News feed của ông chủ Facebook Mark Zuckerberg

Luận văn này sẽ trình về bày bài toán khám phá quan tâm của người dùng dựa vào bài viết cũng như các ý kiến, bình luận trên mạng xã hội để phục vụ định hướng cho công tác tuyển sinh của một trường đại học.

1.2.3 Ý nghĩa của bài toán

Bài toán khám phá quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học nói riêng và bài toán khám phá quan tâm của người dùng trên mạng xã hội nói chung đều có những ý nghĩa mang tính thời sự.

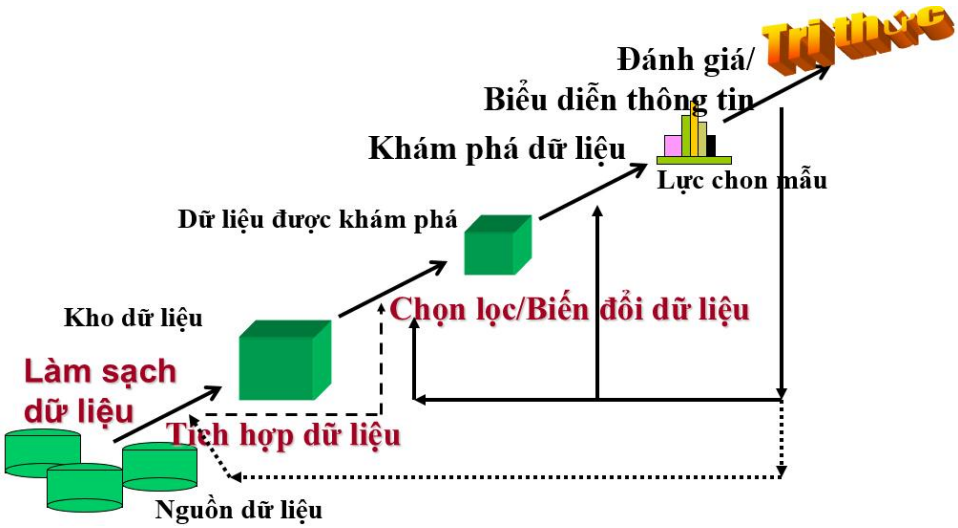
1.2.4 Những thách thức của bài toán

Bài toán khám phá quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học là một bài toán hết sức thiết thực. Tuy nhiên việc giải quyết bài toán cũng gặp nhiều khó khăn thách thức. Điển hình như: việc thu thập dữ liệu rất phức tạp và khó khăn. Đặc biệt là trong hoàn cảnh hiện tại, khi vấn đề bảo mật thông tin và quyền riêng tư của người dùng mạng xã hội đang được thắt chặt, dẫn đến việc truy xuất thông tin bị hạn chế.

1.3 Khai phá dữ liệu và biểu diễn dữ liệu văn bản.

1.3.1 Giới thiệu về khai phá dữ liệu (datamining)

Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD-ROM, băng từ, .v.v.) không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số lượng cũng như kích cỡ của các cơ sở dữ liệu (CSDL) cũng tăng lên một cách nhanh chóng. Nói một cách hình ảnh là chúng ta đang "ngập" trong dữ liệu nhưng lại "đói" tri thức. Data Mining như là một công nghệ tri thức giúp khai thác những thông tin hữu ích từ những kho dữ liệu được tích trữ trong suốt quá trình hoạt động của một công ty, tổ chức nào đó.



Hình 1.8. Các bước trong Data Mining & KDD

1.3.2 Khai phá dữ liệu văn bản

Hiện nay, cơ sở dữ liệu văn bản (text database) đang phát triển nhanh chóng và thu hút sự quan tâm nghiên cứu bởi sự gia tăng nhanh chóng số lượng thông tin ở dạng số, ví dụ như các tài liệu điện tử, email, thư điện tử, cá trang web…. Có thể thấy hầu hết thông tin của các chính phủ, các ngành công nghiệp, kinh doanh, trường học… đều được số hóa và lưu trữ ở dạng cơ sở dữ liệu này. Dữ liệu lưu trữ trong cơ sở dữ liệu văn bản là dữ liệu bán cấu trúc, tức là chúng không hoàn toàn phi cấu trúc nhưng cũng không hoàn toàn có cấu trúc. Ví dụ, một tài liệu có thể chứ một vài trường có cấu trúc chẳng hạn tiêu đề, tên tác giả, ngày xuất bản, phân loại… nhưng cũng có thể chứa một lượng lớn những thành phần văn bản phí cấu trúc như phần tóm tắt hay nội dung của tài liệu. Do đó vấn đề đặt ra là làm sao để có thể tìm kiếm và khai thác tri thứ từ những nguồn dữ liệu như vậy. Các kỹ thuật để giải quyết vấn đề này được gọi là kỹ thuật "Text Mining" hay khai phá dữ liệu văn bản.

Khai phá văn bản chia thành các vấn đề nhỏ hơn bao gồm phân loại văn bản (text categorization), gom cụm văn bản (text clustering), rút trích thực thể (entity extraction), phân tích quan điểm (sentiment analysis), tóm tắt tài liệu (document summarization), và mô hình hóa quan hệ giữa các thực thể (entity relation modeling).

Tìm kiếm văn bản

a. Nội dung

b. Quá trình

Phân loại văn bản

a. Nội dung

b. Quá trình

1.3.3 Mô hình biểu diễn dữ liệu văn bản

1.3.3.1 Tiền xử lý văn bản

Trước khi bắt đầu quá trình biểu diễn văn bản, người ta tiến hành bước tiền xử lý văn bản. Đây là bước hết sức quan trọng vì nó có nhiệm vụ làm giảm số từ có trong biểu diễn văn bản và qua đó sẽ làm giảm kích thước dữ liệu trong biểu diễn văn bản.

