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

LUẬN VĂN THẠC SĨ KỸ THUẬT

(*Theo định hướng ứng dụng*)

HÀ NỘI - 2020

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***)

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS TRẦN ĐÌNH QUẾ

HÀ NỘI - 2020

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi.

Các số liệu, kết quả đưa ra trong luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tác giả luận văn

Đỗ Ngọc Sơn

LỜI CẢM ƠN

Để hoàn thành được luận văn, ngoài sự nghiên cứu và cố gắng của bản thân, tôi xin cảm ơn PGS.TS Trần Đình Quế - người thầy trực tiếp hướng dẫn, tận tình chỉ bảo và định hướng cho tôi trong suốt quá trình thực hiện luận văn. Xin được gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc của tôi tới thầy!

Tôi xin gửi lời cảm ơn chân thành tới tất cả các thầy cô giáo của Học viện Công nghệ Bưu chính Viễn thông đã giảng dạy, hướng dẫn và dìu dắt tôi trong trong suốt quá trình học tập tại trường.

Trong quá trình nghiên cứu và thực hiện luận văn, mặc dù được sự hướng dẫn nhiệt tình của thầy giáo PGS.TS Trần Đình Quế và những nỗ lực của bản thân tuy nhiên không thể tránh khỏi những hạn chế, thiếu sót. Tôi rất mong nhận được ý kiến đóng góp, sửa chữa từ quý Thầy, Cô và các bạn bè đồng nghiệp để luận văn được hoàn thiện hơn.

Trân trọng cảm ơn!

Tác giả

Đỗ Ngọc Sơn

MỤC LỤC

DANH MỤC TỪ VIẾT TẮT

Từ viết tắt	**Tiếng Anh**	**Tiếng Việt**
BOW	Bag of words	Túi từ
IDF	Inverse Document Frequency	Tần số văn bản nghịch đảo
KDD	Knowledge Discovery in Databases	Phát hiện tri thức trong cơ sở dữ liệu
KPDL	Data mining	Khai phá dữ liệu
MXH	Social network	Mạng xã hội
NB	Naïve Bayes	Naïve Bayes
SVD	Singular Value Decomposition	Phân rã giá trị số ít
SVM	Support Vector Machine	Máy vector hỗ trợ
TF	Term Frequency	Tần số xuất hiện từ

DANH MỤC BẢNG BIỂU

DANH MỤC HÌNH VẼ

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

Mạng xã hội với cách gọi đầy đủ là "dịch vụ mạng xã hội" (tiếng Anh là "social network service") hay "trang mạng xã hội", là nền tảng trực tuyến nơi mọi người dùng để xây dựng các mối quan hệ với người khác có chung tính cách, nghề nghiệp, công việc, trình độ,... hay có mối quan hệ ngoài đời thực.

Mạng xã hội có nhiều dạng thức và tính năng khác nhau, có thể được trang bị thêm nhiều công cụ mới, và có thể vận hành trên tất cả các nền tảng như máy tính để bàn, máy tính xách tay, máy tính bảng hay điện thoại thông minh.

Mạng xã hội cho phép người dùng chia sẻ câu chuyện, bài viết, ý tưởng cá nhân, đăng ảnh, video, đồng thời thông báo về hoạt động, sự kiện trên mạng hoặc trong thế giới thực. Nếu như trong mô hình mạng xã hội truyền thống, ví dụ như sự kiện hội chợ, đã tồn tại từ lâu trong lịch sử thì mạng xã hội trên web giúp người dùng kết nối với những người sống ở nhiều vùng đất khác nhau, ở thành phố khác hoặc trên toàn thế giới.

Mạng xã hội được chia làm hai đặc điểm chính đó là: Một là sự góp mặt của những chủ thể hoặc cá nhân. Hai là người dùng sẽ tự tạo ra nội dung của trang web những thành viên còn lại sẽ được xem thông tin của người dùng tạo nên. Tùy vào từng nền tảng mạng xã hội, nhưng thường các thành viên có thể liên hệ với bất kỳ thành viên nào khác. Trong một số trường hợp khác, các thành viên có thể liên hệ bất kỳ ai họ có mối liên hệ.

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.

Từ khi bắt đầu xuất hiện, mạng xã hội mang lại những lợi ích không nhỏ trong nhiều lĩnh vực, ngành nghề, trong liên hệ công việc, tuyển dụng, trao đổi, học hỏi, kinh doanh, mua bán, tương tác xã hội...

Mặc dù vậy trong quá trình vận hành thì các mạng xã hội cũng bộc lộ nhiều vấn đề bất cập liên quan đến spam, quyền riêng tư, thu thập thông tin, bảo mật, nguy cơ sử dụng sai mục đích, hay bảo vệ trẻ em...

1.1.3 Ứng dụng của Mạng xã hội

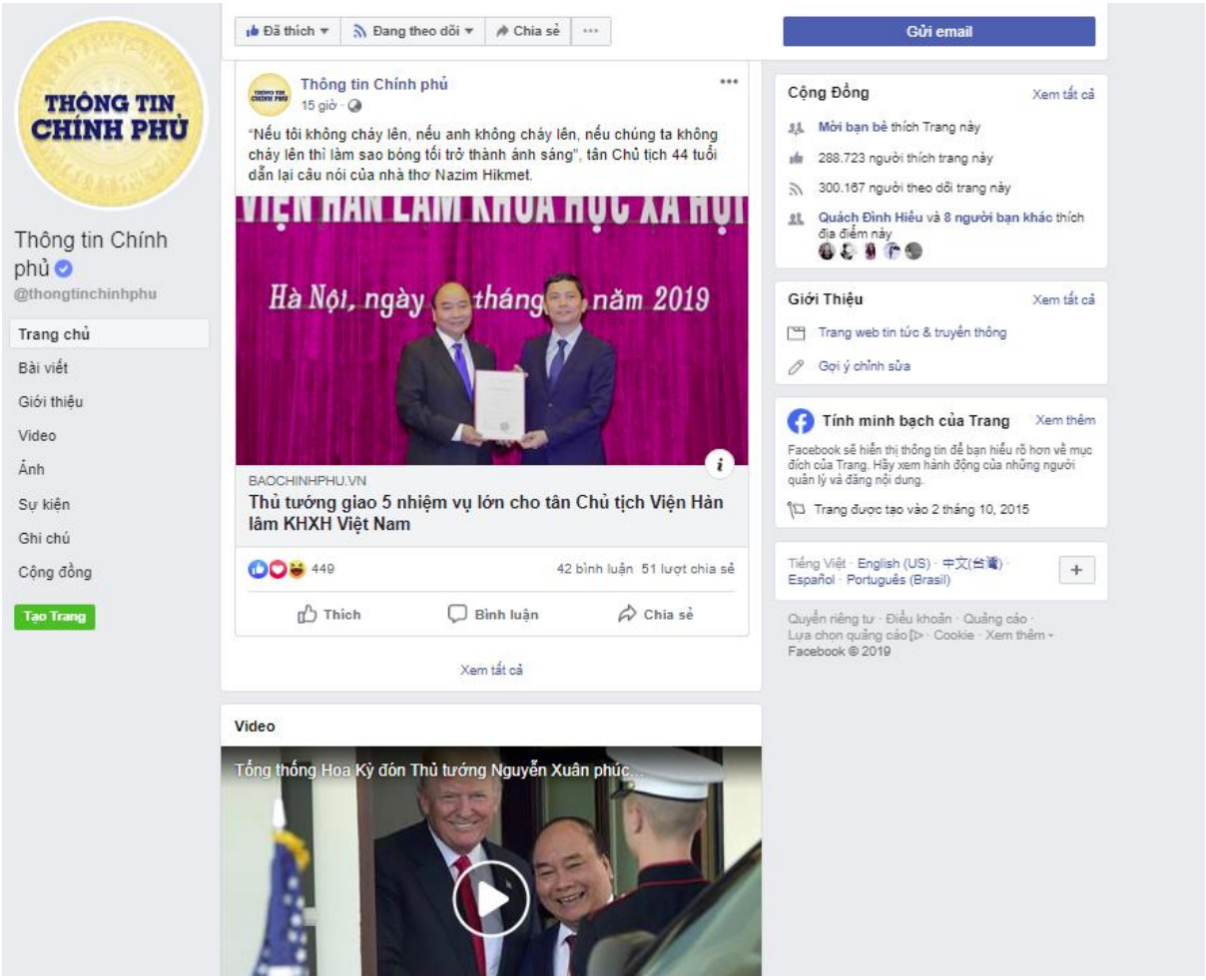
Sự tiếp cận đến từng cá nhân người dùng với tốc độ nhanh tạo ra nhiều cơ hội và lợi ích về truyền tải, tiếp nhận, chia sẻ, thông tin, tri thức; phục vụ các nhu cầu đa dạng của cộng đồng như: kết bạn, giải trí, kinh doanh, bày tỏ quan điểm, phản biện xã hội, lan tỏa những điều tốt đẹp…

Có rất nhiều những ứng dụng hiện nay của mạng xã hội có thể kể đến một số như:

Kênh cập nhật tin tức, kiến thức, xu thế: với tính năng like, theo dõi trang người đọc sẽ nhận được ngay những thông tin cập nhật của trang mạng mình yêu thích hoặc quan tâm về tất các thể loại và nhà cung cấp hay nhà quảng cáo cũng có thể nhanh chóng cập nhật xu thế mới nhất của lĩnh vực mình yêu thích. Người dùng cũng có thể xem tin tức hay cập nhật những bộ phim, những video clip nhạc mới nhất

trên youtube, facebook hay các trang social network khác rất nhanh chóng.

Trở thành một kênh báo chí, dịch vụ công: các cơ quan báo chí và thông tin đại chúng đang tích cực đăng tải cùng một lúc trên báo giấy, trên báo điện tử và trang mạng của mình để theo kịp xu thế của thời đại và giữ số lượng độc giả của mình. Các cơ quan pháp luật hay dịch vụ công cũng đang dần "lên sóng" social network để cập nhật những tin tức và quy định mới của mình hoặc lắng nghe ý kiến phê bình góp ý của người dân để tiến tới một bộ máy hành chính công thông minh và giản tiện hơn.



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

Kết nối bạn bè, gia đình, cộng đồng: Dù ở đâu, gần hay cách xa, một người thân hay bạn bè cập nhật thông tin, hình ảnh, trạng thái, cảm xúc… người dùng sẽ nhận được ngay. Việc chia sẻ thông tin cá nhân trên social network có thể giúp người tự giới thiệu về mình và tìm hiểu những người bạn mới nhanh chóng và dễ dàng hơn rất nhiều so với những cuộc gặp gỡ thông thường.

Kinh doanh và quảng cáo: nhờ vào sự kết nối mọi người cùng tốc độ truyền hình ảnh và thông tin qua Twitter, Zalo, Viber hay Facebook. Người dùng chỉ cần tạo một trang miễn phí, mời các người dùng khác tham gia và ghé thăm từ đó có thể thường xuyên cập nhật và quảng cáo các sản phẩm của mình, nhận các đơn đặt hàng và chăm sóc trực tuyến khách hàng. Và các mạng xã hội cũng đã tạo ra những công cụ giao diện riêng danh cho hoạt động kinh doanh và quảng cáo trên các nền tảng của mình. Không chỉ có người dùng đơn lẻ mà ngay cả những công ty tập đoàn lớn cũng rất tích cực sử dụng không gian mạng xã hội làm nơi quảng cáo, kinh doanh các sản phẩm của mình.

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

Ví dụ mạnh mẽ nhất của ứng dụng mạng xã hội đó là việc Tổng thống Mỹ Donald Trump đã sử dụng công cụ mạng xã hội để có thể bày tỏ quan điểm và tương tác với các cử chi Mỹ, đây nhân tố quyết định cho chiến dịch tranh cử cuối cùng là thắng cử Tổng thống của ông. Phương thức này của Tổng thống Trump đã định hình lại cách thức tranh cử của các chính trị gia trên khắp thế giới.

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]

Thời đại thông tin ngày nay đã tạo những điều kiện cho con người giao lưu, liên kết, chia sẻ những sở thích, sự quan tâm, những ý tưởng, những việc làm bằng các phương tiện truyền thông hiện đại – nhất là sự phát triển ngày càng đa dạng của internet, trong đó có các mạng xã hội. Mối quan tâm của người dùng trên mạng xã hội rất đa dạng. Nó có thể đơn giản là một quan tâm về một cá nhân nào đó, một vấn đề mang tính thời sự, một mặt hàng, một quan điểm, một bài báo, một đánh giá về vấn đề. Mỗi cá nhân, tổ chức hay còn gọi là thực thể sử dụng mạng xã hội đều có thể thể hiện mối quan tâm của mình trên mạng xã hội. Hay nói cách khác, khi một nút trên mạng xã hội có liên kết tới một nút khác thì sẽ được hiểu là nút đó có quan tâm đến đối tượng là một nút khác đó.

