BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO HỌC PHẦN MÔN KHAI PHÁ DỮ LIỆU

Đề tài : Từ tập dữ liệu cho sẵn xây dựng mô hình phân lớp tự động để từ đó xác định và đánh giá các tập dữ liệu khác

NHÓM SINH VIÊN THỰC HIỆN :

- **NGHIÊM THỊ QUỲNH HOA : 19021278 - K64 T-CLC**
- **BÙI HOÀNG NAM : 19021334 - K64 T-CLC**
- **HOÀNG BẢO PHÚC : 19021344 - K64 T-CLC**
- **NGUYỄN CÔNG THÀNH : 19021368 - K64 T-CLC**

HỌC PHẦN : KHAI PHÁ DỮ LIỆU - 2122I_INT3209E_20

Hà Nội 2021

Mục lục

Giới thiệu chung về báo cáo và mục tiêu

Bài báo cáo này được soạn ra để tổng hợp lại những vấn đề mà nhóm chúng em đã tìm hiểu trong quá trình thực hiện bài tập lớn về chủ đề "Topic detection". Cấu trúc của báo cáo sẽ theo dạng một khối thống nhất, đi từng bước một giúp quá trình theo dõi, đánh giá về nội dung dễ dàng hơn. Bài báo cáo sẽ giới thiệu về khâu tiền xử lý để làm sạch dữ liệu trước khi tiến hành khai phá, sau đó áp dụng các phương pháp, giải thuật khác nhau để tiến hành khai phá dữ liệu và đưa ra được mô hình phân loại. Cuối cùng là áp dụng mô hình vừa làm được vào các thử nghiệm, kiểm thử để kiểm tra độ tin cậy, chính xác của mô hình.

"Topic detection" là tập dữ liệu gồm 16000 các quảng cáo phân loại theo nhiều chủ đề như nhà đất, mua sắm, ăn uống,... Dữ liệu gốc là một file text, mỗi đoạn tương ứng với 1 quảng cáo. Mục tiêu của báo cáo khai phá dữ liệu này là từ tập dữ liệu trên (Training Data), phát triển một mô hình phân loại có độ chính xác cao để từ đó áp dụng mô hình vừa xây dựng vào việc tự động phân loại, gán nhãn cho các quảng cáo khác (Testing Data).

Trong quá trình soạn báo cáo do giới hạn về thời gian và các thành viên trong nhóm chưa có nhiều kinh nghiệm về chủ đề này nên dù đã cố gắng hết sức thì cũng không tránh khỏi sai sót, mong thầy thông cảm.

Phần I : Tổng quan về khai phá dữ liệu và các công cụ hỗ trợ

Nội dung trọng tâm của phần này sẽ là trình bày các khái niệm của khai phá dữ liệu, các bước cơ bản của quá trình khai phá và ứng dụng của nó.

1. Khái niệm khai phá dữ liệu

Khai phá dữ liệu (Data mining) là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu, từ đó rút ra được các tri thức quan trọng. Các mô hình khai phá dữ liệu thường được các doanh nghiệp sử dụng, từ đó cho phép họ có thể dự đoán được các xu hướng tương lai, đánh giá tâm lý khách hàng,…

2. Các bước cơ bản trong quy trình khai phá dữ liệu

Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Tuy nhiên, nó sẽ luôn bao gồm 7 bước cơ bản sau đây :

- Bước 1: Làm sạch dữ liệu – Data cleaning

Làm sạch dữ liệu là bước đầu tiên trong quá trình khai phá. Nó đóng vai trò quan trọng bởi lẽ trong thực tế tập dữ liệu đầu vào không phải lúc nào cũng sạch. Nếu dữ liệu không được làm sạch có thể dẫn đến việc quá trình khai khá bị ảnh hưởng, kết quả đem lại sẽ không chính xác

Về cơ bản, có rất nhiều phương pháp khác nhau được sử dụng để làm sạch tùy theo loại dữ liệu đầu vào. Đó có thể là lọc các dữ liệu thừa, trống hay loại bỏ các từ khóa không cần thiết

- Bước 2 : Tích hợp dữ liệu – Data integration

Trong quá trình khai phá dữ liệu, thường sẽ có nhiều bảng dữ liệu, file dữ liệu. Việc kết hợp các dữ liệu này lại với nhau được gọi là tích hợp dữ liệu. Việc này giúp tăng độ chính xác và tốc độ khai phá dữ liệu.

- Bước 3 : Lựa chọn dữ liệu đầu vào – Data reduction

Trong bước này, dữ liệu được trích suất từ cơ sở dữ liệu.

- Bước 4 : Chuyển đổi dữ liệu : Data transformation

Trong bước này, dữ liệu sẽ được chuyển đổi về dạng phù hợp hơn để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.

- Bước 5 : Khai phá dữ liệu : Data mining

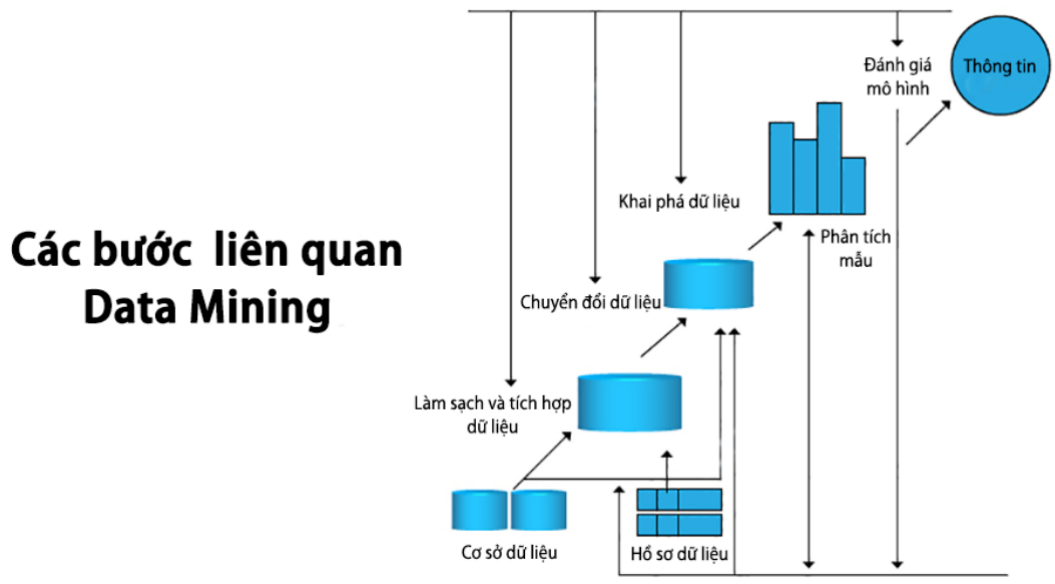
Khai phá dữ liệu là bước xác định và rút ra được các thông tin hữu ích từ một tập dữ liệu lớn. Trong bước này ta sẽ áp dụng các phương pháp và kĩ thuật khác nhau để lấy ra được thông tin cần dùng. Các thông tin này thường sẽ được lưu trữ dưới dạng các model để có thể áp dụng cho sau này.