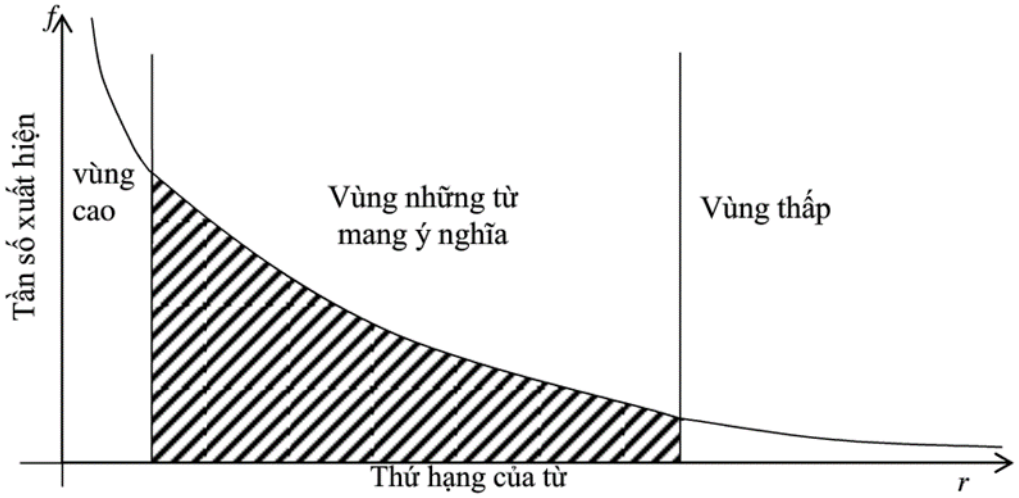
Loại bỏ StopWords

Bảng 1.1. Các từ dừng (stopwords) trong tiếng việt

nhận	rằng	cao	nhà	quá	riêng	gì	muốn
rồi	số	thấy	hay	lên	lần	nào	qua
bằng	điều	biết	lớn	khác	vừa	nếu	thời gian
họ	từng	đây	tháng	trước	chính	cả	việc
chưa	do	nói	ra	nên	đều	đi	tới

tôi	có thể	cùng	vì	làm	lại	mới	ngày
đó	vẫn	mình	chỉ	thì	đang	còn	bị
mà	năm	nhất	hơn	sau	ông	rất	anh
phải	như	trên	tại	theo	khi	nhưng	vào
đến	nhiều	người	từ	sẽ	ở	cũng	không
về	để	này	những	một	các	cho	được
với	có	trong	đã	là	và	của	thực sự
ở trên	tất cả	dưới	hầu hết	luôn	giữa	bất kỳ	hỏi
bạn	cô	tôi	tớ	cậu	bác	chú	dì
thím	cậu	mợ	ông	bà	em	thường	ai
cảm ơn							

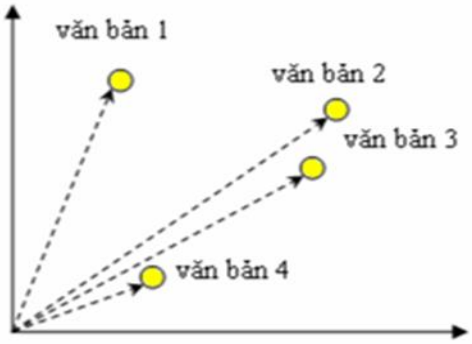
Loại bỏ những từ có tần số xuất hiện thấp



Hình 1.9. Lược đồ thống kê tần số của từ theo định luật Zipf

1.3.3.2 Mô hình không gian vector

Vector space model (Mô hình không gian vector) [6] là một mô hình đại số (algebraic model) thể hiện thông tin văn bản như một vector, các phần tử của vector này thể hiện mức độ quan trọng của một từ và cả sự xuất hiện hay không xuất hiện của nó trong một tài liệu.



Hình 1.10. Biểu diễn các vector văn bản trong không gian 2 chiều

1.3.3.3 Mô hình Boolean

Một mô hình biểu diễn vector với hàm f cho ra giá trị rời rạc với duy nhất hai giá trị đúng và sai (true và false, hoặc 0 và 1) gọi là mô hình Boolean. Hàm f tương ứng với từ khóa t_i sẽ cho ra giá trị đúng nếu và chỉ nếu từ khóa t_i xuất hiện trong văn bản đó.

Mô hình Boolean được xác định như sau:

Giả sử có một cơ sở dữ liệu gồm ***m*** văn bản, ***D = {d₁, d₂,… dₘ}***. Mỗi văn bản được biểu diễn dưới dạng một vector gồm ***n*** từ khóa ***T = {t₁, t₂,…tₙ}***. Gọi ***W = {wᵢⱼ}*** là ma trận trọng số, trong đó ***wᵢⱼ*** là giá trị trọng số của từ khóa ***tᵢ*** trong văn bản ***dⱼ.***

$$W_{ij} = \begin{cases} 1 & \text{nếu } t_i \text{ có mặt trong } d_j \\ 0 & \text{nếu ngược lại} \end{cases}$$

1.3.3.3 Mô hình N-Gram

N-gram được hiểu đơn giản là tần suất xuất hiện của n có thể là âm tiết, chữ cái hoặc từ vựng… liên tiếp xuất hiện trong dữ liệu. Kích thước của một n-grams được gọi là bậc của n-grams chính là số phần tử chứa trong nó. Một số mô hình n-gram phổ biến: unigram mô hình với n=1; bigram với n=2, là mô hình được sử dụng nhiều trong việc phân tích các hình thái cho ngôn ngữ; trigram với n-3, với n càng lớn thì độ chính xác càng cao tuy nhiên đi kèm với đó thì độ phức tạp cũng lớn hơn.

1.4 Kết luận

Trong chương này, tác giả đã giới thiệu về mạng xã hội và các ứng dụng nổi bật của mạng xã hội. Phát biểu về bài toán nghiên cứu quan tâm người dùng dựa vào bài viết trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học và những khó khăn thách thức cũng như ý nghĩa của bài toán. Nội dung của chương còn trình bày các vấn đề liên quan trực tiếp tới bài toán gồm một số vấn đề về khai phá dữ liệu và biểu diễn dữ liệu văn bản.

Chương 2: MÔ HÌNH GIẢI QUYẾT BÀI TOÁN

2.1 Xác định đặc trưng

2.1.1 Tầm quan trọng của Xác định đặc trưng

Các thuộc tính trong tập dữ liệu ảnh hưởng trực tiếp đến mô hình dự đoán, do đó ta cần xác định tốt cấu trúc của các thuộc tính sao cho diễn đạt hiệu quả nhất bản chất của tập dữ liệu.