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

Có rất nhiều cách để người dùng thực hiện quan tâm đến một vấn đề nào đó trên mạng xã hội. Ví dụ họ có thể trực tiếp viết về vấn đề đó, chia sẻ quan điểm cá nhân, những gì họ cảm nhận. Hoặc họ có thể tương tác với vấn đề mình quan tâm thông qua việc thích và chia sẻ vấn đề đó với những người dùng khác (các nút khác)

trong mạng xã hội. Việc truy cập thường xuyên vào một nút nào đó trong mạng xã hội cũng là cách người dùng quan tâm vấn đề đó. Quan tâm của người dùng trên mạng xã hội đã tạo ra lượng dữ liệu rất lớn. Dữ liệu đó bao gồm người sử dụng mạng xã hội, vấn đề mà họ quan tâm, họ quan tâm vấn đề đó như thế nào…

Nghiên cứu phân tích mạng xã hội có thể kể đến phân tích về hành vi của người dùng như công trình nghiên cứu của PGS.TS. Trần Đình Quế và các cộng sự về đề tài "Ước tính sự giống nhau của người dùng mạng xã hội dựa trên hành vi" [8]. Nghiên cứu này trình bày một mô hình để ước tính sự giống nhau giữa người dùng dựa trên hành vi của họ trên mạng xã hội. Các hành vi được xem xét là các hoạt động bao gồm đăng bài, thích những mục này, bình luận và thích bình luận trong những bài này. Hay như phân tích dựa trên nội dung đưa lên như của nhóm tác giả Diana Palsetia và các cộng sự [5] đã mô hình hóa các mạng dựa trên mối quan tâm của người dùng bằng cách suy ra ý định từ các hoạt động truyền thông xã hội như bình luận và tweet của hàng triệu người dùng trong Facebook và Twitter, tương ứng.

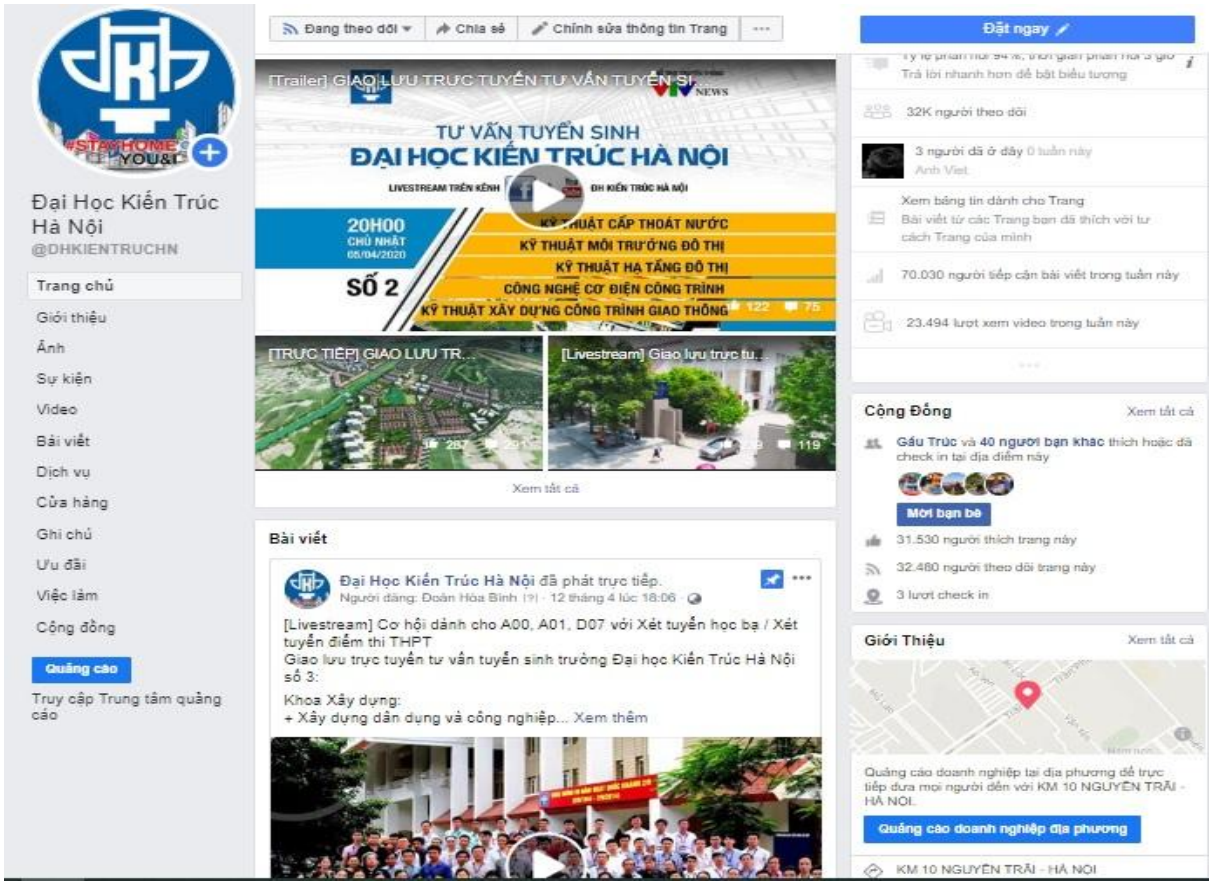
Luận văn sẽ dựa trên phân tích về nội dung đưa lên của người dùng mạng xã hội để trình bày bài toán nghiên cứu quan tâm người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

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.

Ở Việt Nam những năm gần đây các trường đại học đều đang rất tập trung và việc thu hút sinh viên giỏi, xây dựng thương hiệu và tền 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. Hiện nay môi trường mạng xã hội đang là nơi giúp cho các trường đại học kết nối gần nhất, nhanh nhất với tất cả mọi người ở mọi miền đất nước cũng như trên thế giới. Việc nghiên cứu sự quan tâm của người dùng mạng xã hội sẽ đem đến những cơ hội tuyệt vời để một trường đại học có thể tập trung định hướng việc tiếp thị hình ảnh của mình tới những người học tiềm năng qua đó thu hút, định hướng được sự quan tâm của người học đến với trường đại học đó.



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

Thông qua việc sử dụng trang mạng cá nhân, mỗi người có thể đưa ra quan điểm, nhận định, sự quan tâm, sở thích… của mình đối với một đối tượng nào đó bằng việc đăng tải hay chia sẻ bài viết. Qua đó, những người dùng khác có thể tương tác với họ bằng cách comment, like, share… những bài viết này. Bằng việc phân tích những bài viết, comment… đó, chúng ta sẽ phân loại được những nhóm người dùng có chung sự quan tâm đến một chủ đề riêng biệt. Việc tương tác với các nhóm người dùng như vậy sẽ nhanh chóng và hiệu quả rất nhiều hơn so với tương tác với từng cá

nhân đơn lẻ. Với công tác tuyển sinh của một trường đại học hiện nay bên cạnh việc tiếp cận trên môi trường mạng internet cũng phải trực tiếp tiếp cận với người học tại nơi họ học tập, sinh sống. Bởi vậy, nắm bắt được những mối quan tâm của người dùng tại những khu vực nhất định sẽ tạo sự thuận lợi rất lớn, công tác tuyển sinh sẽ đánh đúng và trúng vào những nhu cầu sở thích của người học tiềm năng từ đó nâng cao đáng kể hiệu quả cũng như tránh sự dàn trải cho công tác tuyển sinh.

Dưới đây là một số thuật ngữ thường được dùng trong mạng xã hội Facebook:

- *News feed*: là phần hiển thị nội dung trên trang chủ Facebook của người dùng, đó là một danh sách cập nhật liên tục những bài viết, hoạt động từ bạn bè hoặc các trang mà người dùng theo dõi trên Facebook; cũng là nơi mà người dùng đăng tải bài viết, hình ảnh, video… News feed đôi khi cũng hiển thị những quảng cáo của các trang fanpage nào đó.
- *Fanpage*: là một trang được tạo ra bởi một cá nhân hay một tổ chức, doanh nghiệp nào đó, hình thành nên một nhóm cộng đồng có cùng chung sở thích, quan tâm tới một chủ đề nhất định, giúp những người này có thể tương tác với nhau dễ dàng hơn. Fanpage cũng được coi như một nút mạng, tương đương với một người dùng trong mạng xã hội Facebook.
- *Bài viết (hay status):* là những thông tin người dùng đăng tải trên news feed của mình hoặc đăng tải lên fanpage, news feed của người dùng khác.
- *Bài share*: là một bài post bất kỳ của một nút nào đó trong mạng xã hội Facebook mà người dùng chia sẻ lại trên trang news feed của mình.
- *Comments*: là các ý kiến, bình luận của người dùng để lại dưới một bài viết bất kỳ.
- *Tag*: là một liên kết được dùng để đánh dấu một đối tượng cụ mà người dùng quan tâm và nhắc tới trong bài post, comment của mình.
- *Hash-tag*: cũng tương tự như tag nhưng không phải là một liên kết mà chỉ dùng để đánh dấu hay nhấn mạnh vấn đề mà họ quan tâm. Hash-tag được bắt đầu bởi ký tự "#" và theo sau là chuỗi các ký tự không chứa khoảng trắng.

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.

Mạng xã hội Facebook là một mạng xã hội có quy mô rất lớn, mang tính toàn cầu và ngày càng phát triển mạnh mẽ. Được sử dụng đông đảo bởi rất nhiều người dùng từ rất nhiều các quốc gia khác nhau trên thế giới, dẫn đến ngôn ngữ sử dụng trên mạng xã hội này cũng rất đa dạng. Hiện nay, tại Việt Nam, mạng xã hội Facebook đang là mạng xã hội được sử dụng phổ biến nhất; người dùng Facebook ở Việt Nam thuộc rất nhiều tầng lớp, không phân biệt nghề nghiệp, giới tính hay tuổi tác. Do vậy, luận văn đi sâu vào nghiên cứu bài toán khám phá quan tâm của người dùng trong mạng xã hội Facebook trên miền dữ liệu Tiếng Việt.

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.

Thư nhất, bài toán có ý nghĩa rất lớn đối với các trường đại học khi muốn giới thiệu, tiếp thị hình ảnh của mình tới những người học tiềm năng một cách nhanh chóng hiệu quả.

Thứ hai, qua việc khám phá quan tâm của người dùng sẽ giúp các trường đại học xác định được vị trí các cộng đồng người dùng có những mối quan tâm giống nhau từ đó tập trung giới thiệu tiếp thị các ngành đào tạo phù hợp giúp tiết kiệm thời gian chi phí nhưng đạt hiệu quả cao cho công tác truyền thông tuyển sinh của mình.

Thứ ba, các bài viết và ý kiến bình luận sẽ phản ánh một cách chân thực nhất quan tâm của người dùng nên việc khám phá quan tâm của người dùng qua bài viết và bình luận sẽ cho một kết quả chính xác cao.

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)

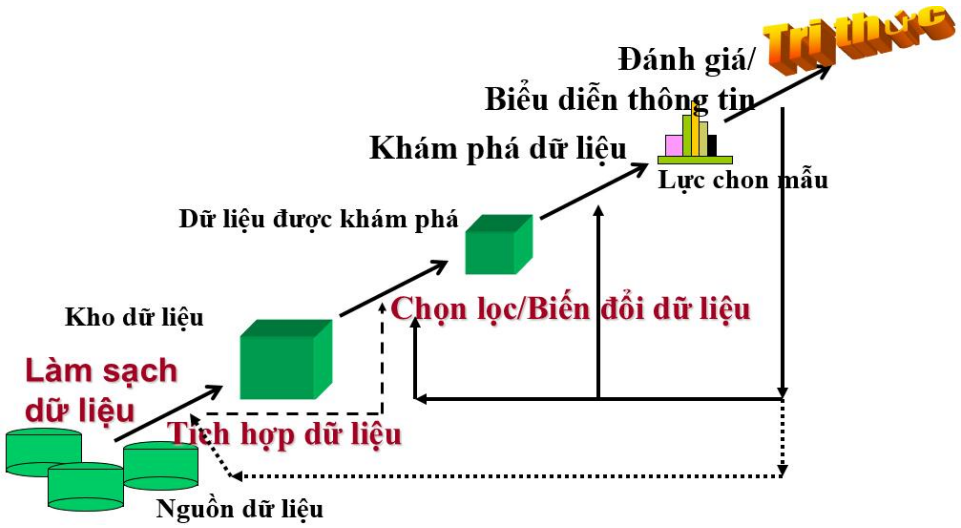
Khoảng hơn một thập kỷ trở lại đây, lượng thông tin được lưu trữ trên các thiết bị điện tử (đĩa cứng, CD-ROM, băng từ, .v.v.) không ngừng tăng lên. Sự tích lũy dữ liệu này xảy ra với một tốc độ bùng nổ. Người ta ước đoán rằng lượng thông tin trên toàn cầu tăng gấp đôi sau khoảng hai năm và theo đó số 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.