- Bước 6 : Đánh giá mẫu : Pattern evaluation

Ta sẽ đánh giá các mẫu rút ra được từ tập dữ liệu và đánh giá chúng dựa trên các thông số. Các thông số này thường được trình bày dưới dạng bảng, sơ đồ để giúp người dùng dễ hiểu hơn

- Bước 7 : Trình bày thông tin : Knowledge presentation

Đây là bước cuối cùng trong quá trình khai phá dữ liệu. Từ các bước trên ta sẽ rút ra được mô hình khai phá dữ liệu hoàn chỉnh, để từ đó áp dụng vào thực tiễn



Hình 1.2.1 : Tóm tắt quy trình khai phá dữ liệu

3. Ứng dụng của khai phá dữ liệu

Có nhiều ứng dụng của Data Mining thường thấy như:

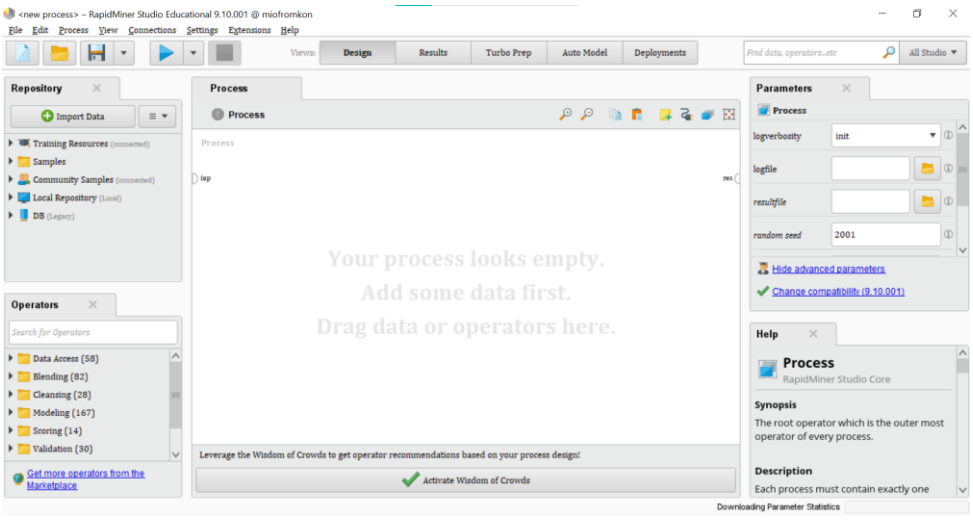
- Phân tích thị trường và chứng khoán
- Phát hiện gian lận
- Quản lý rủi ro và phân tích doanh nghiệp
- Phân tích giá trị trọn đời của khách hàng

Đây là những ứng dụng vô cùng thực tiễn của việc khai phá dữ liệu, nó giúp các doanh nghiệp có cái nhìn tổng quan, đánh giá được thị trường cũng như là đưa ra các dự đoán có độ chính xác cao về xu hướng thị trường trong thời gian sắp tới.

4. Sử dụng công cụ Rapid Miner trong việc khai phá dữ liệu

Việc sử dụng các công cụ phần mềm giúp tự động hóa nhiều phần trong quá trình khai phá dữ liệu, tiết kiệm nhiều thời gian và hơn nữa thông tin được trình bày dưới dạng dễ hiểu, thân thiện với người dùng.

Rapid Miner là một trong những công cụ phổ biến nhất để khai phá dữ liệu. Hơn nữa, nó cung cấp các chức năng khai thác dữ liệu khác nhau như tiền xử lý dữ liệu, biểu diễn dữ liệu, lọc, phân cụm, v.v.



Hình 1.4.1 : Giao diện chính của Rapid Miner

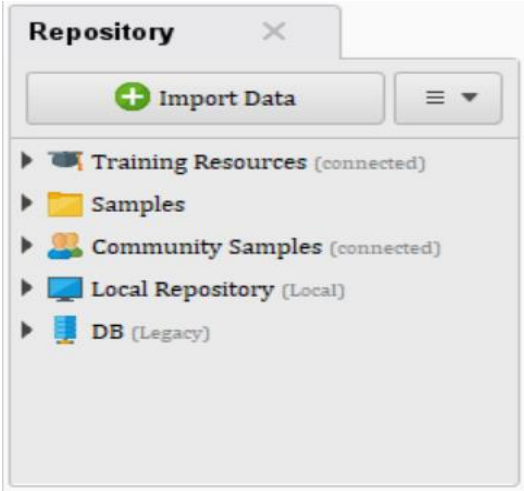
Trên đây là giao diện màn hình chính của RapidMiner. Nó gồm 4 thành phần chính

Giao diện Repository : Từ đây ta có thể lưu trữ, nhập, xuất các dữ liệu, models, process để có thể sử dụng cho sau này. Ngoài ra Rapid Miner còn hỗ trợ 1 kho dữ liệu mẫu và các tài liệu hướng dẫn để người dùng mới có thể tự tìm hiểu và làm quen với ứng dụng.

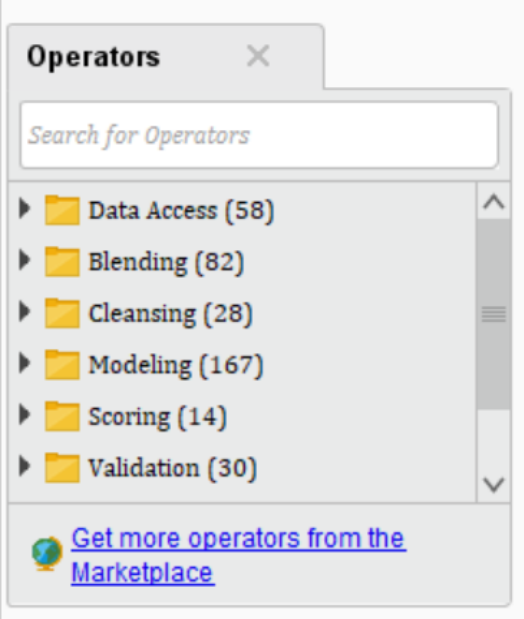
Giao diện Operators : Từ giao diện này ta có thể thấy các phương pháp, kĩ thuật thường được sử dụng trong việc khai phá dữ liệu, chúng được sắp xếp, phân loại tùy theo mục đích sử dụng và loại dữ liệu mà ta sẽ khai phá. Mỗi operator sẽ có 1 chức năng, kiểu đầu vào và đầu ra khác nhau. Ngoài ra, ta cũng có thể tải thêm các operators tùy chọn khác được người dùng tải lên từ thư viện Marketplace. Các operators được sử dụng đơn giản bằng cách kéo thả vào giao diện Process

Giao diện Process : Đây là nơi mà người dùng sẽ thực hiện chủ yếu các bước của khai phá dữ liệu. Ban đầu giao diện này sẽ trống, chỉ có 2 đầu đó là inp-dữ liệu đầu vào và res-kết quả đầu ra thu được. Khi sử dụng, người dùng sẽ kéo thả các operators mình muốn vào giao diện này, kết nối các đầu vào và đầu ra phù hợp để từ đó Process hoạt động.

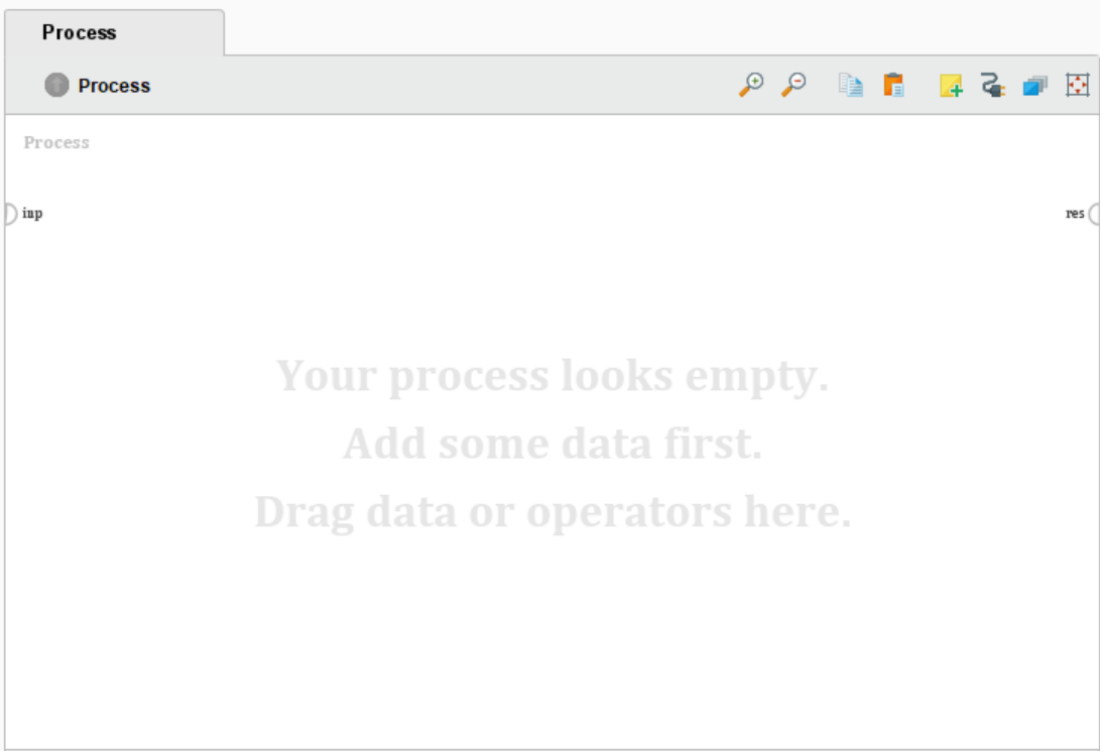
Giao diện Parameters : Như đã nói ở trên, mỗi Operators sẽ có 1 chức năng và thông số riêng. Các thông số này có thể được thay đổi cho phù hợp với mục đích sử dụng. Chính sự tùy chỉnh này vừa giúp Rapid Miner dễ dàng sử dụng, lại không hề bị gò bó.