2.1.2 Một số ví dụ về Xác định đặc trưng

Trực tiếp lấy dữ liệu thô

Lựa chọn đặc trưng

Giảm kích thước

Túi từ

Phương pháp giúp đưa các từ, các câu, đoạn văn ở dạng text trong các văn bản về một vector mà mỗi phần tử là một số.

2.2 Mô hình túi từ

Túi từ (Bag of Words) là một thuật toán hỗ trợ xử lý ngôn ngữ tự nhiên và mục đích của BoW là phân loại text hay văn bản. Ý tưởng của BoW là phân tích và phân nhóm dựa theo "Bag of Words"(corpus). Với test data mới, tiến hành tìm ra số lần từng từ của test data xuất hiện trong "bag". Tuy nhiên BoW vẫn tồn tại khuyết điểm, nên TF-IDF là phương pháp khắc phục.

2.2.1 Túi từ

Bag of word model (BoW) là mô hình được sử dụng trong xử lí ngôn ngữ tự nhiên giúp chúng ta lọc và tìm kiếm các từ quan trọng trong một đoạn văn bản bất kì, từ đó có thể đưa ra đặc trưng và giá trị của nó trong đoạn văn bản đó.

Mỗi từ được tương ứng với 1 chiều trong không gian dữ liệu, mỗi văn bản sẽ trở thành một vector nhiều chiều, mỗi chiều có giá trị không âm. Giá trị của mỗi từ được tính bằng tần suất xuất hiện của từ đó trong văn bản.

2.2.2 Phương pháp Phương pháp Tần số xuất hiện từ - Tần số văn bản nghịch đảo (TF-IDF)

TF-IDF: Giúp thống kê các từ các đoạn từ trọng đoạn văn bản (hay trong các trường của dữ liệu trong dữ liệu của bài này).

(TF) (Term frequency) là tần số xuất hiện của một từ. Số lần xuất hiện của từ đó so với số lần của từ xuất hiện nhiều nhất, giá trị trong khoảng từ [0,1]

Công thức tính:

$$\text{tf(t,d)} = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

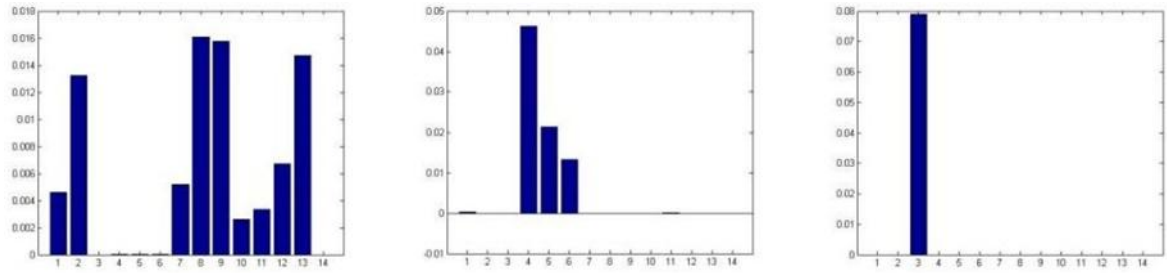
f (t,d): Số lần xuất hiện của từ t trong đoạn d

$\max\{f(w,d) : w \in d\}$: Số lần xuất hiện nhiều nhất của 1 từ bất kì trong văn bản

IDF (Inverse document frequency): Tấn số nghịch của 1 từ trong tập văn bản.

Tính IDF để giảm giá trị của những từ phổ biến. Mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản.

Hình ảnh minh họa 3 thành phần của SDV lấy được từ ma trận trọng số:



Hình 2.1. Ba thành phần của SVD

2.3 Một số thuật toán học có giám sát

Có nhiều thuật toán cho phân lớp như Naïve Bayes, K - láng giềng gần nhất, K-means, cây quyết định (Decision Tree), Máy vector hỗ trợ (Support Vector Machine), Mạng lọc thưa (Sparse Network of Winnows - SNoW), Mô hình Entropy cực đại,… Trong khuôn khổ luận văn, tác giả giới thiệu hai thuật toán học có giám sát là: Naïve Bayes, Máy vector hỗ trợ. Đây cũng là hai thuật toán sẽ tiến hành chạy thử nghiệm cho bài toán đang tìm hiểu tại chương 3.

2.3.1 Thuật toán Naïve Bayes

Naïve Bayes (NB) [10] là phương pháp phân loại có giám sát dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học.

Huấn luyện: tính $P(C_i)$ và $P(x_k|C_i)$

Các bước của giai đoạn huấn luyện được trình bày trong thuật toán 1:

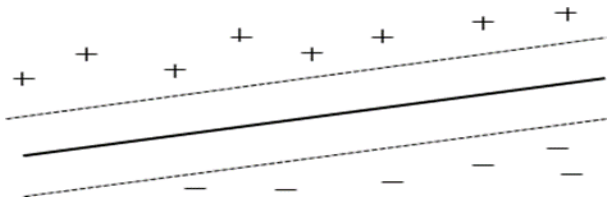
Input: D – tập DL training, C_i – phân lớp i
Output: $P(C_i)$ và $P(x_k|C_i)$.
- Đọc tập DL training
- Đọc C_i
- Với mỗi $c_i \in C$
- $P(C_i) \leftarrow \frac{|docs_i|+1}{|total\ docs|+m}$
- Với mỗi x_k trong phân lớp i
- $d_k \leftarrow$ số giá trị có thể có của đặc trưng thứ k
- $P(x_k|C_i) \leftarrow \frac{|docs_{x_k i}|+1}{|docs_i|+d_k}$
- Kết thúc
- Kết thúc

|docsi|: số văn bản của tập huấn luyện thuộc phân lớp i.
|total docs|: số văn bản trong tập huấn luyện.
m: số phân lớp
$|docs_{x_k i}|$: Số văn bản trong trong phân lớp i có đặc trưng thứ k mang giá trị xk. (hay số văn bản trong lớp i, có xuất hiện/không xuất hiện đặc trưng k)
$|docs_i|$: Số văn bản của tập huấn luyện thuộc phân lớp i.