Khai phá dữ liệu (datamining) được định nghĩa như là một quá trình chắt lọc hay khai phá tri thức từ một lượng lớn dữ liệu. Một ví dụ hay được sử dụng là là việc khai thác vàng từ đá và cát, Dataming được ví như công việc "Đãi cát tìm vàng" trong một tập hợp lớn các dữ liệu cho trước. Thuật ngữ Dataming ám chỉ việc tìm kiếm một tập hợp nhỏ có giá trị từ một số lượng lớn các dữ liệu thô. Có nhiều thuật ngữ hiện được dùng cũng có nghĩa tương tự với từ Datamining như Knowledge Mining (khai phá tri thức), knowledge extraction (chắt lọc tri thức), data/patern analysis (phân tích dữ liệu/mẫu), data archaeoloogy (khảo cổ dữ liệu), datadredging (nạo vét dữ liệu),...

Khai phá dữ liệu là một bước trong bảy bước của quá trình KDD (Knowleadge Discovery in Database) và KDD được xem như 7 quá trình khác nhau theo thứ tự sau:

1. Làm sạch dữ liệu (data cleaning & preprocessing): Loại bỏ nhiễu và các dữ liệu không cần thiết.
2. Tích hợp dữ liệu: (data integration): quá trình hợp nhất dữ liệu thành những kho dữ liệu (data warehouses & data marts) sau khi đã làm sạch và tiền xử lý (data cleaning & preprocessing).
3. Trích chọn dữ liệu (data selection): trích chọn dữ liệu từ những kho dữ liệu và sau đó chuyển đổi về dạng thích hợp cho quá trình khai thác tri thức. Quá trình này bao gồm cả việc xử lý với dữ liệu nhiễu (noisy data), dữ liệu không đầy đủ (incomplete data), .v.v.
4. Chuyển đổi dữ liệu: Các dữ liệu được chuyển đổi sang các dạng phù hợp cho quá trình xử lý
5. Khai phá dữ liệu (data mining): Là một trong các bước quan trọng nhất, trong đó sử dụng những phương pháp thông minh để chắt lọc ra những mẫu dữ liệu.
6. Ước lượng mẫu (knowledge evaluation): Quá trình đánh giá các kết quả tìm được thông qua các độ đo nào đó.
7. Biểu diễn tri thức (knowledge presentation): Quá trình này sử dụng các kỹ thuật để biểu diễn và thể hiện trực quan cho người dùng.



Hình 1.7. Các bước trong Data Mining & KDD

Các chức năng chính của khai phá dữ liệu

Data Mining được chia nhỏ thành một số hướng chính như sau:

- Mô tả khái niệm (concept description): thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.
- Luật kết hợp (association rules): là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: "60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kính doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, .v.v.
- Phân lớp và dự đoán (classification & prediction): xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), .v.v. Người ta còn gọi phân lớp là học có giám sát.
- Phân cụm (clustering): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Người ta còn gọi phân cụm là học không giám sát (học không thầy).
- Khai phá chuỗi (sequential/temporal patterns): tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này

được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.

Ứng dụng của khai phá dữ liệu

Data Mining tuy là một hướng tiếp cận mới nhưng thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của nó. Chúng ta có thể liệt kê ra đây một số ứng dụng điển hình:

- Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)
- Điều trị y học (medical treatment)
- Text mining & Web mining
- Tin-sinh (bio-informatics)
- Tài chính và thị trường chứng khoán (finance & stock market)
- Bảo hiểm (insurance)
- Nhận dạng (pattern recognition)
- .v.v…

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

Tìm kiếm văn bản là quá trình tìm kiếm văn bản theo yêu cầu của người dùng. Các yêu cầu được thể hiện dưới dạng các câu hỏi (query), dạng câu hỏi đơn giản nhất là các từ khóa. Có thể hình dung hệ tìm kiếm văn bản sắp xếp văn bản thành hai lớp: Một lớp cho ra những các văn bản thỏa mãn với câu hỏi đưa ra và một lớp không hiển thị những văn bản không được thỏa mãn. Các hệ thống thực tế hiện nay không hiển thị như vậy mà đưa ra các danh sách văn bản theo độ quan trọng của văn bản tuỳ theo các câu hỏi đưa vào, ví dụ điển hình là các máy tìm tin như Google, Altavista,…

b. Quá trình

Quá trình tìm tin được chia thành bốn quá trình chính :

Đánh chỉ số (indexing): Các văn bản ở dạng thô cần được chuyển sang một dạng biểu diễn nào đó để xử lý. Quá trình này còn được gọi là quá trình biểu diễn văn bản, dạng biểu diễn phải có cấu trúc và dễ dàng khi xử lý.

Định dạng câu hỏi: Người dùng phải mô tả những yêu cầu về lấy thông tin cần thiết dưới dạng câu hỏi. Các câu hỏi này phải được biểu diễn dưới dạng phổ biến cho các hệ tìm kiếm như nhập vào các từ khóa cần tìm. Ngoài ra còn có các phương pháp định dạng câu hỏi dưới dạng ngôn ngữ tự nhiên hoặc dưới dạng các ví dụ, đối với các dạngnày thì cần có các kỹ thuật xử lý phức tạp hơn. Trong các hệ tìm tin hiện nay thì đại đa số là dùng câu hỏi dưới dạng các từ khóa.

So sánh: Hệ thống phải có sự so sánh rõ ràng và hoàn toàn câu hỏi các câu hỏi của người dùng với các văn bản đượcl ưu trữ trong CSDL. Cuối cùng hệ đưa ra một quyết định phân loại các văn bản có độ liên quan gầnvới câu hỏi đưa vào và thứ tự của nó. Hệ sẽ hiển thị toàn bộ văn bản hoặc chỉ một phần văn bản.

Phản hồi: Nhiều khi kết quả được trả về ban đầu không thỏa mãn yêu cầu của người dùng, do đó cần phải có qua trình phản hồi để người dùng có thết hay đổi lại hoặc nhập mới các yêu cầu của mình. Mặt khác, người dùng có thể tương tác với các

hệ về các văn bản thỏa mãn yêu cầu của mình và hệ có chức năng cập nhậu các văn bản đó. Quá trình này được gọi là quá trình phản hồi liên quan (Relevance feeback).

Các công cụ tìm kiếm hiện nay chủ yếu tập trung nhiều vào ba quá trình đầu, còn phần lớn chưa có quá trình phản hồi hay xử lý tương tác người dùng và máy. Quá trình phản hồi hiện nay đang được nghiên cứu rộng rãi và riêng trong quá trình tương tác giao diện người máy đã xuất hiện hướng nghiên cứu là interface agent.

Phân loại văn bản

a. Nội dung

Phân lớp văn bản được xem như là quá trình gán các văn bản vào một hay nhiều văn bản đã xác định từ trước. Người ta có thể phân lớp các văn bản một cách thủ công, tức là đọc từng văn bản một và gán nó vào một lớp nào đó. Cách này sẽ tốn rất nhiều thời gian và công sức đối với nhiều văn bản và do đó không khả thi. Do vậy mà phải có các phương pháp phân lớp tự động. Để phân lớp tự động người ta sử dụng các phương pháp học máy trong trí tuệ nhân tạo (Cây quyết định, Bayes, k người láng giềng gần nhất).

Một trong những ứng dụng quan trọng nhất của phân lớp văn bản là trong tìm kiếm văn bản. Từ một tập dữ liệu đã phân lớp các văn bản sẽ được đánh chỉ số đôi với từng lớp tương ứng. Người dùng có thể xác định chủ đề hoặc phân lớp văn bản mà mình mong muốn tìm kiếm thông qua các câu hỏi.

Một ứng dụng khác của phân lớp văn bản là trong lĩnh vực tìm hiểu văn bản. Phân lớp văn bản có thể được sử dụng để lọc các văn bản hoặc một phần các văn bản chứa dữ liệu cần tìm mà không làm mất đi tính phức tạp của ngôn ngữ tự nhiên.

Trong phân lớp văn bản, một lớp có thể được gán giá trị đúng sai hoặc được tính theo mức độ phụ thuộc (văn bản có môt mức độ phụ thuộc vào lớp). Trong trương hợp có nhiều lớp thì phân loại đúng sai sẽ là việc xem một văn bản có thuộc vào một lớp duy nhất nào đó hay không..

b. Quá trình

Quá trình phân lớp văn bản tuân theo các bước sau:

Đánh chỉ số (Indexing): Quá trình đánh chỉ số văn bản cũng giống như trong

quá trình đánh chỉ số của tìm kiếm văn bản. Trong phần này thì tốc độ đánh chỉ số đóng vai trò quan trọng vì một số các văn bản mới có thể cần đươc xử lý trong thời gían thực

Xác định độ phân lớp: Cũng giống như trong tìm kiếm văn bản, phân lớp văn bản yêu cầu quá trình diễn tả việc xác định văn bản đó thuộc lớp nào đó như thế nào, dựa trên cấu trúc biểu diễn của nó. Đối với hệ phân lớp văn bản, chúng ta gọi quá trình này là bộ phân lớp (Categorization hoặc classifier). Nó đóng vai trò như những câu hỏi trong hệ tìm kiếm. Nhưng trong khi những câu hỏi mang tính nhất thời, thì bộ phân loại được sử dụng một cách ổn định và lâu dài cho quá trình phân loại.

So sánh: Trong hầu hết các bộ phân loại, mỗi văn bản đều được yêu cầu gán đúng sai vào một lớp nào đó. Sự khác nhau lớn nhất đối với quá trình so sánh trong hệ tìm kiếm văn bản là mỗi văn bản chỉ được so sánh với một số lượng các lớp một lần và việc chọn quyết đnịh phù hợp còn phụ thuộc vào mối quan hệ giữa các lớp văn bản.

Phản hồi (Hay thích nghi): Quá trình phản hồi đóng vai trò trong hệ phân lớp văn bản. Thứ nhất là khi phân loại thì phải có môt số lượng lớn các văn bản đã được xếp loại bằng tay trước đó, các văn bản này được sử dụng làm mẫu huấn luyện để hỗ trợ xây dựng bộ phân loại. Thứ hai là đối với việc phân loại văn bản này không dễ dàng thay đổi các yêu cầu như trong quá trình phản hồi của tìm kiếm văn bản , người dùng có thể thông tin cho người bảo trì hệ thống về việc xóa bỏ, thêm vào hoặc thay đổi các phân lớp văn bản nào đó mà mình yêu cầu.

Một số bài toán khác

Ngoài hai bài toán kể trên, còn có các bài toán sau:

- Tóm tắt văn bản
- Phân cụm văn bản
- Phân cụm các từ mục
- Phân lớp các từ mục
- Đánh chỉ mục các từ tiềm năng
- Dẫn đường văn bản

Trong các bài toán xử lý văn bản đã nêu ở trên, chúng tra thấy vai trò của biểu diễn văn bản rất lớn, đặc biệt trong các bài toán tìm kiếm, phân lớp, phân cụm, dẫn đường.

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

Có những từ xuất hiện nhiều trong ngôn ngữ tự nhiên, tuy nhiên lại không mang nhiều ý nghĩa. StopWords (từ dừng) [Eduard Dragut et. al. (2009), "Stop Word and Related Problems in Web Interface Integration", VLDB '09, Lyon, France, Copyright 2009 VLDB Endowment.] là những từ thường xuyên xuất hiện trong văn bản mà không có nhiều thông tin nội dung (ví dụ: giới từ, liên từ, v.v.). Ở tiếng việt StopWords là những từ như: để, này, kia... Tiếng anh là những từ như: is, that, this...

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

Chúng ta có thể nhận ra rằng, trong văn bản có những từ xuất hiện rất ít lần. Nếu mục tiêu của chúng ta là xác định độ tương tự và sự khác nhau trong toàn bộ tập hợp các văn bản thì những từ xuất hiện rất ít đó (một hoặc hai lần) có ảnh hưởng không đáng kể tới văn bản đang xử lý.