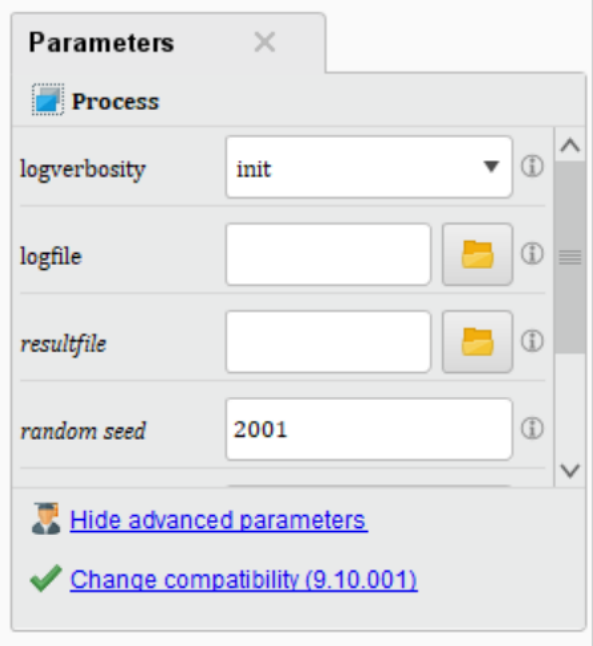
Hình 1.4.2 : Giao diện Repository



Hình 1.4.3 : Giao diện Operators



Hình 1.4.4 : Giao diện Process



Hình 1.4.5 : Giao diện Parameters

Phần II : Sử dụng thuật toán K-NN (K-Nearest Neighbor) trong việc phân loại dữ liệu

Bài toán phân loại dữ liệu là một trong những bài toán thường gặp trong cuộc sống và kĩ thuật, có rất nhiều cách tiếp cận và giải thuật được đưa ra để giải quyết bài toán phân lớp. Các kĩ thuật tiêu biểu thường được sử dụng là Decision Trees, K-NN và Naive Bayes. Với báo cáo này, nhóm chúng em sẽ sử dụng phương pháp K-NN

1. Tổng quan về K-NN

Thuật toán K-Nearest Neighbor (viết tắt là K-NN) là thuật toán có mục đích phân loại lớp cho một mẫu mới (Testing Data) dựa trên các thuộc tính và các mẫu sẵn có (Training Data) , các mẫu này được nằm trọng một hệ gọi là không gian mẫu.

Một đối tượng được phân lớp dựa vào K gần nhất của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng với mẫu mới, sau đó chuẩn đoán mẫu mới thuộc phân lớp nào dựa vào số K láng giềng xác định trước có khoảng cách gần mẫu mới nhất so với các mẫu khác.

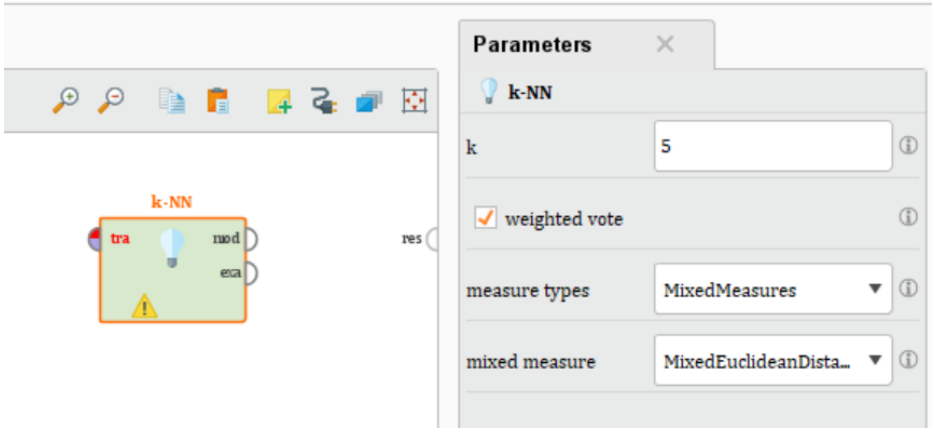
2. Mô tả về thuật toán K-NN

Các mẫu được mô tả bằng n – chiều thuộc tính số. Mỗi mẫu đại diện cho một điểm trong một chiều không gian n – chiều. Theo cách này tất cả các mẫu được lưu trữ trong một mô hình không gian n – chiều.

Các bước thực hiện của thuật toán K-NN được mô tả như sau :

- Xác định tham số K (số láng giềng gần nhất) với giá trị càng lớn thì độ chính xác càng cao, nhưng sẽ tỉ lệ thuận với thời gian chạy (runtime).
- Tính khoảng cách giữa đối tượng cần phân lớp (Testing Data) với tất cả các đối tượng trong các mẫu có sẵn (Trainning Data) (Thường sử dụng khoảng cách Euclidean).
- Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Testing data.
- Lấy tất cả các lớp của K láng giềng gần nhất đã xác định.
- Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Testing Data.

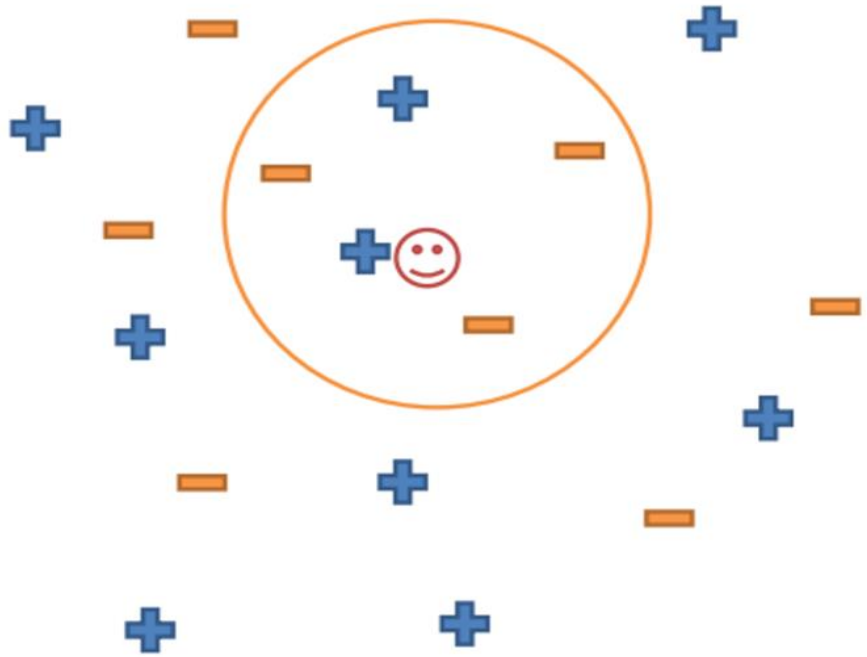
Trong Rapid Miner, K-NN có thể được tìm và sử dụng dưới dạng 1 Operator, với đầu vào là tập dữ liệu training và đầu ra là một model để có thể sử dụng sau này. Ta có thể điều chỉnh thông số K (số láng giềng gần nhất thông qua bảng Parameters) :



Hình 2.2.1 : Operator K-NN trong RapidMiner

3. Minh họa về thuật toán K-NN

Trong hình dưới đây, Trainning Data được mô tả bằng dấu (+) và dấu (-) tương đương với 2 lớp, đối tượng cần được xác định lớp cho nó (Testing Data) là hình mặt cười đỏ. Nhiệm vụ của ta là ước lượng lớp của Testing Data dựa vào việc lựa chọn số láng giềng gần nhất với nó. Nói cách khác ta muốn biết liệu Testing Data sẽ được phân vào lớp (+) hay lớp (-).