2.3.2 Thuật toán vector hỗ trợ

Thuật toán máy vector hỗ trợ (Support Vector Machines - SVM) [10] được Corters và Vapnik giới thiệu vào năm 1995. SVM rất hiệu quả để giải quyết các bài toán với dữ liệu có số chiều lớn như các vector biểu diễn văn bản. Thuật toán SVM ban đầu chỉ được thiết kế để giải quyết bài toán phân lớp nhị phân tức là số lớp hạn chế là hai lớp. Hiện nay, SVM được đánh giá là bộ phân lớp chính xác nhất cho bài toán phân lớp văn bản, bởi vì đó là bộ phân lớp tốc độ rất nhanh và hiệu quả đối với bài toán phân lớp văn bản.

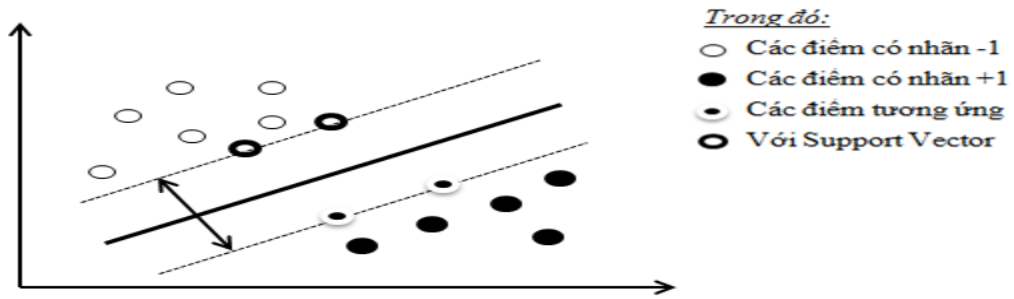
Mục đích của phương pháp SVM là tìm được khoảng cách biên lớn nhất, điều này được minh họa như sau:



Hình 2.2. Hình Siêu phẳng phân chia dữ liệu học thành 2 lớp + và – với khoảng cách biên lớn nhất

Bài toán phân 2 lớp với SVM

Bài toán đặt ra là: Xác định hàm phân lớp để phân lớp các mẫu trong tương lai, nghĩa là với một mẫu dữ liệu mới x_i thì cần phải xác định x_i được phân vào lớp +1 hay lớp -1



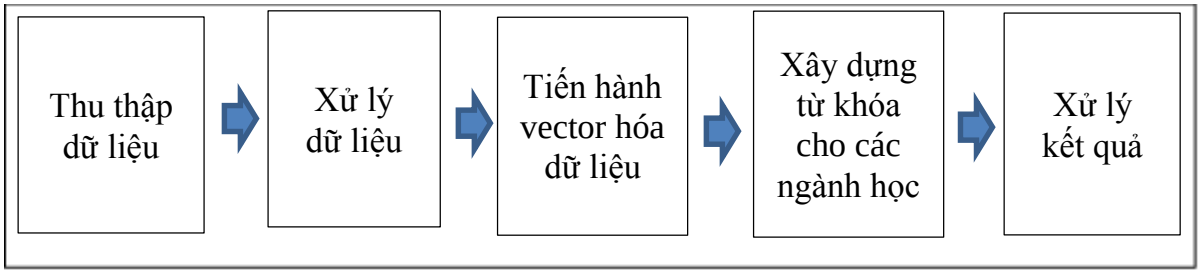
Hình 2.3. Minh họa bài toán phân 2 lớp bằng phương pháp SVM

Bài toán nhiều phân lớp với SVM

Các bước chính của phương pháp SVM

2.4 Phương pháp khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học dựa trên xử lý ngôn ngữ tự nhiên.

Sau quá trình nghiên cứu, tác giả đưa ra phương pháp khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học như sau:



Bước 1: Thu thập dữ liệu: tiến hành thu thập dữ liệu là các bài viết của người dùng mạng xã hội.

Bước 2: Xử lý dữ liệu: tiến hành tách từ, chuẩn hóa từ, loại bỏ từ dừng.

Bước 3: Tiến hành vector hóa dữ liệu bằng thuật toán túi từ - bag of words.

Bước 4: Xây dựng từ khóa về các ngành học: dựa trên hệ thống từ khoá các ngành học, đào tạo được cung cấp bởi Phòng Đào tạo – ĐH Kiến trúc Hà Nội.

Bước 5: Xử lý kết quả sau phân loại: Đánh giá độ lớn vector đại diện của các từ khóa.

2.5 Kết luận

Trong chương này luận văn đã trình bày tổng quan về học máy, hai thuật toán học máy có giám sát là Naïve Bayes và Support Vector Machines - SVM, kỹ thuật Xác định đặc trưng với mô hình Túi từ được sử dụng trong việc giải quyết bài toán đang tìm hiểu. Đồng thời, đưa ra phương pháp khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

Chương 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Dựa trên cơ sở lý thuyết đã trình bày ở chương 1 và chương 2, chương 3 sẽ mô tả chi tiết về bài toán khám phá quan tâm của người dùng trên mạng xã hội, dữ liệu, các phần mềm và công cụ sử dụng, một số kết quả thử nghiệm và đánh giá.