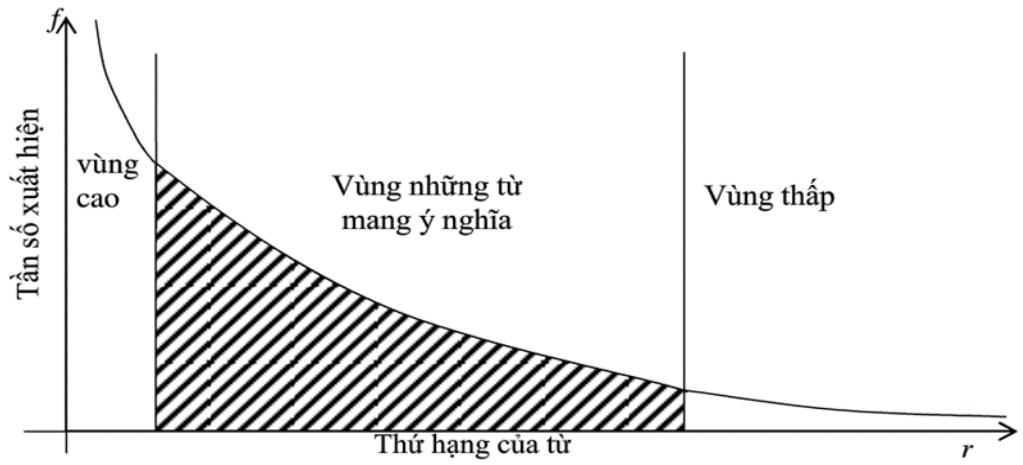
Năm 1949, với sự quan sát đó, Zipf đã phát biểu điều này và được mọi người coi đó như một định luật. Đó là một hiện tượng xấp xỉ toán học về tần số xuất hiện của một t trong tài liệu D. Dưới đây sẽ là mô tả chi tiết hơn về định luật này.

Gọi tần số xuất hiện của từ t trong tài liệu D là f_t. Sau đó sắp xếp các từ trong tập hợp các từ đó theo chiều giảm dần của tần số xuất hiện f và thứ hạng mỗi từ t là rt. Định luật Zipf được phát hiểu như sau:

$r_t . f_t \approx K$ (với K là một hằng số)

Giả sử t_i được xếp ở ô vị trí thấp nhất với tần số xuất hiện là b nào đấy và t_j cũng được xếp ở vị trí thấp kế tiếp với tần số $b + 1$. Khi đó ta có thể thu được xấp xỉ đối với các từ riêng biệt có tần số xuất hiện là b. Một cách tổng quát, một từ chỉ xuất hiện một lần trong tập hợp ta có $r_{max} = K$.

Cũng từ công thức trên, phân bố của các từ duy nhất xuất hiện b lần trong tập hợp sẽ là K/b.



Hình 1.8. Lược đồ thống kê tần số của từ theo định luật Zipf

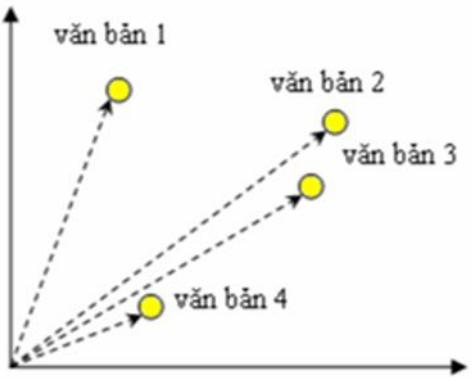
1.3.3.2 Mô hình không gian vector

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.

Mô hình này biểu diễn văn bản như những điểm trong không gian Euclid n-chiều, mỗi chiều tương ứng với một từ trong tập hợp các từ. Phần tử thứ i, là di của vector văn bản cho biết số lần mà từ thứ i xuất hiện trong văn bản. Sự tương đồng của hai văn bản được định nghĩa là khoảng cách giữa các điểm, hoặc là góc giữa những vector trong không gian.

Mỗi từ trong không gian vector sẽ có một trọng số, có nhiều phương pháp xếp hạng khác nhau, nhưng tf-idf (term frequency–inverse document frequency) là một phương pháp phổ biến để đánh giá và xếp hạng một từ trong một tài liệu. Về cơ bản thì tf-idf là một kỹ thuật giúp chuyển đổi thông tin dưới dạng văn bản thành một mô hình không gian vector thông qua các trọng số. Mô hình không gian vector và tf-idf được phát triển bởi Gerard Salton vào đầu thập niên 1960s.

Mặc dù đơn giản, nhưng mô hình không gian vector và những biến thể của nó hiện nay vẫn là cách phổ biến để biểu diễn văn bản trong Data mining và Information retrieval.



Hình 1.9. Biểu diễn các vector văn bản trong không gian 2 chiều

Term frequency tft,d xác định số lần từ t xuất hiện trong tài liệu d. Nhưng chỉ tần suất xuất hiện của một từ thôi thì chưa đủ.

Ví dụ trong một tài liệu, sự xuất hiện của một từ 10 lần thì tài liệu đó được coi là phù hợp hơn tài liệu mà từ đó chỉ xuất hiện 1 lần. Nhưng không phải là phù hợp hơn tài liệu kia 10 lần. Sự phù hợp không tỷ lệ thuận với số lần xuất hiện của từ đó trong một tài liệu.

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.

Ví dụ với câu: "Thời tiết rất đẹp" ta có các mô hình N-Gram như sau:

- Unigram (với n=1) sẽ bao gồm: thời; tiết; rất; đẹp.
- Bigram (với n=2) sẽ bao gồm: thời tiết; tiết rất; rất đẹp.
- Trigram (với n-3) sẽ bao gồm: thời tiết rất; tiết rất đẹp.

N-grams được áp dụng rộng rãi trong xử lý ngôn ngữ tự nhiên mang tính chất thống kê như hệ thống tách từ, gán nhãn từ loại… nó thường được dùng để ước lượng xác suất xuất hiện của một yếu tố dựa vào các yếu tố xung quanh nó trong câu. Trong phạm vi luận văn này, N-gram được sử dụng với yếu tố là từ vựng.

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.

Mặc dù không chọn được mô hình dự đoán tốt nhất nhưng ta vẫn có thể đạt được kết quả dự đoán cao. Hầu hết các thuật toán sẽ tự điều chỉnh các thông số phù hợp theo cấu trúc các thuộc tính hiện tại. Tuy nhiên, việc có được tập thuộc tính tinh gọn sẽ góp phần làm đơn giản hoá độ phức tạp tính toán của mô hình nhờ vậy mà tính toán được nhanh hơn và dễ dàng để diễn giải cho người dùng. Ví dụ, khi sử dụng mô hình cây quyết định, nếu ta sử dụng quá nhiều thuộc tính vào quá trình dự đoán, mặc dù cho kết quả rất tốt tuy nhiên, người dùng sẽ rất khó quan sát và diễn giải kết quả dự đoán.

Trong quá trình tối ưu hoá tham số, mặc dù không đạt được mục tiêu này nhưng với tập thuộc tính được thiết kế tốt, ta vẫn có thể đạt được kết quả dự đoán cao. Ta không cần phải cực lực tìm kiếm mô hình nào phù hợp nhất cũng như bộ trọng số được tối ưu nhất. Chỉ với tập thuộc tính được thiết kế tốt, ta đã mô tả được tập dữ liệu hiện có cũng như tiếp cận với bài toán thực tế dễ dàng và rõ ràng hơn rất nhiều.

2.1.2 Một số ví dụ về Xác định đặc trưng

Trực tiếp lấy dữ liệu thô

Với bài toán phân loại chữ số viết tay trong bộ cơ sở dữ liệu MNIST, mỗi bức ảnh có số chiều là 28 pixel x 28 pixel (tất nhiên việc crop và chỉnh sửa mỗi bức ảnh đã được thực hiện từ trước rồi, đó đã là một phần của feature engineering rồi). Một cách đơn giản thường được dùng là kéo dài ma trận 28x28 này để được 1 vector có số chiều 784. Trong cách này, các cột (hoặc hàng) của ma trận ảnh được đặt chồng lên (hoặc cạnh nhau) để được 1 vector dài. Vector dài này được trực tiếp sử dụng làm feature đưa vào các bộ classifier/clustering/regression/… Lúc này, giá trị của mỗi

pixel ảnh được coi là một feature.

Việc làm đơn giản này đã làm mất thông tin về không gian (spatial information) giữa các điểm ảnh, tuy nhiên, trong nhiều trường hợp, nó vẫn mang lại kết quả khả quan.

Lựa chọn đặc trưng

Giả sử rằng các điểm dữ liệu có số đặc trưng khác nhau (do kích thước dữ liệu khác nhau hay do một số feature mà điểm dữ liệu này có nhưng điểm dữ liệu kia lại không thu thập được), và số lượng đặc trưng là cực lớn. Chúng ta cần chọn ra một số lượng nhỏ hơn các đặc trưng phù hợp với bài toán.

Giảm kích thước

Một phương pháp làm giảm số chiều của dữ liệu để giảm bộ nhớ và khối lượng tính toán. Việc giảm số chiều này có thể được thực hiện bằng nhiều cách, trong đó random projection là cách đơn giản nhất. Tức chọn một ma trận chiếu ngẫu nhiên (ma trận béo) rồi nhân nó với từng điểm dữ liệu (giả sử dữ liệu ở dạng vector cột) để được các vector có số chiều thấp hơn. Ví dụ, vector ban đầu có số chiều là 784, chọn ma trận chiếu có kích thước (100x784), khi đó nếu nhân ma trận chéo này với vector ban đầu, ta sẽ được một vector mới có số chiều là 100, nhỏ hơn số chiều ban đầu rất nhiều. Lúc này, có thể ta không có tên gọi cho mỗi feature nữa vì các feature ở vector ban đầu đã được trộn lẫn với nhau theo một tỉ lệ nào đó rồi lưu và vector mới này. Mỗi thành phần của vector mới này được coi là một feature (không tên).

Việc chọn một ma trận chiếu ngẫu nhiên đôi khi mang lại kết quả tệ không mong muốn vì thông tin bị mất đi quá nhiều. Một phương pháp được sử dụng nhiều để hạn chế lượng thông tin mất đi có tên là Principle Component Analysis.

Feature learning không nhất thiết phải làm giảm số chiều dữ liệu, đôi khi feature vector còn có số chiều lớn hơn raw data. Random projection cũng có thể làm được việc này nếu ma trận chiếu là một ma trận cao (số cột ít hơn số hàng).

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.

Xét tập n văn bản D = $\{d_1, d_2, ..., d_n\}$ và tập T = $\{t_1, t_2, ..., t_n\}$ là tập các từ riêng biệt được trích ra từ tập văn bản D. Mỗi văn bản được biểu diễn thành một vector m chiều:

$$t_d = (tf(d, t_1), tf(d, t_2), ... tf(d, t_m))$$

trong đó $tf(d,t_i)$ là tần suất xuất hiện của t_i trong văn bản d.

* Các vấn đề của Bag of Words (Bow)

- Với bài toán này, từ điền có nhiều hơn 10 từ rất nhiều, như vậy vector đặc trưng thu được sẽ rất dài. Một văn bản mô tả đều được biểu diễn bằng các vector có số chiều rất lớn.

- Những từ hiếm đôi khi lại mang những thông tin qua trọng nhất mà chỉ loại văn bản đó có. Đây là một nhược điểm của BoW. Có một phương pháp cải tiến khác giúp khắc phục nhược điểm này có tên là Term Frequency-Inverse Document Frequency (TF-IDF) dùng để xác định tầm quan trọng của một từ trong một văn bản dựa trên toàn bộ văn bản trong cơ sở dữ liệu.

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.

Công thức tính:

$$\text{idf (t,D)} = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

$|D|$:Tổng số văn bản trong tập D

$|\{d \in D : t \in d\}|$: Số văn bản chứa từ nhất định, với điều kiện {\displaystyle t} xuất hiện trong văn bản d. Nếu từ đó không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho không không hợp lệ, vì thế người ta thường thay bằng mẫu thức $1 + |\{d \in D : t \in d\}|$

Cơ số logarit trong công thức này không thay đổi giá trị của 1 từ mà chỉ thu hẹp khoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi một số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF). Tuy nhiên việc thay đổi khoảng giá trị sẽ giúp tỷ lệ giữa IDF và TF tương đồng để dùng cho công thức TF-IDF như bên dưới.

Giá trị TF-IDF:

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

Ứng dụng

IDF có ứng dụng trong máy tìm kiếm. Ví dụ, khi người dùng gửi một truy vấn đến máy tìm kiếm, hệ thống cần biết từ nào là từ người dùng quan tâm nhất. Chẳng hạn: truy vấn của người dùng là "làm thế nào để sửa máy ủi". Sau khi tách từ, chúng ta sẽ có tập các từ: làm, thế nào, để, sửa, máy ủi. Trong các từ này, "máy ủi" sẽ có IDF cao nhất. Hệ thống sẽ lấy ra tất cả các văn bản có chứa từ máy ủi và sau đó mới thực hiện việc đánh giá và so sánh dựa trên toàn bộ câu truy vấn.