Hình 2.3.1 : Ví dụ minh họa về thuật toán K-NN

Từ hình trên, ta có thể thấy rằng Testing Data có 5 láng giềng, trong đó có 3 lớp (-) và 2 lớp (+). Từ đó ta kết luận rằng Testing Data thuộc lớp (-) vì lớp (-) có nhiều đối tượng hơn lớp (+).

4. Đánh giá ưu, nhược điểm của thuật toán K-NN

- Ưu điểm :

+ Đơn giản, dễ hiểu, dễ cài đặt và sử dụng.

+ Việc dự đoán kết quả của dữ liệu mới rất đơn giản.

- Nhược điểm :

+ Thời gian chạy lâu khi tập dữ liệu đầu vào lớn.

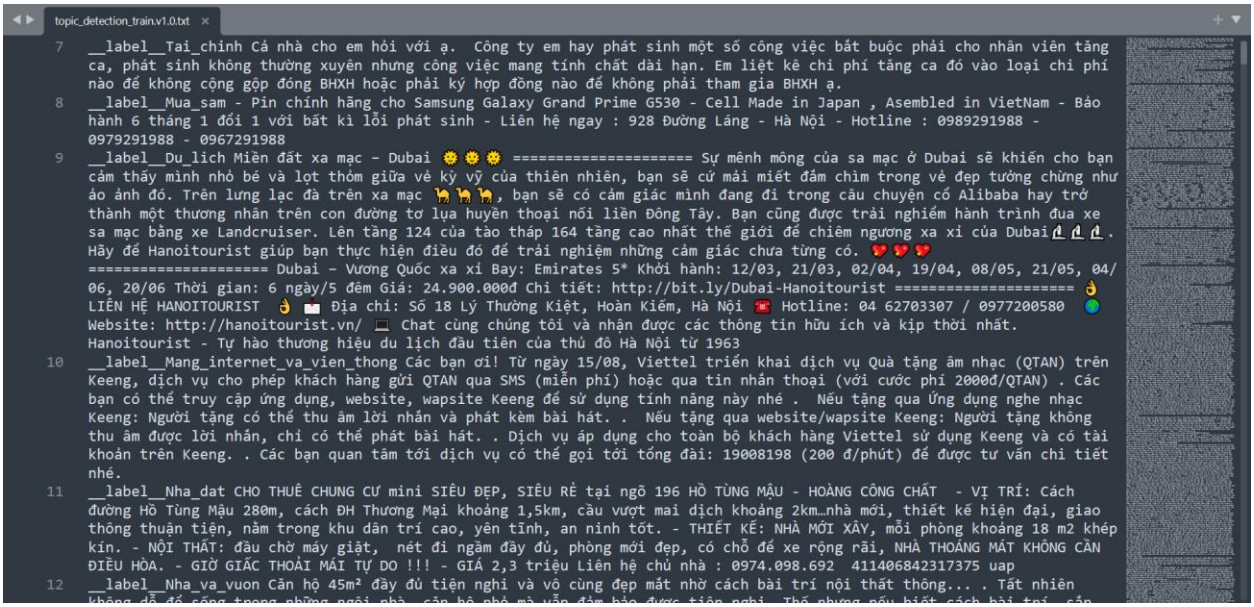
+ Độ chính xác có thể thấp với K đầu vào nhỏ (trung bình thì K=5)

Phần III : Xây dựng mô hình khai phá dữ liệu

Ta sẽ bắt đầu khai phá tập dữ liệu "Topic detection" với quy trình các bước như đã trình bày ở phần I với sự hỗ trợ của công cụ RapidMiner

1. Làm sạch dữ liệu

Khi giải nén tập dữ liệu được cho, ta sẽ có được 1 file text có dạng như sau :



Hình 3.1.1 : Tập dữ liệu đầu vào

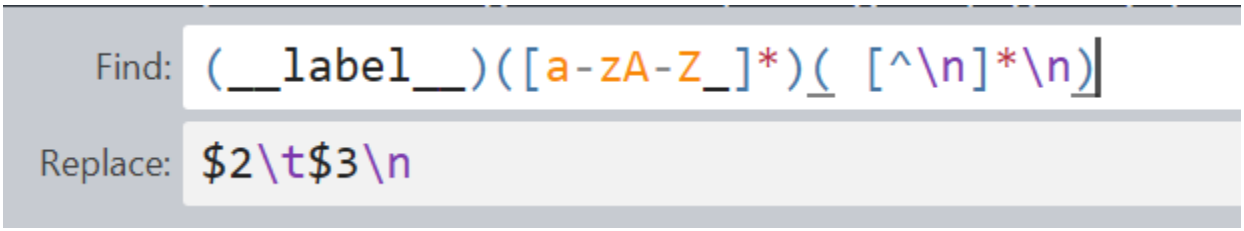
Ta có thể thấy các vấn đề cần xử lí như sau :

+ Đây là một file text chứ không phải bảng dữ liệu nên cần phải chuyển đổi để các phần mềm có thể đọc được.

+ Dữ liệu đầu vào chứa những kí tự không cần thiết. Bởi lẽ đây là bài toán phân lớp

chủ yếu dựa vào text nên những thứ như icon, kí tự đặc biệt, số điện thoại, link, hashtag,… là không cần thiết, do đó cần phải được loại bỏ.

Khi đã nắm rõ được các vấn đề trên, ta bắt tay vào việc làm sạch dữ liệu. Ở đây nhóm em sẽ chủ yếu sử dụng regex để lọc kí tự không cần thiết và biến đổi dữ liệu về dạng bảng như sau :



Hình 3.1.2 : Lọc dữ liệu thành dạng bảng

Ta chia các đoạn text thành 2 phần topic và content, ngăn cách nhau bởi dấu "Tab". Khi đó khi chúng ta sao chép đoạn text và đưa vào excel, nó sẽ tự động chia thành 2 cột như hình dưới