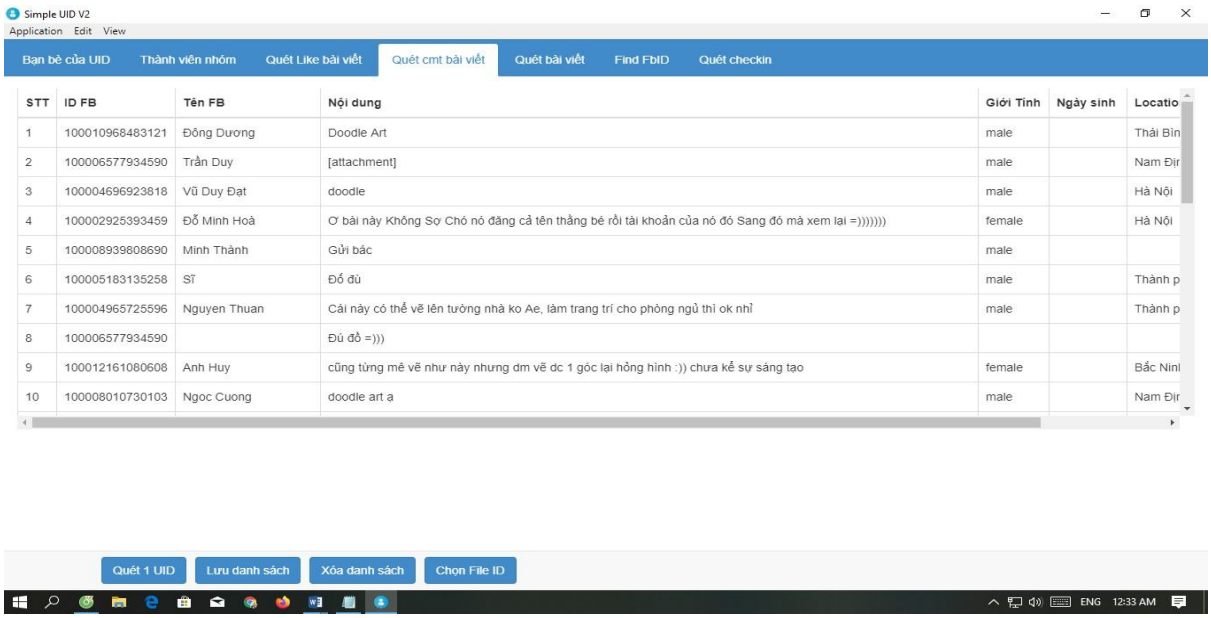
3.1 Phát biểu bài toán

Bài toán khám phá quan tâm của người dùng trên mạng xã hội Facebook dựa vào bài viết Tiếng Việt, vấn đề làm thế nào để biết được mối quan tâm của người dùng tới ngành học của một trường đại học thông qua việc phân tích các từ rút trích ra được từ những bài viết Tiếng Việt do họ tạo ra. Bài toán được phát biểu như sau:

3.2 Dữ liệu

3.2.1 Thu thập dữ liệu

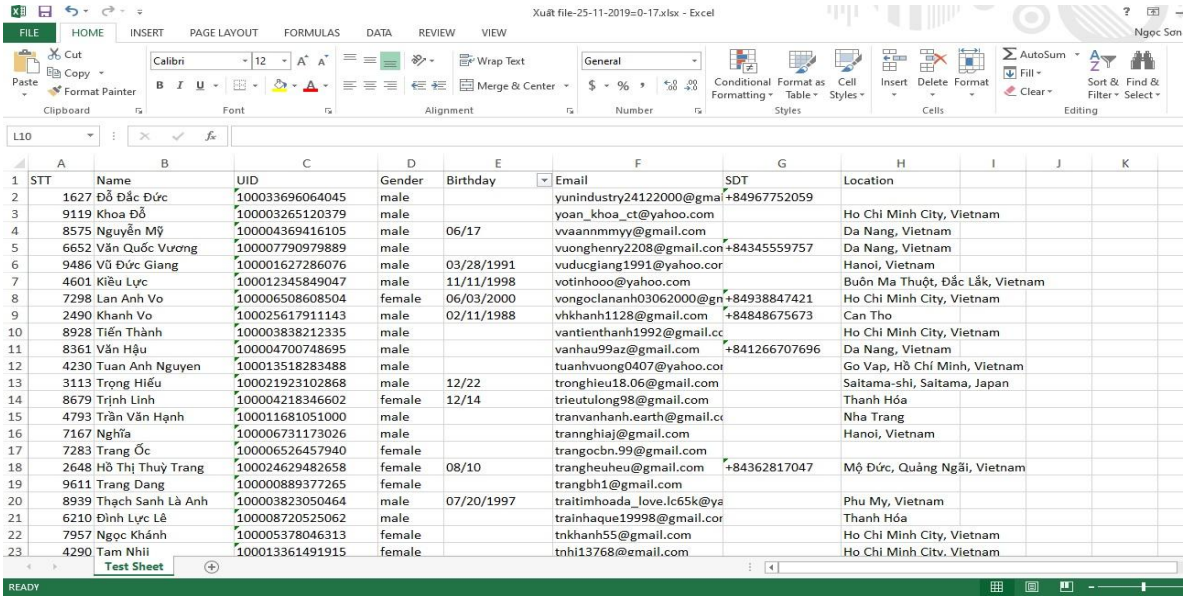
Hiện nay việc thu thập dữ liệu trên mạng xã hội Facebook khá là khó khăn; một phần là do vấn đề bảo mật dữ liệu, mặt khác là vì chúng không được tổng hợp để sẵn có như một số dịch vụ mạng xã hội khác. Ở đây, tác giả đã sử dụng công cụ Simple UID được chia sẻ miễn phí theo địa chỉ: https://atpsoftware.vn/simple-uid



Hình 3.1. Giao diện phần mềm Simple UID

3.2.2 Mô tả dữ liệu

Dữ liệu thử nghiệm của luận văn gồm hàng trăm bài viết từ người dùng trên mạng xã hội Facebook. Mỗi người dùng có một ID và có các trường thông tin sau: tên, địa chỉ, giới tính, ngày sinh, bài viết.