2.2.3 Phương pháp Phân rã giá trị số ít - SVD

SVD là viết tắt của thuật ngữ tiếng anh *singular value decomposition*, giải thuật SVD được Golub và Kahan giới thiệu năm 1965, đó là một công cụ phân rã ma trận hiệu quả được sử dụng để giảm hạng (hay số chiều) của ma trận. Kỹ thuật này được áp dụng vào nhiều bài toán xử lý văn bản khác nhau như tóm tắt văn bản, phát hiện sao chép, lập chỉ mục và truy vấn. SVD cho phép phân tích một ma trận phức tạp thành ba ma trận thành phần. Ứng dụng SVD trong thu nhập thông tin website với mục đích nhằm đưa việc giải quyết bài toán liên quan đến ma trận lớn, phức tạp về những bài toán nhỏ hơn đây là mô hình không gian vector riêng cho phép giảm kích thước từ ma trận giá trị của TF-IDF và loại bỏ nhiễu giúp cho quá trình xử lý tính toán được nhanh hơn. SVD dùng để giảm số chiều của ma trận mà vẫn giữ được các thông tin quan trọng và loại bỏ nhiễu... SVD khai triển văn bản thành r vector độc lập tuyến tính. Với ma trận trọng số X đã xây dựng ở trên, ta có thể khai triển giá trị riêng như sau:

$$X = U \Sigma V^T$$

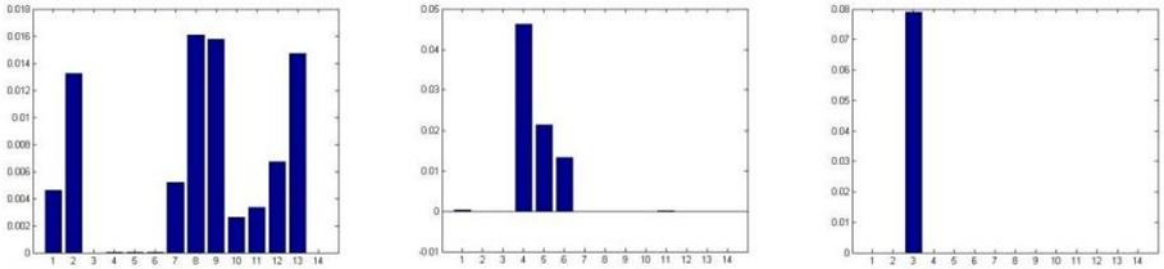
Trong đó

- U là ma trận trực giao cấp m × r (m là số từ chỉ mục) các vector dòng của

U là các vector từ chỉ mục

- Σ là ma trận đường chéo cấp r × r có các giá trị suy biến (singular value) $\sigma 1 \geq \sigma 2 \geq \cdots \geq \sigma r$, với r = rank (A)
- V là ma trận trực giao cấp r × n (n là số văn bản trong tập văn bản) các vector cột của V là các vector 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

Ta có thể thấy các thành phần của SVD mô tả số lần các từ xuất hiện trong tất cả các văn bản, bên cạnh đó ta còn có thể thấy được sự khác nhau giữa các từ xuất hiện nhiều lần trong một văn bản mà không xuất hiện ở văn bản khác. Ta có thể xem mỗi văn bản là một vector mà mỗi thành phần tương ứng với mỗi thuật ngữ trong tập dữ liệu cùng với trọng số của nó. Với thuật ngữ không xuất hiện trong văn bản thì trọng số bằng 0. Ta sẽ biểu diễn các văn bản này trong mô hình không gian vector và sẽ sử dụng cách tính điểm trong câu truy vấn.

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

Thuật toán Naïve Bayes dựa trên định lý Bayes được phát biểu như sau:

$$P(Y|X) = \frac{P(XY)}{P(X)} = \frac{P(X|Y)P(Y)}{P(X)}$$

Áp dụng trong bài toán phân loại, các dữ kiện gồm có:

- *D*: tập dữ liệu huấn luyện đã được vector hóa dưới dạng $X = (x_1, x_2, \dots, x_n)$.
- *Ci*: phân lớp *i*, với *i = {1,2,...,m}*.
- Các thuộc tính độc lập điều kiện đôi một với nhau.

Theo định lý Bayes:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Theo tính chất độc lập điều kiện:

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$$

Trong đó:

- $P(C_i|X)$ là xác suất thuộc phân lớp i khi biết trước mẫu *X*.
- $P(C_i)$ xác suất là phân lớp *i*.
- $P(x_k|C_i)$ xác suất thuộc tính thứ *k* mang giá trị x_k khi đã biết *X* thuộc phân lớp *i*.

Các bước thực hiện thuật toán Naïve Bayes:

Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu), tính $P(C_i)$ và $P(x_k|C_i)$.

Bước 2: Phân lớp $X^{new} = (x_1, x_2, \dots, x_n)$, ta cần tính xác suất thuộc từng phân lớp khi đã biết trước X^{new}. X^{new} được gán vào lớp có xác suất lớn nhất theo công thức:

$$\max_{C_i \in C} \left(P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \right)$$

Thuật toán Naïve Bayes với bài toán phân loại văn bản

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm

quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề. Kết quả dự đoán bị ảnh hưởng bởi kích thước tập dữ liệu, chất lượng của không gian đặc trưng.

Mô tả vector đặc trưng của văn bản: là vector có số chiều là số đặc trưng trong toàn tập dữ liệu, các đặc trưng này đôi một khác nhau. Nếu văn bản có chứa đặc trưng đó sẽ có giá trị 1, ngược lại là 0.

Thuật toán gồm hai giai đoạn huấn luyện và phân lớp rất rõ ràng:

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.

Phân lớp

Input:

- Vector đặc trưng của văn bản cần phân lớp
- Các giá trị xác suất tính $P(C_i)$ và $P(x_k|C_i)$.

Output:

- Giá trị xác suất thuộc phân lớp *I* khi biết trước mẫu *X*.

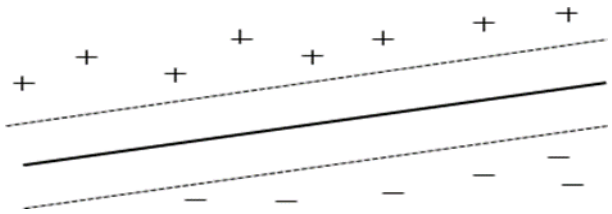
Dựa vào vector đặc trưng của văn bản cần phân lớp, áp dụng công thức tính xác suất thuộc từng phân lớp cho văn bản và chọn ra lớp có xác suất cao nhất.

2.3.2 Thuật toán vector hỗ trợ

Thuật toán máy vector hỗ trợ (Support Vector Machines - SVM) [4] 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.

Cho trước một tập huấn luyện, được biểu diễn trong không gian vector, trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu phẳng quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng là lớp + và lớp -. Chất lượng của siêu phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khi đó, khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác.

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

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian F và siêu phẳng quyết định f trên F sao cho sai số phân loại là thấp nhất.

Cho tập mẫu $(x_1, y_1), (x_2, y_2), \ldots (x_f, y_f)\}$ với $x_i \in R_n$, thuộc vào hai lớp nhãn: $y_i \in$

{-1,1} là nhãn lớp tương ứng của các xi (-1 biểu thị lớp I, 1 biểu thị lớp II).

Ta có, phương trình siêu phẳng chứa vector $\vec{x_l}$ trong không gian $\vec{x_l} . \vec{w} + b = 0$

Đặt $f\left(\vec{x_l}\right) = \text{si}\,gn(\vec{x_l}.\vec{w} + b) = \begin{cases} +1, & \vec{x_l} . \vec{w} + b > 0 \\ -1, & \vec{x_l} . \vec{w} + b < 0 \end{cases}$

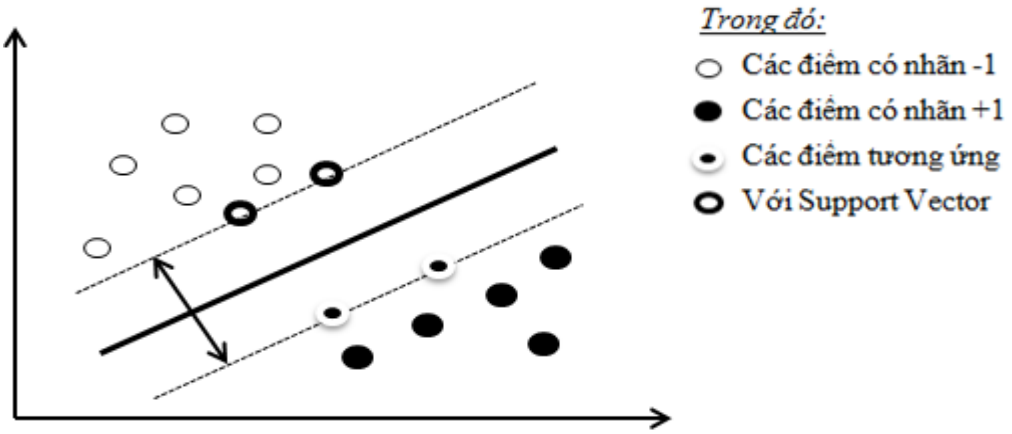
Như vậy, $f(X_i)$ biểu diễn sự phân lớp của X_i vào hai lớp như đã nêu. Ta nói $y_i = +1$ nếu X_i thuộc lớp I và $y_i = -1$ nếu X_i thuộc lớp II. Khi đó, để có siêu phẳng f ta sẽ phải giải bài toán sau: Tìm min w với W thỏa mãn điều kiện sau:

$y_i(\sin(\vec{x_l} . \vec{w} + b)) \geq 1$ với $\forall i \in \overrightarrow{1,n}$

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi về thành dạng đẳng thức. Một đặc điểm thú vị của SVM là mặt phẳng quyết định chỉ phụ thuộc các Support Vector và nó có khoảng cách đến mặt phẳng quyết định là $\frac{1}{\| w \|}$. Cho dù các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Đây chính là điểm nổi bật của phương pháp SVM so với các phương pháp khác vì tất cả các dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả.

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

Để xác định hàm phân lớp dựa trên phương pháp SVM, ta sẽ tiến hành tìm hai siêu phẳng song song sao cho khoảng cách y giữa chúng là lớn nhất có thể để phân tách hai lớp này ra làm hai phía. Hàm phân tách tương ứng với phương trình siêu phẳng nằm giữa hai siêu phẳng tìm được.

Các điểm mà nằm trên hai siêu phẳng phân tách được gọi là các Support Vector. Các điểm này sẽ quyết định đến hàm phân tách dữ liệu.

Bài toán nhiều phân lớp với SVM

Để phân nhiều lớp thì kỹ thuật SVM nguyên thủy sẽ chia không gian dữ liệu thành 2 phần và quá trình này lặp lại nhiều lần. Khi đó hàm quyết định phân dữ liệu vào lớp thứ i của tập n, 2-lớp sẽ là:

$$f_i(x) = \mathrm{w}_i^i x + b_i$$

Những phần tử x là support vector sẽ thỏa điều kiện:

$$f_i(x) = \begin{cases} +1 & \textit{nếu thuộc lớp } i \\ -1 & \textit{nếu thuộc lớp còn lại} \end{cases}$$

Như vậy, bài toán phân nhiều lớp sử dụng phương pháp SVM hoàn toàn có thể thực hiện giống như bài toán 2 lớp. Bằng cách sử dụng cách thức "một - đối - một".

Giả sử bài toán cần phân loại có k lớp (k > 2), chiến lược "một-đối-một"sẽ tiến hành $\frac{k(k-1)}{2}$ lần phân lớp nhị phân sử dụng phương pháp SVM. Mỗi lớp sẽ tiến hành phân tách với k-1 lớp còn lại để xác định k-1 hàm phân tách dựa vào bài toán phân

hai lớp bằng phương pháp SVM.

Các bước chính của phương pháp SVM

+ Phương pháp SVM yêu cầu dữ liệu được biểu diễn như các vector của các số thực. Như vậy nếu đầu vào chưa phải là số thì ta cần phải tìm cách chuyển chúng về dạng số của SVM.

+ Tiền xử lý dữ liệu: Thực hiện biến đổi dữ liệu phù hợp cho quá trình tính toán, tránh các số quá lớn mô tả các thuộc tính. Thường nên co giãn (scaling) dữ liệu để chuyển về đoạn [-1, 1] hoặc [0, 1].

+ Chọn hàm hạt nhân: Lựa chọn hàm hạt nhân phù hợp tương ứng cho từng bài toán cụ thể để đạt được độ chính xác cao trong quá trình phân lớp.