	A	B
1	Topic	Content
2	Du_lich	Theo hành trình tour du lịch Mỹ Bờ Đông, du khách của Lữ hành Saigontourist sẽ đến New York giấc mơ được gọi tên của hàng triệu người chinh phục du khách bằng tượng đài Nữ thần Tự do duyên dáng trên vịnh Hudson, trung tâm Manhattan trụ sở tài chính lớn nhất thế giới với phố Wall, khu Rockeffeler...Và thủ đô Washington, dạo bước trên National Mall, du khách sẽ bị choáng ngợp bởi các công trình kỳ vĩ nơi đây. Từ điện Capitol, tượng đài Lincoln, đài tưởng niệm Washington... cho đến Nhà Trắng, mỗi công trình đều là một tuyệt tác kiến trúc được phối cảnh hài hòa khiến cho Washington DC từ lâu đã được ghi nhận là một trong những thành phố đẹp nhất Hợp chủng quốc Hoa Kỳ. Đặc biệt hơn khi đi du lịch Mỹ, đến bất kỳ thành phố nào vào bất cứ thời gian nào, du khách cũng có thể trải nghiệm thú vui mua sắm bất tận
3	Nha_dat	mình cần tìm 1 phòng cho khoảng 3 người quanh khu vực hồ tùng mậu. phòng khép kín có điều hòa .Mng nếu quen ai cho thuê thì gthieu giúp mình . mình cảm ơn !! Nếu hợp lí thì mình chuyển luôn trong chủ nhật hoặc thứ 2.
4	Nha_dat	Cho thuê nhà riêng dt 60m/sàn. Có 4 phòng ngủ. 1p khách thoáng mát an ninh tốt. Nhà ngõ 401 cố nhuế gần đại học mỏ .hv tài chính hn. Lh A Hoàng
5	Nha_dat	Cho thuê nhà ở tầng 4 khép kín, 4/295 Nguyễn Khoái Có bếp riêng đầy đủ tủ lạnh, lò vi sóng, bếp gas. có máy giặt sân phơi. Điện nước dân dụng chia đầu người Giowf giấc thoải mái Đóng tiền 3 tháng 1 lần đặt cọc 1 tháng (2.5tr / tháng) Liên hệ: inb hoặc Miễn trung gian
6	Mua_sam	Crumpler jackpack full photo giá : 800.000 vnđ giảm 20 % còn : 650.000 vnđ Đựng vừa laptop 15 " Màu sắc : Như Hình Cân nặng 1,2 kg Kích thước ba lô 28.0 x 45.0 x 14.5 Tải trọng tối đa 17 kg Chất liệu vải 420d chống thấm tuyết đối của crumpler Ba lô có 1 ngăn chống sock để vừa laptop 15 " có 1 ngăn riêng được thiết kế 1 cái hộp riing để chống sock cho máy ảnh , lens ba lô có thể để vừa 2 body và 1 lens 700/200 và 4 lens tầm trung Các đầu khóa kéo chắc chắn , có in logo độc quyền Crumpler Ngăn trước có nhiều ngăn nhỏ bên trong , ngăn sau đựng laptop có mút PU dày chống sốc Mặt sau có lớp mút Ari Mesh dày , thoát hơi nhanh , được thiết kế theo thể hình người Việt Nam Quai xách tay va quai đeo ôm sát vai làm cho tải trọng balo được chia đều cho cả lưng , vì vậy người mang cảm thấy rất nhẹ nhàng BẢO HÀNH : 1 năm tại BALOHPHUOT.COM Chi tiêt liên hệ : www.balophuot.com đt : hoặc
		Đăng hộ chủ nhà 58 ngõ 83 Đào Tấn gần Lotte, công viên Thủ Lệ 2 phòng tầng 2 trong nhà 5 tầng đang trống cần người thuê S: 1314m, có

Hình 3.1.3 : Dữ liệu dưới dạng bảng, có thể sử dụng

Tương tự, sau đó ta lọc và xóa các số liệu, kí tự không cần thiết sử dụng các câu lệnh regex như sau :

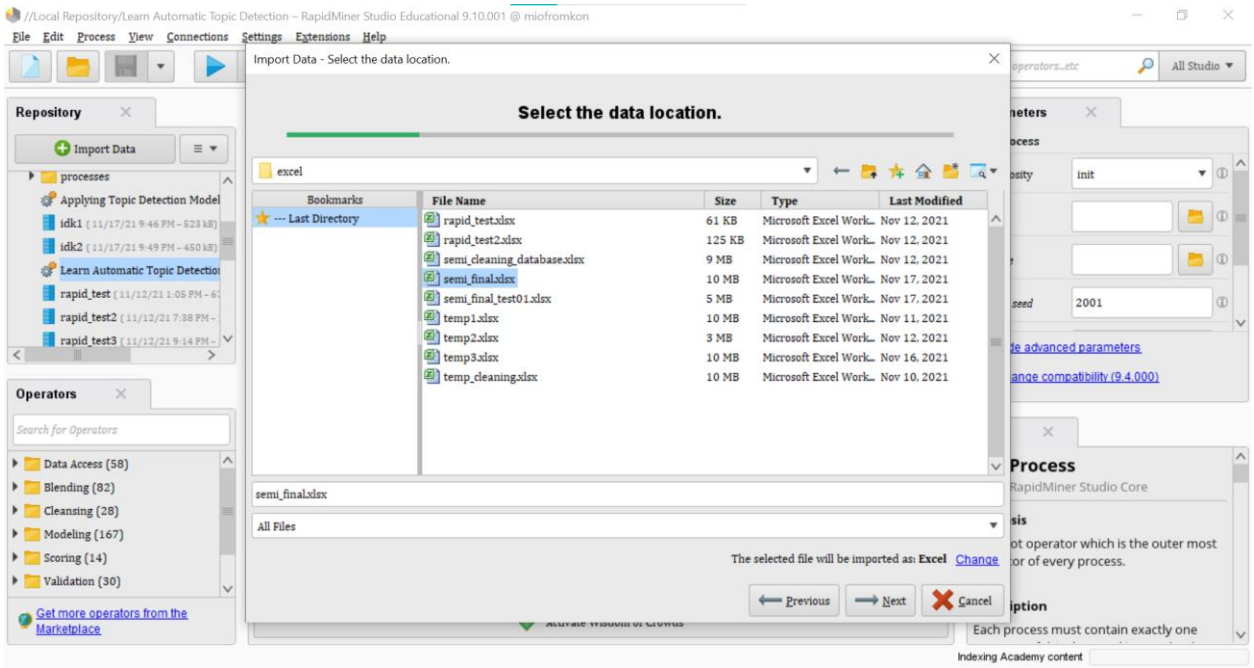
- (((09|03|07|08|05|01|02|04)+([\d.]{6, 11})\d\b)) : Lọc các số điện thoại ở Việt Nam
- (https?:\/\/(www\.)?[-a-zA-Z0-9@:%._\+~#=]{1,256}\.[a-zA-Z0-9()]{1,6}\b([-a-zA-Z0-9()@:%_\+.~#?&//=]*)) : Lọc các đường link http
- ((^|\B)#(?![0-9_]+\b)([a-zA-Z0-9_]{1,30})(\b|\r)) : Lọc các hashtag

Ta làm tương tự đối với các icon, kí tự đặc biệt, email , …

2. Tích hợp, lựa chọn dữ liệu và biến đổi dữ liệu

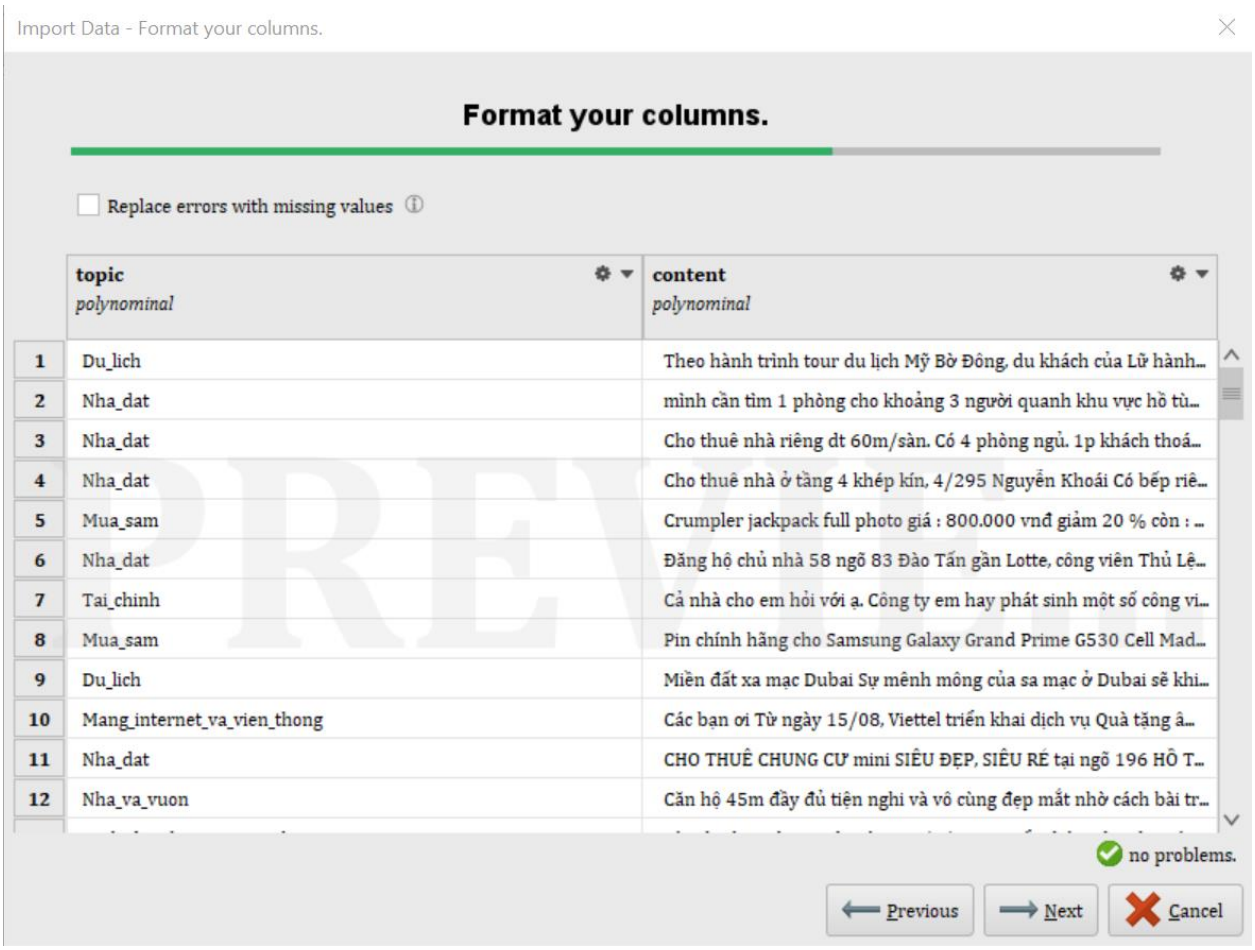
Sau khi đã làm sạch qua dữ liệu, ta có thể bắt đầu import dữ liệu vào Rapid Miner và tiến hành khai phá.