STT	Name	UID	Gender	Birthday	Email	SDT	Location
1627	Đỗ Đắc Đức	100033696064045	male		yunindustry24122000@gmai	+84967752059	
9119	Khoa Đỗ	100003265120379	male		yoan_khoa_ct@yahoo.com		Ho Chi Minh City, Vietnam
8575	Nguyễn Mỹ	100004369416105	male	06/17	vvaannmmyy@gmail.com		Da Nang, Vietnam
6652	Văn Quốc Vương	100007790979889	male		vuonghenry2208@gmail.com	+84345559757	Da Nang, Vietnam
9486	Vũ Đức Giang	100001627286076	male	03/28/1991	vuducgiang1991@yahoo.cor		Hanoi, Vietnam
4601	Kiều Lực	100012345849047	male	11/11/1998	votinhooo@yahoo.com		Buôn Ma Thuột, Đắc Lắk, Vietnam
7298	Lan Anh Vo	100006508608504	female	06/03/2000	vongoclananh03062000@gn	+84938847421	Ho Chi Minh City, Vietnam
2490	Khanh Vo	100025617911143	male	02/11/1988	vhkhanh1128@gmail.com	+84848675673	Can Tho
8928	Tiến Thành	100003838212335	male		vantienthanh1992@gmail.cc		Ho Chi Minh City, Vietnam
8361	Văn Hậu	100004700748695	male		vanhau99az@gmail.com	+841266707696	Da Nang, Vietnam
4230	Tuan Anh Nguyen	100013518283488	male		tuanhvuong0407@yahoo.cor		Go Vap, Hồ Chí Minh, Vietnam
3113	Trọng Hiếu	100021923102868	male	12/22	tronghieu18.06@gmail.com		Saitama-shi, Saitama, Japan
8679	Trịnh Linh	100004218346602	female	12/14	trieutulong98@gmail.com		Thanh Hóa
4793	Trần Văn Hạnh	100011681051000	male		tranvanhanh.earth@gmail.c		Nha Trang
7167	Nghĩa	100006731173026	male		trannghiaj@gmail.com		Hanoi, Vietnam
7283	Trang Ốc	100006526457940	female		trangocbn.99@gmail.com		
2648	Hồ Thị Thuỳ Trang	100024629482658	female	08/10	trangheuheu@gmail.com	+84362817047	Mộ Đức, Quảng Ngãi, Vietnam
9611	Trang Dang	100000889377265	female		trangbh1@gmail.com		
8939	Thạch Sanh Là Anh	100003823050464	male	07/20/1997	traitimhoada_love.lc65k@ya		Phu My, Vietnam
6210	Đinh Lực Lê	100008720525062	male		trainhaque19998@gmail.cor		Thanh Hóa
7957	Ngọc Khánh	100005378046313	female		tnkhanh55@gmail.com		Ho Chi Minh City, Vietnam
4290	Tam Nhii	100013361491915	female		tnhi13768@gmail.com		Ho Chi Minh City, Vietnam

Hình 3.2. Dữ liệu người dùng thu được từ Facebook

Luận văn sẽ sử dụng dữ liệu bài viết thu thập của người dùng tại khu vực Hà Nội, từ 22 tuổi trở xuống. Các dữ liệu từ tất cả các bài viết của người dùng sẽ được lưu tại một file là "dulieumxh.txt"

Hình 3.3. Dữ liệu bài viết của một người dùng Facebook

3.3 Phần mềm và các công cụ sử dụng

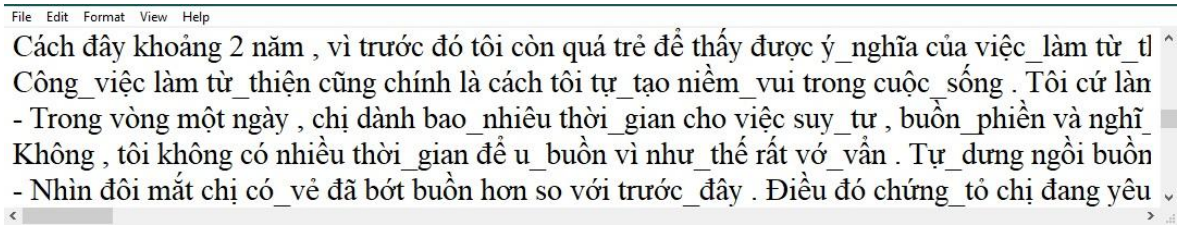
Hệ thống cài đặt chương trình thử nghiệm được thể hiện trong bảng sau:

Bảng 3.1 Môi trường thử nghiệm

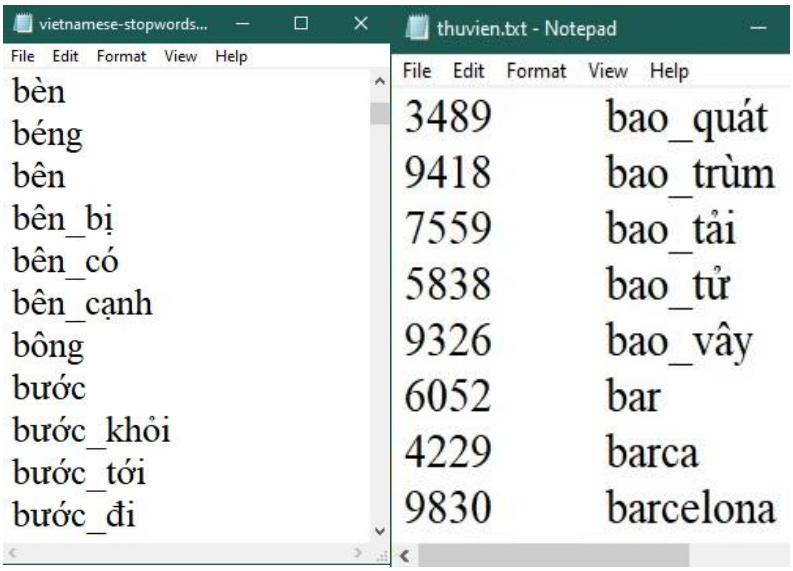
Thành phần	**Thông số**
Hệ điều hành	Windows 10 Pro 64bit
Bộ vi xử lý	Intel Core i3-3220 3.3GHz
RAM	8Gb
Ổ cứng	500Gb

3.4 Xử lý dữ liệu

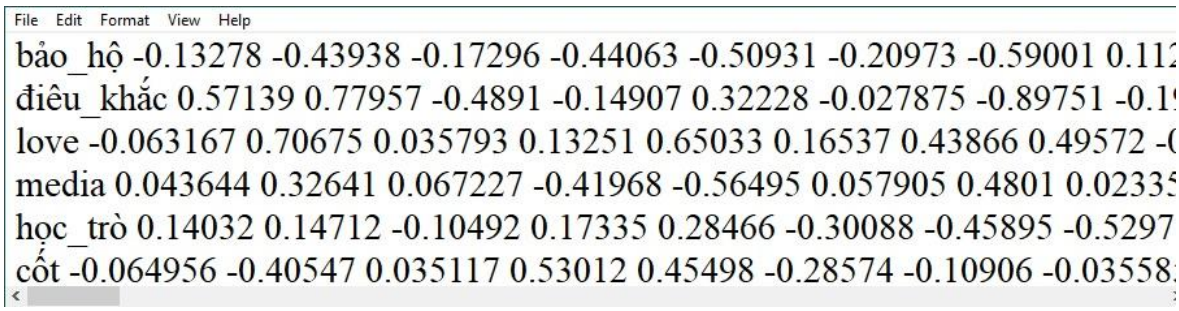
Dưới đây tác giả sẽ trình bày chi tiết cách xử lý dữ liệu cho mô hình giải quyết bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.