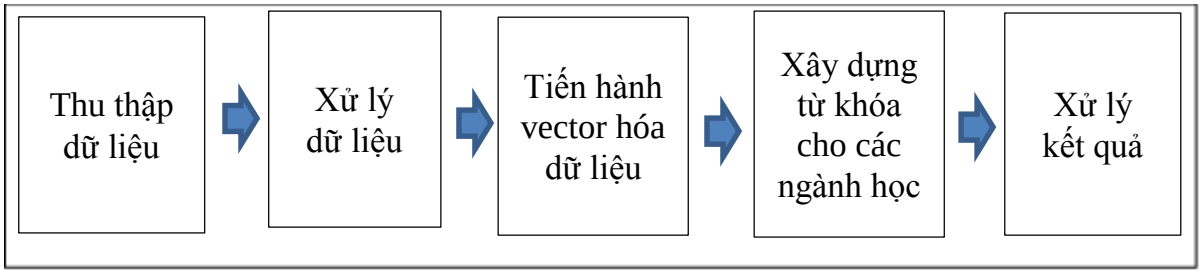
+ Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng đụng. Điều này cũng quyết định đến tính chính xác của quá trình phân lớp.

+ Sử dụng các tham số cho việc huấn luyện với tập mẫu. Trong quá trình huấn luyện sẽ sử dụng thuật toán tối ưu hóa khoảng cách giữa các siêu phẳng trong quá trình phân lớp, xác định hàm phân lớp trong không gian đặc trưng nhờ việc ánh xạ dữ liệu vào không gian đặc trưng bằng cách mô tả hạt nhân, giải quyết cho cả hai trường hợp dữ liệu là phân tách và không phân tách tuyến tính trong không gian đặc trưng.

+ Kiểm thử tập dữ liệu Test.

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 tiếng Việt của người dùng mạng xã hội tại Việt Nam.

Bước 2: Xử lý dữ liệu:

+ Tiến hành tách từ với dữ liệu là các bài viết thu được từ Facebook.

+ Loại bỏ từ dừng (stopwords) và các ký tự đặc biệt.

+ Xây dựng bộ từ điển các từ.

Bước 3: Tạo vector thuộc tính của các từ trong tập văn bản với mô hình Túi từ - Bag of words đánh lại trọng số với tf-idf và sử dụng SVD (singular value decomposition) nhằm mục đích giảm chiều dữ liệu của ma trận đã được trình bày trong mục 2.2 của luận văn.

Bước 4: Xây dựng từ khóa về các ngành học: Ngành Xây dựng; Ngành Kỹ thuật hạ tầng và Môi trường Đô thị; Ngành Quản lý đô thị; Ngành Kiến trúc, Quy hoạch; Ngành Mỹ thuật và nội thất 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: Với các bộ từ khóa đã được xây dựng ở bước 4, tiến hành trích xuất dữ liệu vector đại diện của các từ khóa này từ CSDL thu được ở bước 3. Đánh giá độ lớn vector đại diện của các từ khóa theo công thức tính độ lớn của vector trong không gian:

$$||A|| = \sqrt{\sum_{k=1}^{n} A_k^2}$$

- ||A|| là độ lớn của vector A=$[A_1, A_2, \dots, A_n]$;

- $A_1, A_2, \dots, A_n$là các trọng số của vector A.

Độ lớn của các vector sẽ thể hiện cho tần suất xuất hiện của từ khoá trong bộ

dữ liệu văn bản. Bởi vậy vector đại diện cho từ khoá 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 từ đó đánh giá quan tâm của họ để có giải pháp phục vụ công tác truyền thông tuyển sinh.

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:

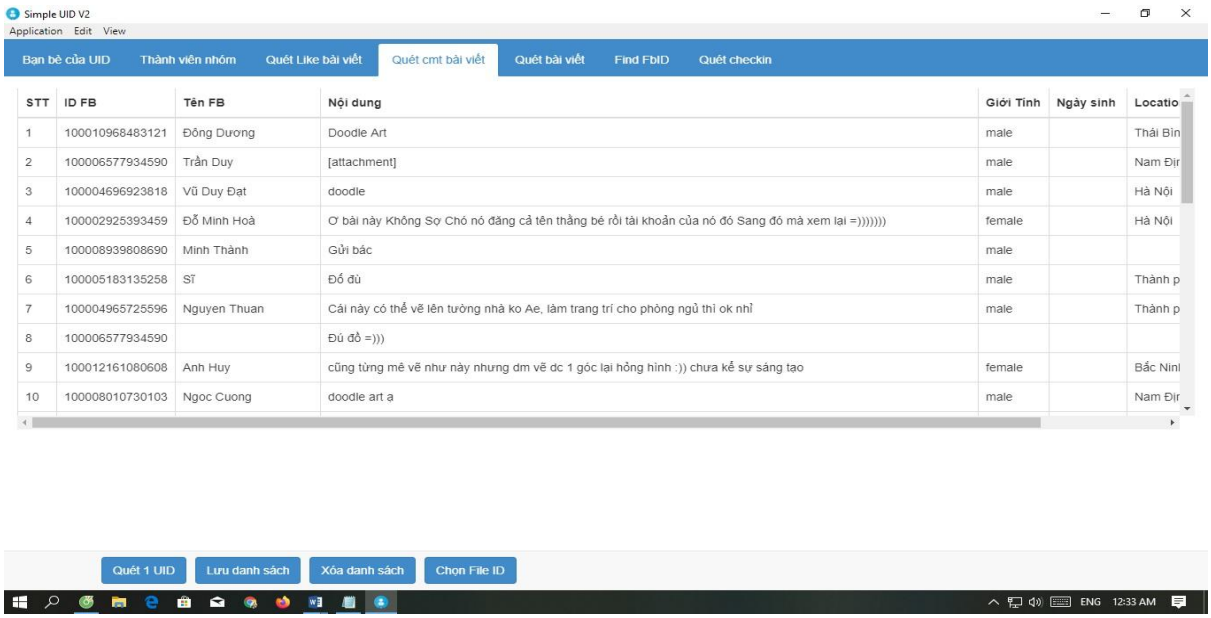
Input: Các bài viết của người dùng tạo ra trên mạng xã hội Facebook, các từ khoá đặc trưng của các ngành học của trường đại học.

Output: Người dùng A quan tâm đến ngành học nào? Ngành học H có những người dùng nào quan tâm?

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



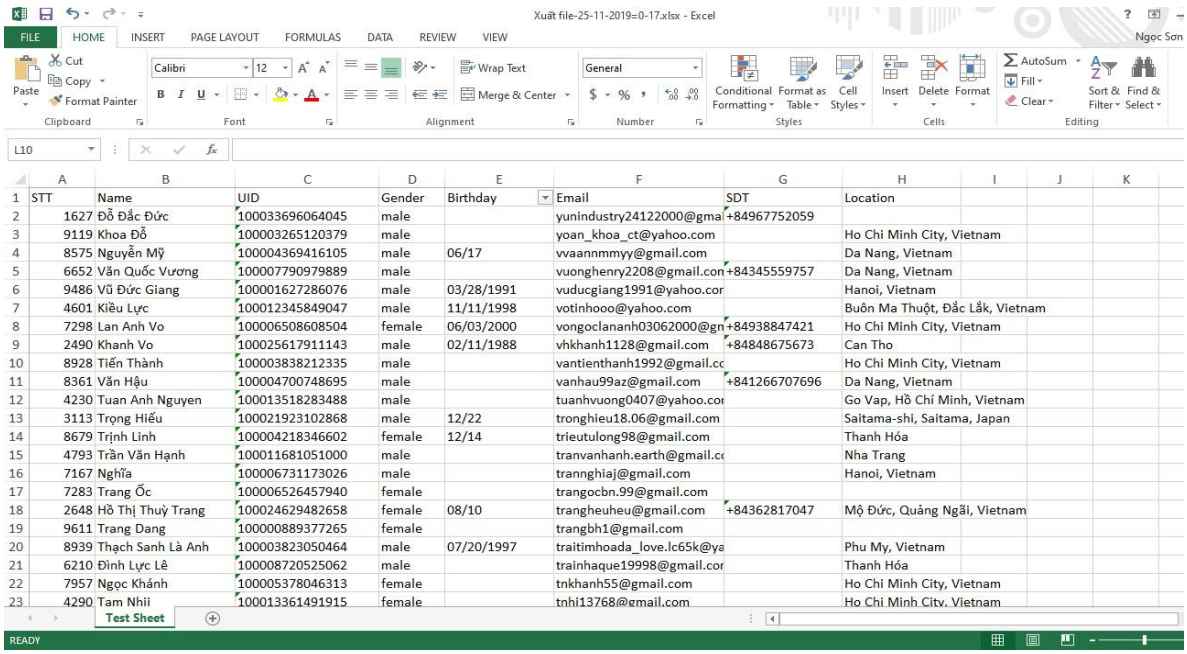
Hình 3.1. Giao diện phần mềm Simple UID

Mỗi tài khoản của một người dùng Facebook đều được cung cấp một mã gọi là mã token. Khi người dùng đăng xuất khỏi Facebook hoặc ngắt kết nối Internet, mã token sẽ không còn giá trị. Với mã token này, chúng ta có thể thu được các thông tin công khai của người dùng sở hữu mã. Đồng thời cũng thu được những thông tin của người dùng khác có kết nối với người sở hữu mã token này trên mạng xã hội Facebook. Hay nói cách cách khác, là thu được thông tin của các nút có liên kết trực tiếp với người dùng đó trong mạng. Tất nhiên, những thông tin này phải được người dùng công khai, trường hợp những thông tin là riêng tư thì trình khám phá sẽ không thu thập được.

Phần mềm khám phá Facebook Simple UID cung cấp một số trường thông tin chúng ta có thể lấy như: ID, tên, địa chỉ, ngày sinh, giới tính, feed... Trong đó, Feed ở đây là những thông tin mà người dùng thực hiện trên trang Newsfeed cá nhân, có thể bao gồm các status, bài share, comment…

3.2.2 Mô tả dữ liệu

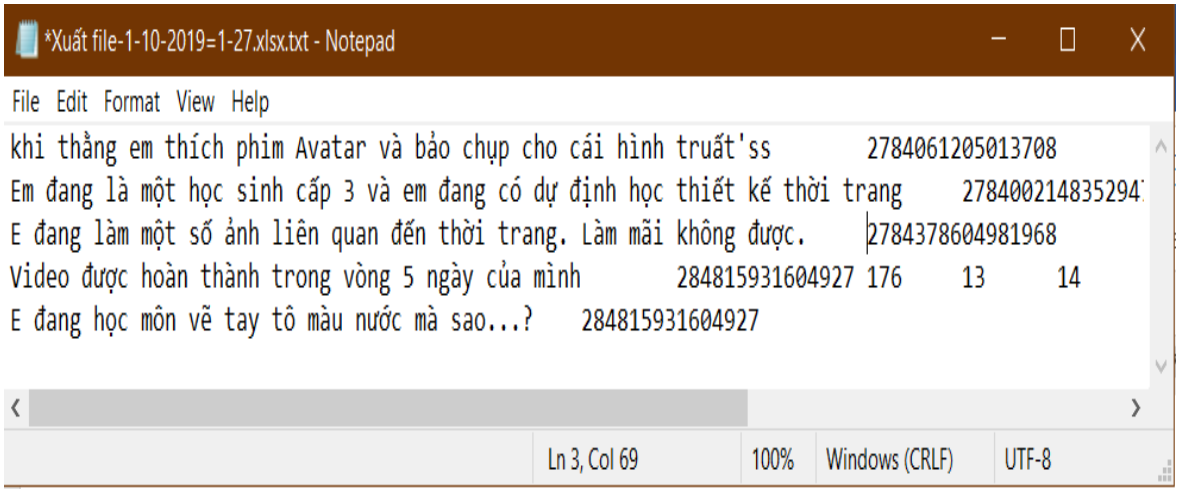
Dữ liệu thử nghiệm của luận văn gồm hàng trăm bài viết từ người dùng trên mạng xã hội Facebook. Mỗi người dùng có một ID và có các trường thông tin sau: tên, địa chỉ, giới tính, ngày sinh, bài viết.