Ta có thể import bằng cách từ giao diện Repository, chọn Import Data và chọn bảng dữ liệu được lưu trữ trên máy.



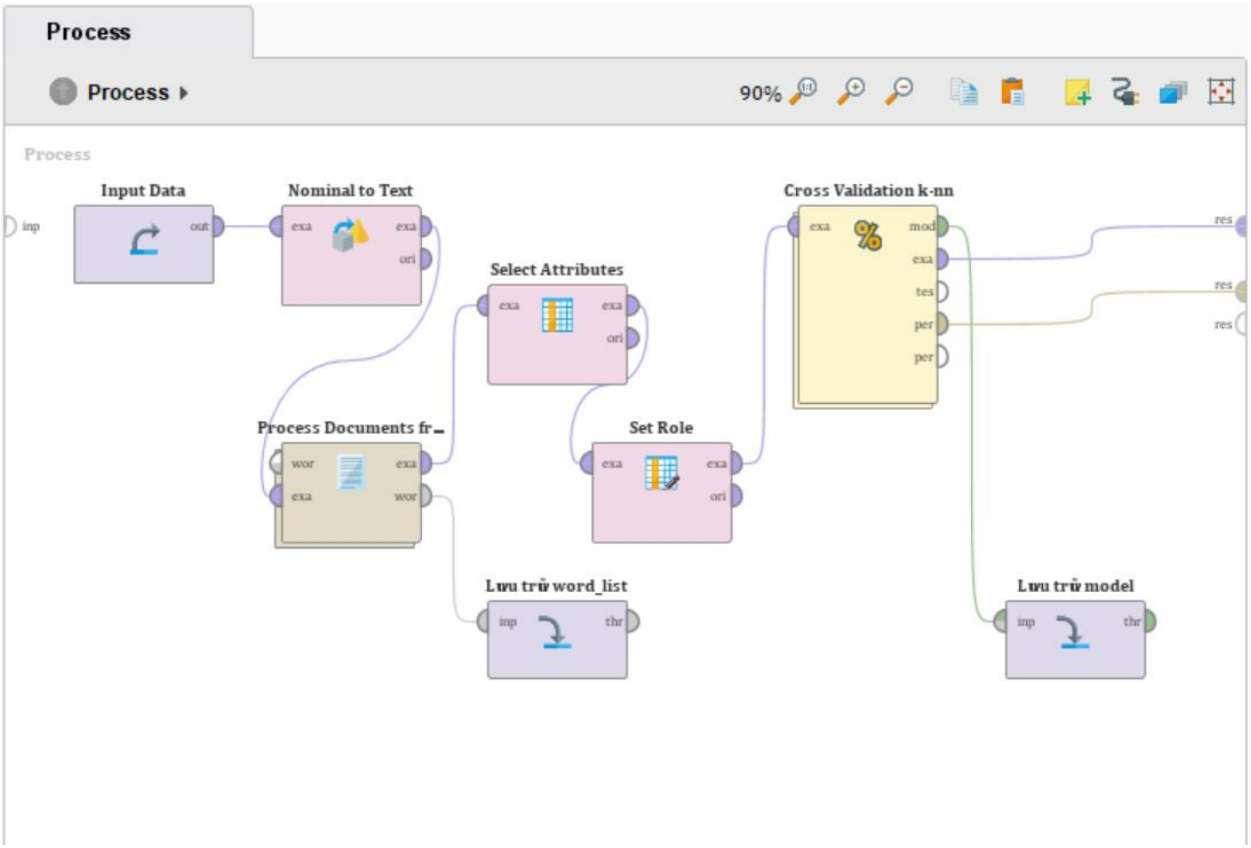
Hình 3.2.1 : Import database vào RapidMiner

Kết quả thu được sẽ là 1 bảng với 2 cột là topic và content :



3. Khai phá dữ liệu

Sau khi đã nhập dữ liệu xong, ta tiến hành khai phá để xây dựng model tự động phân loại. Để thuận tiện, nhóm em sẽ gửi toàn bộ quá trình sau đó sẽ giải thích từng bước :

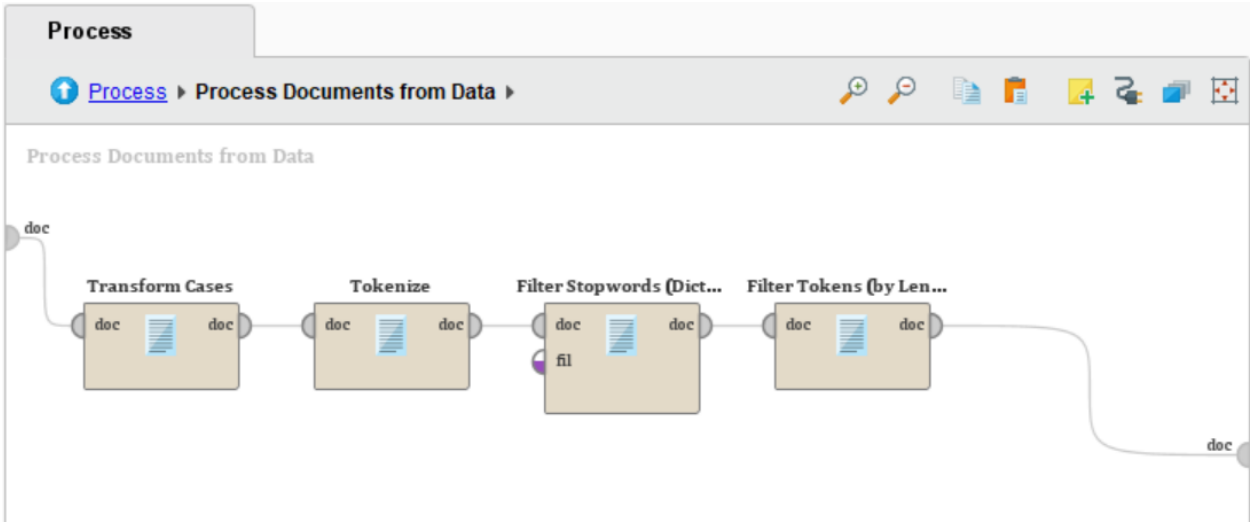


Hình 3.3.1: Quá trình xây dựng mô hình phân loại

+ Đầu tiên, ta sử dụng Operator Retrieve để nhập vào bảng dữ liệu Training vừa được Import.