Hình 3.4. Dữ liệu sau quá trình tách từ



Hình 3.5. File stopwords và từ điển



Hình 3.6. Dữ liệu thu được sau khi vector hóa

3.5 Kết quả thử nghiệm và đánh giá

Luận văn tiến hành thử nghiệm mô hình Bag of Words (BoW) trên bộ dữ liệu thu được từ mạng xã hội Facebook. Đồng thời đánh giá kết quả và áp dụng vào thực tiễn công tác truyền thông tuyển sinh của Trường Đại học Kiến trúc Hà Nội.

Sau quá trình Xử lý dữ liệu như đã nêu ở trên, tác giả đã thu được thông số cho từ khóa của từng ngành học.

Ngành Xây dựng											
vật_liệu	0.11825	-0.22825	-0.41305	0.11068	-0.04938	-0.35523	-0.27335	-0.21385	-0.23932	0.5645	0.08191
công_trình	0.024897	0.42179	-0.09055	-0.23157	0.16239	-0.43276	-0.42523	-0.12106	0.050621	0.16321	0.28918
kết_cấu	0.2147	-0.20546	-0.19028	0.60609	0.053737	-0.5029	-0.49851	-0.3593	0.12243	-0.0487	0.37172
thi_công	-0.32523	0.10789	0.33556	-0.17188	0.2692	-0.29029	-0.09969	-0.21028	-0.04171	0.21429	-0.38031
xây_dựng	0.019169	0.13665	0.17505	-0.23332	0.084166	0.11793	-0.32035	0.14137	-0.02936	0.2019	0.075332
Ngành Kiến trúc											
kiến_trúc	0.63067	0.39237	-0.11765	0.24409	-0.16288	-0.65607	-0.36931	0.016235	-0.08975	0.13193	0.027729
quy_hoạch	-0.01672	-0.24207	0.41791	-0.33353	-0.13357	-0.04928	-0.29123	-0.02221	-0.11835	-0.22239	-0.40372
thiết_kế	0.20904	0.008976	0.091165	-0.02448	0.50396	-0.00098	-0.19158	-0.52637	-0.12885	0.34529	0.33582
cấu_trúc	0.29731	0.15952	-0.19508	0.38325	-0.37031	-0.28364	-0.33279	-0.26296	-0.07663	-0.214	0.096339
cảnh_quan	-0.03962	-0.08081	-0.36495	-0.18348	-0.34059	-0.66444	-0.52442	0.087442	0.36015	0.19057	-0.21567
Ngành Hạ tầng Đô thị											
môi_trường	0.005382	-0.1181	0.005521	-0.05279	-0.38462	-0.56886	-0.21188	-0.12375	0.080534	0.048914	0.047379
cấp_nước											
thoát_nước											
giao_thông	-0.13601	-0.38686	0.48404	0.055456	0.051649	-0.16636	-0.17199	0.3091	-0.57228	-0.01076	-0.07962
hạ_tầng	0.10422	-0.14114	0.24324	0.33448	-0.25958	-0.35476	-0.04816	0.006758	-0.32539	0.005047	0.38408
Ngành Mỹ thuật - Nội thất											
mỹ_thuật	0.69189	0.3583	-0.25009	-0.2782	0.15529	-0.69523	-0.70028	-0.14766	-0.35032	0.48532	0.30353
nghệ_thuật	0.60611	0.24823	-0.31098	0.087602	-0.13693	-0.31117	-0.35657	-0.02013	-0.12537	0.040088	0.49566
đồ_họa	0.15245	0.18944	0.49924	0.12226	-0.28422	0.26981	-0.8734	-0.82665	-0.40272	0.01173	0.488
điêu_khắc	0.57139	0.77957	-0.4891	-0.14907	0.32228	-0.02788	-0.89751	-0.19313	0.07891	0.57056	0.19281
hội_họa	0.70376	0.36276	-0.53892	0.43506	-0.03438	-0.43883	-0.71	-0.40511	-0.20867	0.24452	0.040738

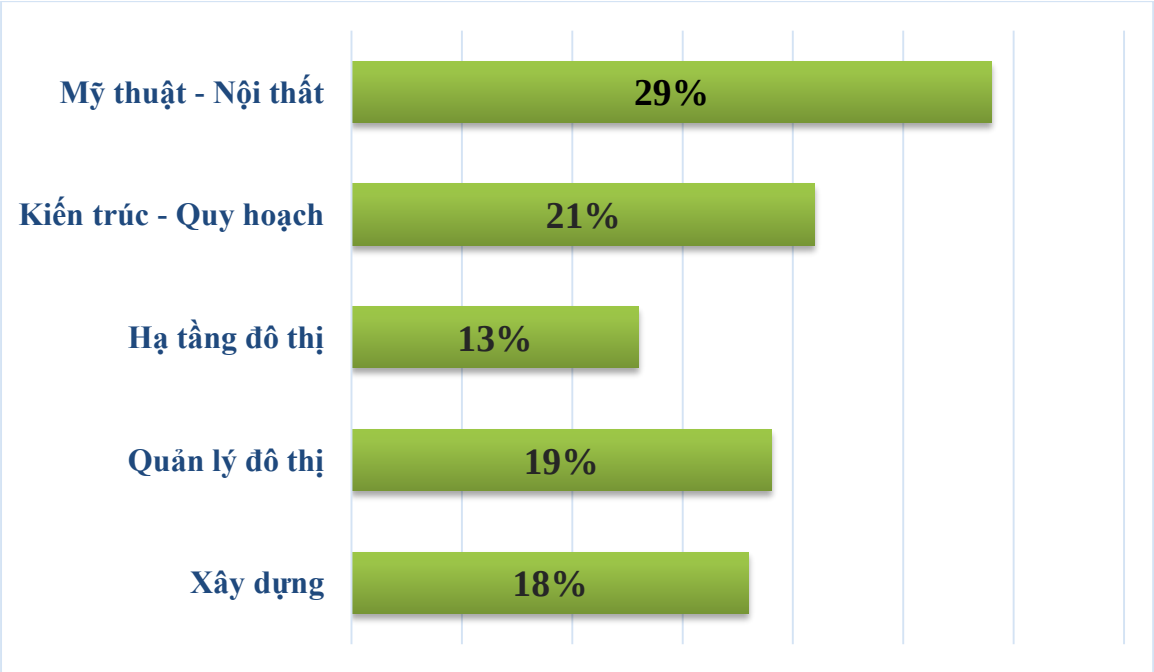
Hình 3.7. Kết quả thu được với từ khóa của từng ngành học