STT	Name	UID	Gender	Birthday	Email	SDT	Location
1627	Đỗ Đắc Đức	100033696064045	male		yunindustry24122000@gma	+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

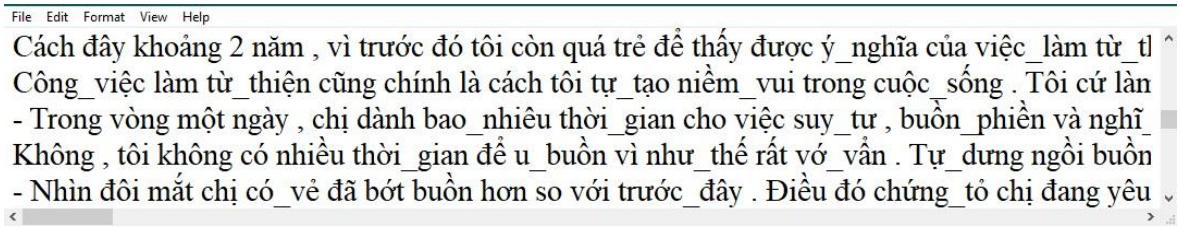
Kết quả thử nghiệm của luận văn được hoàn thiện bởi các phần mềm và công cụ sau:

- Lấy dữ liệu Facebook bằng phần mềm Simple UID
- Công cụ VnTokenizer từ https://github.com/
- Sử dụng ngôn ngữ lập trình Python 3.6.
- Xây dựng từ khóa các ngành học với sự giúp đỡ của Phòng Đào tạo ĐH Kiến trúc Hà Nội.

3.4 Xử lý dữ liệu

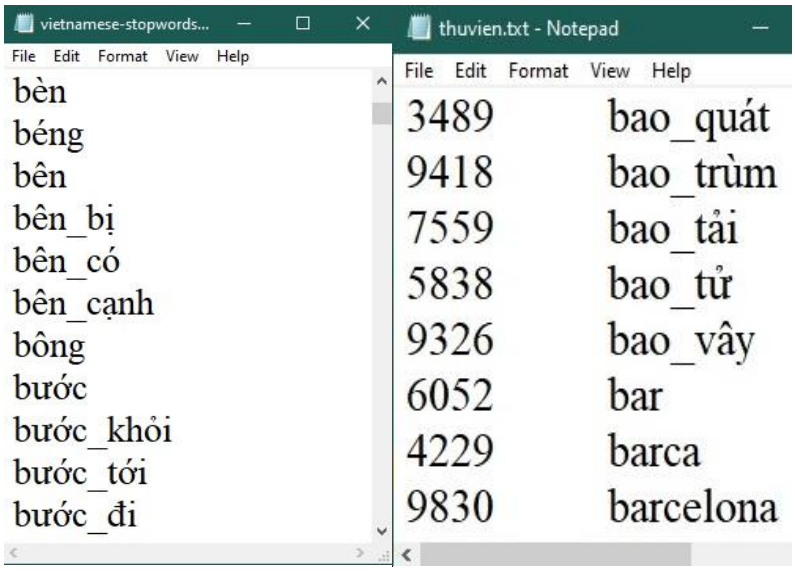
Dưới đây tác giả sẽ trình bày chi tiết cách xử lý dữ liệu cho mô hình giải quyết bài toán khám phá quan tâm của người dùng trên mạng xã hội phục vụ công tác truyền thông tuyển sinh của trường đại học.

Bước 1: Tiến hành tách từ với dữ liệu là các bài viết thu được từ Facebook. Tuy nhiên tiếng Việt không đơn giản như tiếng Anh vì nó có những từ ghép. Tác giả sử dụng công cụ VnTokenizer để thực hiện việc tách từ tiếng Việt. Dữ liệu thu được sau quá trình tách từ được thể hiện ở hình dưới.



Hinh 3.4. Dữ liệu sau quá trình tách từ

Bước 2: loại bỏ từ dừng (stopwords) và các ký tự đặc biệt. Sau quá trình loại bỏ từ dừng và các ký tự đặc biệt dữ liệu thu được 10.563 từ sau đó xây dựng bộ từ điển các từ bằng gensim trên python. Dữ liệu từ dừng và từ điển các từ được lưu trữ trong file vietnamese-stopwords.txt và thuvien.txt



Hình 3.5. File stopwords và từ điển

Bước 3: Tạo vector thuộc tính với mô hình Túi từ - Bag of words đánh lại trọng số với tf-idf và sử dụng SVD (singular value decomposition) nhằm mục đích giảm chiều dữ liệu của ma trận.

Ví dụ: ta có 2 câu

- "Em đang là học sinh cấp 3 và em đang có ý định học thiết kế thời trang."
- "Em đang làm một số ảnh liên quan đến thời trang."

Tập các từ trong văn bản:

Vb1= {Em, đang, là, học_sinh, cấp, 3, và, em, có, ý_định, học, thiết_kế, thời_trang}

Vb2= {Em, đang, làm, một, số, ảnh, liên_quan, đến, thời_trang}

Tập V là tập các từ có trong 2 văn bản: V= {em, đang, là, làm, học_sinh, cấp, 3, và, có, ý_định, học, thiết_kế, thời_trang, một, số, ảnh, liên_quan, đến}

Tạo vector lưu trữ số lần xuất hiện của từ trong V ứng với Vb1 và Vb2:

$\overrightarrow{Vb1}$= [2, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0]

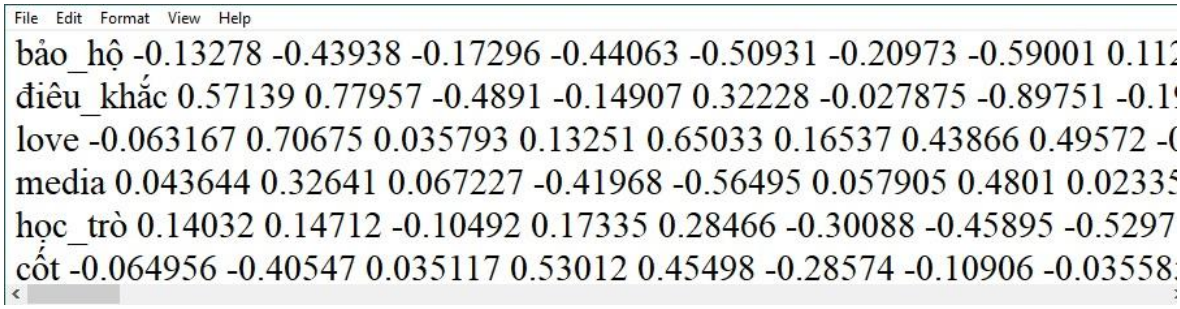
$\overrightarrow{Vb2}$= [1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,1]

Tiến hành đánh lại trọng số với tf-idf đã được trình bày tại mục 2.2.3 của chương 2:

$\overrightarrow{Vb1}$= [-0.02709096, -0.01354548, 0.023156153, 0.0, 0.023156153, 0.023156153, 0.023156153, 0.023156153, 0.023156153, 0.023156153, 0.023156153, 0.023156153, 0.023156153, -0.01354548, 0.0, 0.0, 0.0, 0.0, 0.0]

$\overrightarrow{Vb2}$= [-0.01956569, -0.01956569, 0.0, 0.033447777, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, -0.01956569, 0.033447777, 0.033447777, 0.033447777, 0.033447777, 0.033447777]

Từ $\overrightarrow{Vb1}$ và $\overrightarrow{Vb2}$ có thể thu được trọng số vector biểu diễn của một từ có trong hệ thống văn bản ví dụ: thời_trang= [-0.02709096, -0.01956569]; học_sinh= [0.0, 0.033447777].



Hình 3.6. Dữ liệu thu được sau khi vector hóa

Bước 4: Xây dựng bộ từ khóa cho các ngành học (tác giả xây dựng bộ từ khóa cho các ngành học thuộc trường ĐH Kiến trúc Hà Nội):

- Ngành Xây dựng: vật liệu, công trình, kết cấu, thi công, xây dựng.
- Ngành Kỹ thuật hạ tầng và Môi trường Đô thị: môi trường, cấp nước, thoát nước, giao thông, hạ tầng.
- Ngành Quản lý đô thị: kinh tế, quản lý, môi giới, đầu tư, kinh doanh.
- Ngành Kiến trúc, Quy hoạch: kiến trúc, quy hoạch, thiết kế, cấu trúc, cảnh quan.
- Ngành Mỹ thuật và nội thất: nghệ thuật, đồ họa, điêu khắc, mỹ thuật, hội họa, thời trang.

Bước 5: Tiến hành đánh giá kết quả thu được của **bước 3** và **bước 4.**

Với các bộ từ khóa đã được xây dựng ở bước 4, tiến hành trích xuất dữ liệu vector đại diện của các từ khóa này từ CSDL thu được ở bước 3. Như đã được trình bày trong chương II của đồ án, độ lớn vector đại diện của các từ khóa này sẽ thể hiện cho cường độ xuất hiện của chúng trong bộ dữ liệu văn bản. Bởi vậy 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. Độ lớn của vector trong không gian được tính theo công thức:

$$||A|| = \sqrt{\sum_{k=1}^{n} A_k^2}$$

- ||A|| là độ lớn của vector A=$[A_1, A_2, \dots, A_n]$;

- $A_1, A_2, \dots, A_n$là các trọng số của vector A.

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

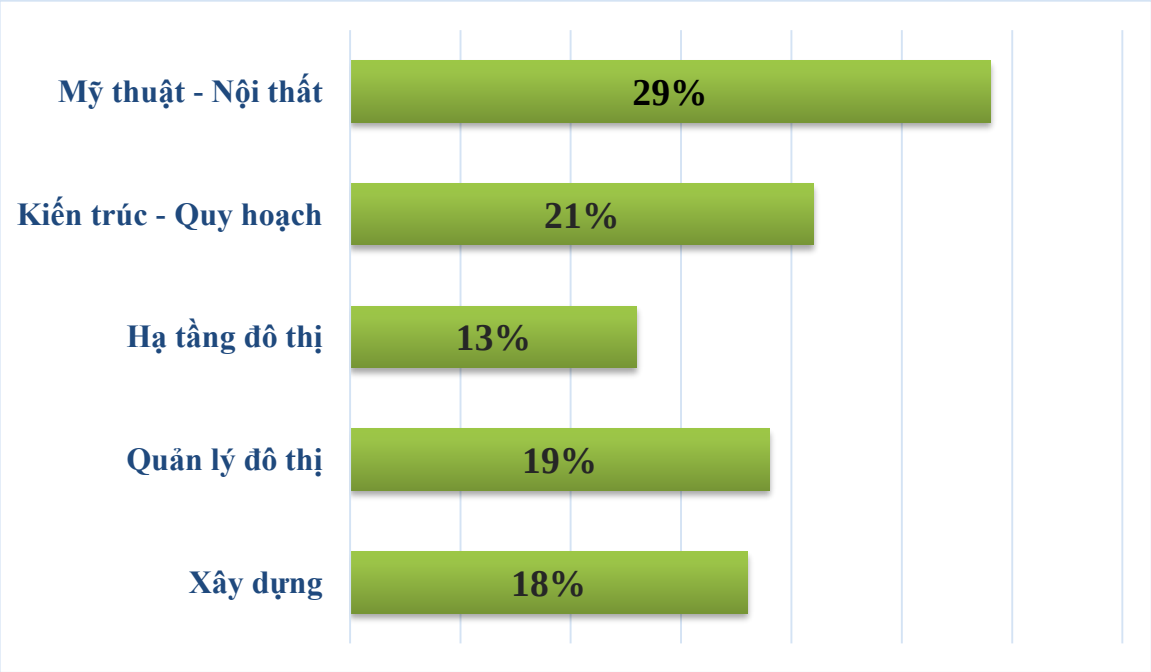
Hình 3.7. Kết quả thu được với từ khóa của từng ngành học

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

Với kết quả thu được sau quá trình xử lý dữ liệu tác giả nhận thấy người dùng dưới 22 tuổi tại khu vực Hà Nội có sự quan tâm nhiều tới ngành học Mỹ thuật – Nội thất cao nhất chiếm 29% tiếp đến là Kiến trúc – Quy hoạch 21%, còn lại là Quản lý đô thị 19%, Xây dựng 18% và Hạ tầng đô thị 13%

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.

DANH MỤC CÁC TÀI LIỆU THAM KHẢO

Danh mục tài liệu tiếng Việt:

[1] Nguyễn Thị Hội, Trần Đình Quế, '2018', "Ước lượng quan tâm người dùng trên mạng xã hội dựa trên tương tự bài viết", *Tạp chí khoa học và công nghệ Đại học Đà Nẵng*, 'số 7'.

[2] Nguyễn Thị Hội, Đàm Gia Mạnh, Trần Đình Quế (2017), *Độ tương đồng ngữ nghĩa các bài viết trên mạng xã hội dựa trên Wikipedia*, Hội nghị Khoa học Quốc gia lần thứ X – Nghiên cứu cơ bản và ứng dụng CNTT (FAIR'10), Đà Nẵng.

[3] Bùi Khánh Linh, Nguyễn Thị Thu Hà, Nguyễn Thị Ngọc Tú, Đào Thanh Tĩnh (2016), *Phân loại văn bản tiếng việt dựa trên mô hình chủ đề*, Hội nghị Khoa học Quốc gia lần thứ IX ―Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9), Cần Thơ.