+ Tiếp đến, ta biến đổi dữ liệu của bảng về dạng text thay vì dạng polynominal như mặc định. Có 2 cách để thay đổi đó là sử dụng Operator hoặc tự Edit bảng dữ liệu

+ Tiếp sau đó sẽ là xử lý dữ liệu. Đây là 1 tập gồm các bước khác nhau để xử lí dữ liệu đã được làm sạch.



Hình 3.3.2 : Các bước để xử lí dữ liệu

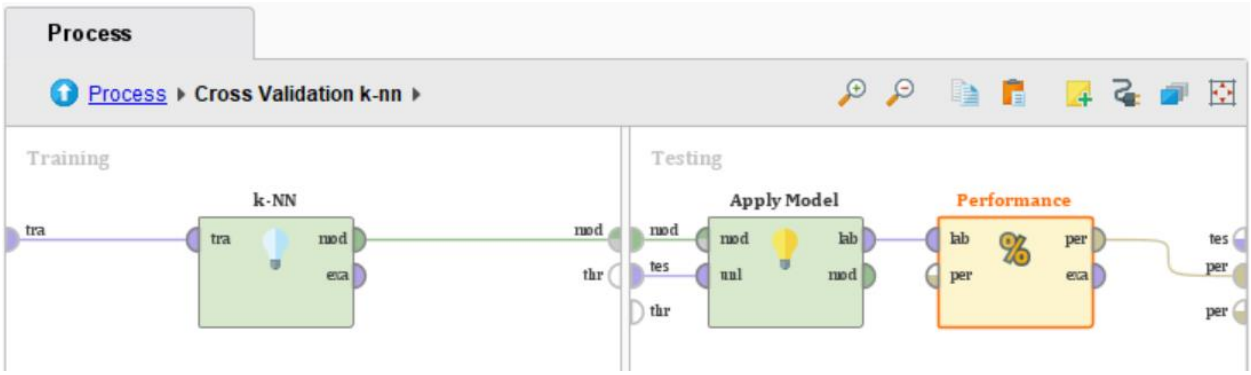
- Transform Cases : Biến đổi tất cả text về in thường (Hoa => hoa,…)
- Tokenize : Tách đoạn text thành các dãy các token. Ở đây nhóm em chọn tokenize sử dụng các kí tự không phải là chữ (?, =, !, …), từ đó thu được dãy token là các từ tiếng Việt
- Filter Stopwords : Lọc các từ ít nghĩa hoặc không có giá trị sử dụng cho việc phân loại như ngôi xưng hô, các từ cảm thán,… Do RapidMiner không hỗ trợ Stopwords cho ngôn ngữ tiếng Việt, nên bọn em đã tham khảo file Stopwords tìm được trên github. Link github sẽ được đặt ở phần Tài liệu tham khảo
- Filter Tokens : Đây là bước cuối cùng trong quá trình xử lí dữ liệu. Khi tách đoạn dữ liệu thành các token, thường sẽ xảy ra những lỗi như token chỉ có 1 chữ hay token là một từ quá dài dẫn đến không hợp lệ. Việc filter này sẽ giúp giảm thiểu lỗi đó xảy ra.

+ Sau khi xử lí dữ liệu, ta sẽ lưu các token word dưới dạng bag-of-words để tránh mất thời gian xử lí lại dữ liệu khi áp dụng model lên một bảng dữ liệu mới.

+ Tiếp đến là Select Attributes : Ở đây ta chỉ lựa chọn những dữ liệu nào đủ, không bị thiếu sót, từ đó giúp giảm thiểu độ sai sót khi xây dựng model

+ Bước tiếp theo là Set Role : Bởi đây là bài toán phân loại, việc Set Role này sẽ giúp RapidMiner biết được đâu là cột cần được phân loại và dự đoán

+ Bước cuối cùng là sử dụng thuật toán K-NN để RapidMiner bắt đầu học quá trình tự động phân loại : Ta sẽ sử dụng Operator kiểm chứng chéo Cross Validation với K=10. Cross Validation sẽ xáo trộn tập dữ liệu 1 cách ngẫu nhiên, chia thành K phần. Sau đó sử dụng các nhóm hiện tại để huấn luyện mô hình và sử dụng các nhóm khác để đánh giá hiệu năng. Từ đó sẽ giúp ta có cái nhìn tổng quan về độ hiệu quả của model vừa được xây dựng. Cross Validation gồm 3 bước như sau :



Hình 3.3.3 : Quy trình tự động học phân lớp sử dụng K-NN

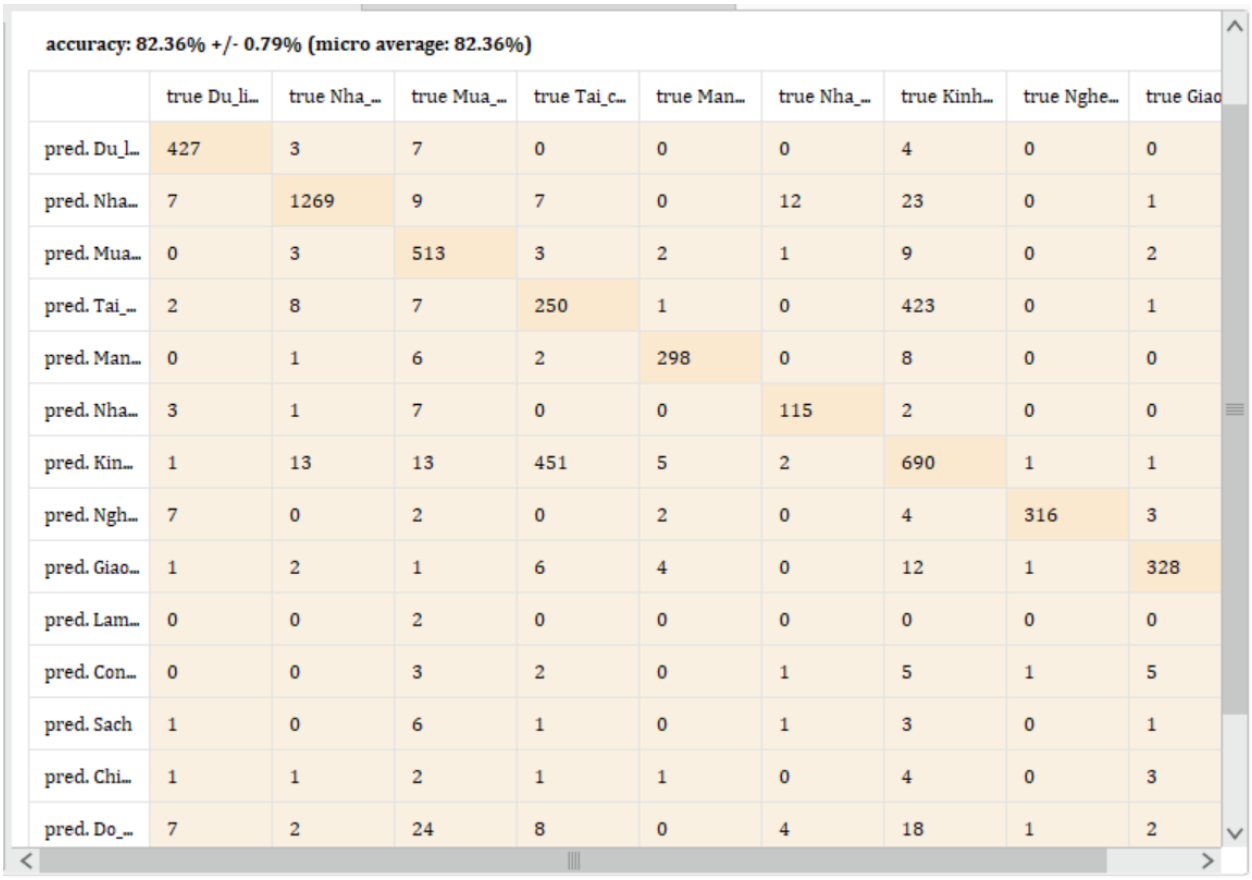
Với đầu vào là tập dữ liệu đã được xử lí (Training), sử dụng thuật toán K-NN với K = 10 ta sẽ ra được một model tự động phân lớp, sau đó RapidMiner sẽ sử dụng chính dữ liệu từ bảng Training để tiến hành kiểm thử (Testing) lên model vừa được xây dựng.