Từ kết quả sau quá trình vector hóa văn bản, tác giả có thể biết được độ lớn của vector biểu diễn các từ khoá từ đó có thể tính toán được tỉ lệ tương quan sự quan tâm của người dùng giữa các ngành học. Từ khóa cấp_nước và thoát_nước không thu được kết quả do không xuất hiện trong bộ cơ sở dữ liệu. Độ lớn vector đại diện cho các từ khoá như sau:

Bảng 3.2 Bảng độ lớn vector của các từ khóa thuộc ngành học

Ngành học	Từ khóa	Độ lớn vector
Xây dựng	vật_liệu	3.31786
	công_trình	2.77719
	kết_cấu	3.237185
	thi_công	3.109306
	xây_dựng	2.462262
Kiến trúc – Quy hoạch	kiến_trúc	3.346985
	quy_hoạch	3.622615
	thiết_kế	2.846713
	cấu_trúc	2.888134
	cảnh_quan	3.343251
Kỹ thuật Hạ tầng đô thị	môi_trường	3.070959
	cấp_nước	0.0
	thoát_nước	0.0
	giao_thông	3.383439
	hạ_tầng	3.632892
Mỹ thuật – Nội thất	mỹ_thuật	4.282763
	đồ_họa	4.183288
	điêu_khắc	3.427316
	hội_họa	3.707231
	thời_trang	3.906045
Quản lý Đô thị	kinh_tế	3.048389
	quản_lý	2.621017
	môi_giới	3.359173
	đầu_tư	3.468024
	kinh_doanh	2.848544

Như đã trình bày ở trên vector đại diện cho từ khoá của ngành học càng lớn thì cường độ xuất hiện của từ càng nhiều và nó cũng là thước đo cho sự quan tâm của người dùng tới từ khoá, ngành học đó. Vì vậy từ độ lớn vector thuộc tính của từ khóa, tác giả có thể biết được tỉ lệ sự quan tâm của người dung tới các ngành đào tạo như hình dưới.



Hình 3.8. Biểu đồ tỉ lệ sự quan tâm của người dùng tới các ngành học

3.6 Kết luận

Chương 3 của luận văn đã trình bày về phương pháp lấy dữ liệu bài viết của người dùng trên mạng xã hội Facebook, cách xử lý dữ liệu. Kết quả thử nghiệm phương pháp xử lý ngôn ngữ tự nhiên với mô hình Túi từ, với kết quả thu được đã cho biết tỉ lệ quan tâm của người dùng mạng xã hội tới các ngành học của trường ĐH Kiến trúc Hà Nội. Kết quả này cũng tương đồng với thực tế ghi nhận được tại Trường Đại học Kiến trúc những năm gần đây khi các ngành liên quan đến nghệ thuật đang rất được đón nhận và có số lượng thí sinh dự thi tăng đột biến.

KẾT LUẬN

1. Những kết quả đạt được:

Với mục tiêu nghiên cứu đề ra, luận văn đã đi sâu nghiên cứu các vấn đề xung quanh bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học và đã đạt được những kết quả sau:

- Khảo sát một số thuật toán học có giám sát và các vấn đề về biểu diễn và xử lý dữ liệu văn bản.
- Phát biểu và xây dựng mô hình xử lý bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học và triển khai giải quyết bài toán theo mô hình.
- Xây dựng bộ từ khóa tương ứng của các ngành học của Trường ĐH Kiến trúc Hà Nội áp dụng trong chương trình thử nghiệm. Tiến hành thử nghiệm với mô hình xử lý ngôn ngữ tự nhiên, đưa ra kết quả phân loại quan tâm của người dùng đối với các ngành học này và áp dụng vào công tác truyền thông tuyển sinh của Trường Đại học Kiến trúc sẽ triển khai tập trung mạnh vào việc thông tin, quảng bá về hai ngành học Mỹ thuật – Nội thất và Kiến trúc – Quy hoạch tại địa bàn thành phố Hà Nội.

2. Hướng phát triển luận văn:

Trong quá trình thực hiện luận văn, không tránh khỏi có một số hạn chế do điều kiện về mặt thời gian và trình độ của học viên. Vì vậy, hướng nghiên cứu tiếp theo của học viên là:

- Ngoài việc sử dụng bài viết, luận văn có thể hướng đến việc sử dụng những thông tin khác mà người dùng chia sẻ trên mạng xã hội để khám phá vấn đề mà họ quan tâm. Cũng như, không chỉ dừng lại ở mạng xã hội Facebook mà còn có thể mở rộng cho các mạng xã hội phổ biến khác.
- Bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học có ứng dụng rất rộng rãi cho nhiều lĩnh vực. Kết quả bài toán sẽ là công cụ đắc lực để các cơ sở giáo dục có những điều chỉnh, định hướng công tác truyền thông sau đó là tác động đến quan điểm của người dùng. Do đó, luận văn có thể tiếp tục phát triển theo hướng một trong những ứng dụng của bài toán.