Danh mục tài liệu tiếng Anh:

[4] Bing Liu (2007), *Web Data Mining*, Department of Computer Science University of Illinois at Chicago, Springer, Berlin, Germany.

[5] Diana Palsetia, Md. Mostofa, Ali Patwary, Kunpeng Zhang, Kathy Lee, Christopher Moran, Yves Xie, Daniel Honbo, Ankit Agrawal, Wei-keng Liao, Alok Choudhary, User-Interest based Community Extraction in Social Networks, ACM, NY, USA, 2012.

[6] Mehdi Allahyari, Seyed Amin Pouriyeh, Mehdi Assefi, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys J. Kochut, '2017', "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", *arXiv:1707.02919 [cs.CL]*.

[7] M. F. Schwartz and D. C. M. Wood (1993), "Discovering shared interests using graph analysis", Communications of the ACM, pp. 78–89.

[8] TH Nguyen, DQ Tran, GM Dam, MH Nguyen, Estimating the similarity of social network users based on behaviors, Vietnam Journal of Computer Science, 2018.

[9] X. Li, L. Guo, and Y. E. Zhao (2008), "Tag-based social interest discovery", *World Wide Web. New York, NY, USA: ACM*, pp. 675–684.

Danh mục website tham khảo:

[10] https://vinaresearch.net/, tuy nhập ngày 24/02/2020

PHỤ LỤC

Danh sách các từ stop words của tiếng Việt:

a_ha	bập_bà_bập_bõm	bỗng	chết_nỗi	có
a-lô	bập_bõm	bỗng_chốc	chết_tiệt	có chăng là
à_ơi	bất_chợt	bỗng_dưng	chết_thật	có dễ
á	bất_cứ	bỗng_đâu	chí_chết	có thể
à	bất_đồ	bỗng_không	Chín	có_thể
á_à	bất_giác	bỗng_nhiên	chỉn	có thế
ạ	bất_kể	bỏ_bố	chính	có vẻ
ạ_ơi	bất_kì	bỏ_mẹ	chính_là	cóc khô
ai	bất_kỳ	bớ	chính_thị	coi bộ
ai_ai	bất_luận	bởi	chỉ	coi mòi
ai_nấy	bất_nhược	bởi_chung	chỉ do	con
ái	bất_quá	bởi_nhưng	chỉ là	còn
ái_chà	bất_thình_lình	bởi_thế	chỉ tại	cô mình
ái_dà	bất_tử	bởi_vậy	chỉ vì	cổ lai
ái_khanh	bây_bẩy	bởi_vì	chiếc	công nhiên
alô	bây_chừ	bức	cho đến	cơ
amen	bây_giờ	cả	cho đến khi	cơ chừng
anh_ta	bây_giờ	cả_thảy	cho nên	cơ hồ
áng	bây_nhiêu	cái	cho tới	cơ mà
ào	bấy	các	cho tới khi	cơn
ăn	bấy_giờ	cả_thảy	Choa	cu cậu
ắt	bấy_chầy	cả_thể	chốc chốc	của
ắt_hẳn	bấy_chừ	càng	chớ	cùng
ắt_là	bấy_giờ	cần	chớ chi	cùng cực
âu_là	bấy_lâu	căn	chợt	cùng nhau
ầu_ơ	bấy_lâu_nay	căn_cắt	Chú	cùng với
ấy	bấy_nay	cật_lực	chu cha	cũng
ba	bấy_nhiêu	cật_sức	chú mày	cũng như
bài	bèn	cây	chú mình	cũng vậy
bàn	béng	cha_chả	chui cha	cũng vậy thôi
bán_mạng	bển	chành_chạnh	chùn chùn	cứ
bản	bệt	chao_ôi	chùn ch n	cứ việc
bao_giờ	bị	chắc	chủn	cực kì

bao_lăm	biết_bao	chắc_hẳn	chung cục	cực kỳ
bao_lâu	biết_bao_nhiêu	chăn_chắn	chung qui	cực lực
bao_nả	biết_chừng_nào	chăng	chung quy	cuộc
bao_nhiêu	biết_đâu	chằn_chặn	chung quy lại	cuốn
bay_biến	biết_đâu_chừng	chẳng_lẽ	chúng mình	dành
bằng	biết_đâu_đấy	chẳng_những	chúng ta	dào
bằng_ấy	biết_mấy	chẳng_nữa	chúng tôi	dạ
bằng_không	bộ	chẳng_phải	chứ	dần dà
bằng_nấy	bội_phần	chậc	chứ lại	dần dần
bắt_đầu_từ	bông	chầm_chập	chứ lị	dầu sao
dẫu	để cho	ha	hồi lùng	không thể
dẫu sao	đêm đêm	hả	hốt nhiên	không_thể
dễ sợ	đến	hà rằm	Hơi	không thể nào
dễ thường	đến cùng	hà rầm	hỡi	không trách
dĩ chí	đến đây	hàng	hỡi ôi	khư khư
do	đến nỗi	hay là	Hòn	khứ hồi
do vì	đều	hãy	Hơn	khướt
do đó	đều đều	hắn	Hung	kia
do vậy	đi	hẳn là	húng hắng	kia mà
dở chừng	đích thị	hằng	huống	kì cùng
dù cho	đó	h ng	huống chi	kì thật
dù rằng	đó đây	hiện_nay	huống gì	kì thực
dùng	đoá	hầu như	huống hồ	kì tình
duy	đôi	hcm	huống nữa	kìa
dữ	đôi khi	hé	hừ	kìn kìn
dọ	đối với	hèn chi	hử	kỳ thật
đã	đổng	hèn gì	hứ	kỳ thực
đại để	đột nhiên	hèn nào	hừm	kỳ tình
đại loại	đơ	hén	Ít	lá
đại nhân	đúng	hê	Kém	là
đại phàm	đúng là	hề	kẻo mà	lạ lùng
đạt	đúng như	hễ	kẻo nữa	lại
đoạn	đúng thật	hềnh hệch	kẻo rồi	lại thế nữa
đang	đúng thật là	hết	kế đến	làm
đáng lẽ	đúng vậy	hết mình	kế tiếp	làm sao
đáng lí	đùng đùng	hết mực	kha khá	lắm

đáng lý	đưa	hết nước	Khá	lắm lắm
đành đạch	được	hết nước hết cái	khả dĩ	lắm lúc
đánh đùng	eo ôi	hết sảy	khẩu	lắm khi
đáo để	êu	hết sức	khi nãy	lẳng lặng
đặc cách	gặp	hết thảy	khi đó	lầm lụi
đăm đắm	gì	hết ý	khi ấy	làn
đằng ấy	gia dĩ	hình	Khi	lần hồi
đâu	giả như	hình như	khi không	lần lượt
đâu đâu	giả phỏng	hiếm	khiến	lập tức
đâu đây	giả tỉ như	hiện	khó	lâu lâu
đâu đấy	giả tỷ như	hiện nay	khôn cùng	lẻn
đâu đó	giá mà	hiện tại	khôn xiết	lén
đầu	giá mà	hò khoan	không biết	leo lẻo
đầu tiên	giá như	hoài của	không có	lẽo đẽo
đây	giá phỏng	hoặc	không dưng	lêu
đây đó	giờ	hoặc giả	không khéo	lí láu
đấy	giờ đây	hoặc là	không mấy khi	lịa
để	giữa	hồ dễ	không những	lóc cóc
để mà	giữa chừng	hộc tốc	không phải	lọc cọc
long lóc	năng	nhỉ	phải chi	rất mực
lông lốc	nằng nặc	nhiều	phải chăng	ren rén
lũ lượt	nẫy	nhiên hậu	phăn phắt	rén
lui lủi	nấy	nhiệt liệt	phắt	rích
lùi lũi	nên	nhóm	phè	riệt
lùi lụi	nên chi	nhón nhén	phỉ phui	riu ríu
lủi thủi	nền	nhỡ ra	pho	rón rén
luôn	nếu	nhung nhăng	phóc	rồi
luôn luôn	nếu như	như	phỏng	rốt cục
luôn thể	ngay	như chơi	phỏng như	rốt cuộc
luôn tiện	ngay cả	như không	phót	rút cục
lý láu	ngay lập tức	như quả	phốc	rứa
mãi	ngay lúc	như thể	Phụt	sa sả
mãi mãi	ngay khi	như tuồng	phương chi	sạch
mặc dầu	ngay từ	như vậy	phứt	sao
mặc dù	ngay tức khắc	nhưng	qua quít	sau chót
mặc dù vậy	ngày càng	nhưng mà	qua quýt	sau cùng

mặc nhiên	ngày ngày	những	quả	sau cuối
mặc sức	ngày xưa	những ai	quả đúng	sau đó
mặc tình	ngày xửa	những như	quả là	sắp
mặt	ngăn ngắt	nhược bằng	quả tang	sất
màn	nghe chừng	nlđ	quả thật	sẽ
mất	nghe đâu	nó	quả tình	sì
mấy	nghen	nóc	quả vậy	so
mầy	nghiễm nhiên	nọ	quá	so_sánh
mậy	nghỉm	nổi	quá chừng	song le
mèm	ngõ hầu	nớ	quá độ	số là
miễn là	ngoải	nữa	quá đỗi	sống
miễn sao	ngoài	nức nở	quá lắm	sốt sột
min	ngôi	oai oái	quá sá	sở dĩ
mình	ngọn	oái	quá thể	suýt
mọi	ngọt	ô hay	quá trời	sự
món	ngộ nhỡ	ô hô	quá ọi	tà tà
mô phật	ngọơi	ô kê	quá xá	tại
mô tê	nhau	ô kìa	quý hồ	tại vì
một	nhân dịp	ồ	quyển	tấm
một mực	nhân tiện	ôi chao	quyết	tấn
một phép	nhất	ôi thôi	quyết nhiên	tự
một số	nhất đán	ối dào	ra	tự vì
mỗi	nhất định	ối giời	ra phết	tanh
mới	nhất loạt	ối giời ơi	ra trò	tăm tắp
muốn	nhất luật	ôkê	ráo	tắp
mựa	nhất mực	ổng	ráo trọi	tắp lự
nào	nhất nhất	ơ	rày	tất cả
nào là	nhất quyết	ơ hay	răng	tất tần tật
nay	nhất sinh	ơ kìa	rằng	tất tật
này	nhất tâm	ờ	rằng là	tất thảy
nãy	nhất tề	ớ	rất	tênh
nãy giờ	nhất thiết	ở	rất chi là	tha hồ
năm thì mười	nhé	ơi	rất đỗi	thà
thà là	thỏm	trên	úi dào	vô luận
thà rằng	thọt	trển	ư	vô vàn
thái quá	thốc	trệt	ứ hự	vốn dĩ

than ôi	thốc tháo	trếu tráo	ứ ừ	với lại
thanh	thộc	trệu trạo	ử	vở
thành ra	thôi	trong	ừ	vung tàn tán
thành thử	thốt	trỏng	và	vung tán tàn
thảo hèn	thốt nhiên	trời đất ơi	vả chăng	vung thiên địa
thảo nào	thuần	trừ phi	vả lại	vụt
thậm	thục mạng	tp	vạn nhất	vừa
thậm chí	thúng thắng	tphcm	văng tê	vừa mới
thật lực	thửa	tù tì	vẫn	xa xả
thật vậy	thực ra	tuần	vâng	xăm xăm
thật ra	thực vậy	tuần tự	vậy	xăm xắm
thẩy	thương ôi	tuốt luốt	vậy là	xăm xúi
thêm	tiện thể	tuốt tuồn tuột	vậy thì	xềnh xệch
thế	tiếp đó	tuốt tuột	veo	xệp
thế à	tiếp theo	tuy	veo veo	xin
thế là	tít mù	tuy nhiên	vèo	xiết bao
thế mà	tỏ ra	tuy rằng	về	xoành xoạch
thế nào	tỏ vẻ	tuy thế	vì	xoẳn
thế nên	tò te	tuy vậy	vì chọng	xoét
thế ra	toà	tuyệt nhiên	vì thế	xoẹt
thế thì	toé khói	từ	vì vậy	xon xón
thếch	toẹt	từ_tốn	ví bằng	xuất kì bất ý
thi thoảng	tọt	từng	ví dù	xuất kỳ bất ý
thì	tốc tả	tức thì	ví phỏng	xuể
thình lình	tôi	tức tốc	ví thử	ý chừng
thỉnh thoảng	tối ư	tựu trung	vị tất	ý da
thoạt	tông tốc	ủa	vn	
thoạt nhiên	tột	úi	vô hình trung	
thoắt	tràn cung mây	úi chà	vô kể	