Sau đó ta sẽ lưu lại model vừa được xây dựng và trả về kết quả của quá trình học máy.

4. Đánh giá mẫu

Sau khi xây dựng xong mô hình tự động phân loại, ta lưu lại và bắt đầu chạy

Kết quả thu được sẽ là 1 bảng như sau :



accuracy: 82.36% +/- 0.79% (micro average: 82.36%)

	true Du_li...	true Nha_...	true Mua_...	true Tai_c...	true Man...	true Nha_...	true Kinh...	true Nghe...	true Giao
pred. Du_l...	427	3	7	0	0	0	4	0	0
pred. Nha...	7	1269	9	7	0	12	23	0	1
pred. Mua...	0	3	513	3	2	1	9	0	2
pred. Tai_...	2	8	7	250	1	0	423	0	1
pred. Man...	0	1	6	2	298	0	8	0	0
pred. Nha...	3	1	7	0	0	115	2	0	0
pred. Kin...	1	13	13	451	5	2	690	1	1
pred. Ngh...	7	0	2	0	2	0	4	316	3
pred. Giao...	1	2	1	6	4	0	12	1	328
pred. Lam...	0	0	2	0	0	0	0	0	0
pred. Con...	0	0	3	2	0	1	5	1	5
pred. Sach	1	0	6	1	0	1	3	0	1
pred. Chi...	1	1	2	1	1	0	4	0	3
pred. Do_...	7	2	24	8	0	4	18	1	2

Hình 3.4.1 : Kết quả thu được từ thuật toán K-NN với K = 10

Từ bảng kết quả trên ta có thể thấy độ chính xác trung bình của mô hình vừa được xây dựng là 82,36% với sai số là 0.79%. 2 chiều của bảng kết quả sẽ là Topic của bảng dữ liệu Training (true_topic) và kết quả mà model dự đoán (pred_topic). Số lượng dự đoán đúng của từng loại chủ đề sẽ được bôi đậm.

Thực hiện tương tự với K=9,8,7,6,5 ta được thu được kết quả như sau:

K	10	9	8	7	6	5
Độ chính xác	82,36%	82,78%	82,02%	82,45%	81,77%	82,05
Sai số	0.79%	0,66%	0.81%	0,61%	0.58%	0,82%

Khi lướt xuống phía dưới, ta sẽ thấy được độ chính xác của việc dự đoán đối với từng loại topic, với K = 10 nhóm em sẽ thống kê chúng dưới dạng bảng như sau :

Topic	Độ chính xác
Du_lich	93.03%
Nha_dat	97.39%
Mua_sam	84.51%
Tai_chinh	33.97%
Mang_internet_va_vien_thong	94.90%
Nha_va_vuon	84.56%
Kinh_doanh_va_cong_nghiep	56.74%
Nghe_thuat	98.44%
Giao_duc	93.54%
Lam_dep_va_the_hinh	90.96%
Con_nguoi_va_xa_hoi	92.55%
Sach	88.37%

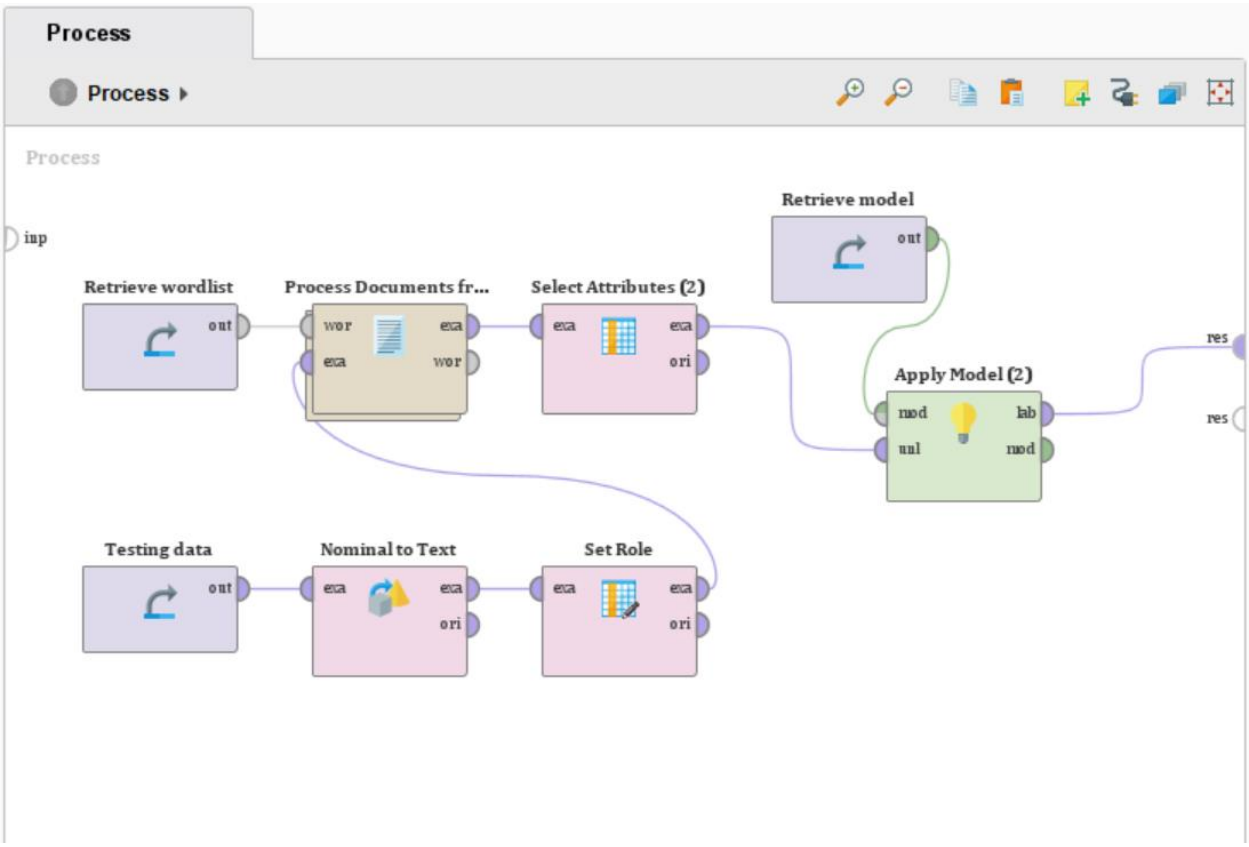
Topic	Độ chính xác
Chinh_tri	91.03%
Do_an_va_do_uong	98.54%
Giao_thong	96.88%
Thoi_quen_va_so_thich	92.86%
Giai_tri	100%
Suc_khoe	77.32%
Phap_luat	35.94%
Khoa_hoc	70.83%
May_tinh_va_thiet_bi_dien_tu	91.18%
Cong_nghe_moi	60.00%
The_thao	81.40%

Ta có thể thấy, những topic có độ dự đoán chính xác thấp là : Tài chính, Kinh doanh và công nghiệp, Pháp luật, Công nghệ mới. Việc không cân đối về tỉ lệ dự đoán chính xác giữa các topic với nhau có thể là do các topic này có nhiều keyword trùng lặp với nhau, dẫn đến việc thuật toán không thể dự đoán chính xác được. Có thể xử lí bằng cách tăng thông số K của thuật toán K-NN, nhưng điều này tương đương với việc thời gian chạy cũng sẽ tăng theo.

Phần IV : Áp dụng mô hình tự động phân loại

Sau khi đã lưu trữ được word_list và model tự động phân loại, giờ là bước ta áp dụng lên các tập dữ liệu test.

1. Xây dựng Process áp dụng model tự động phân loại



Hình 4.1.1 : Quy trình áp dụng model tự động phân loại

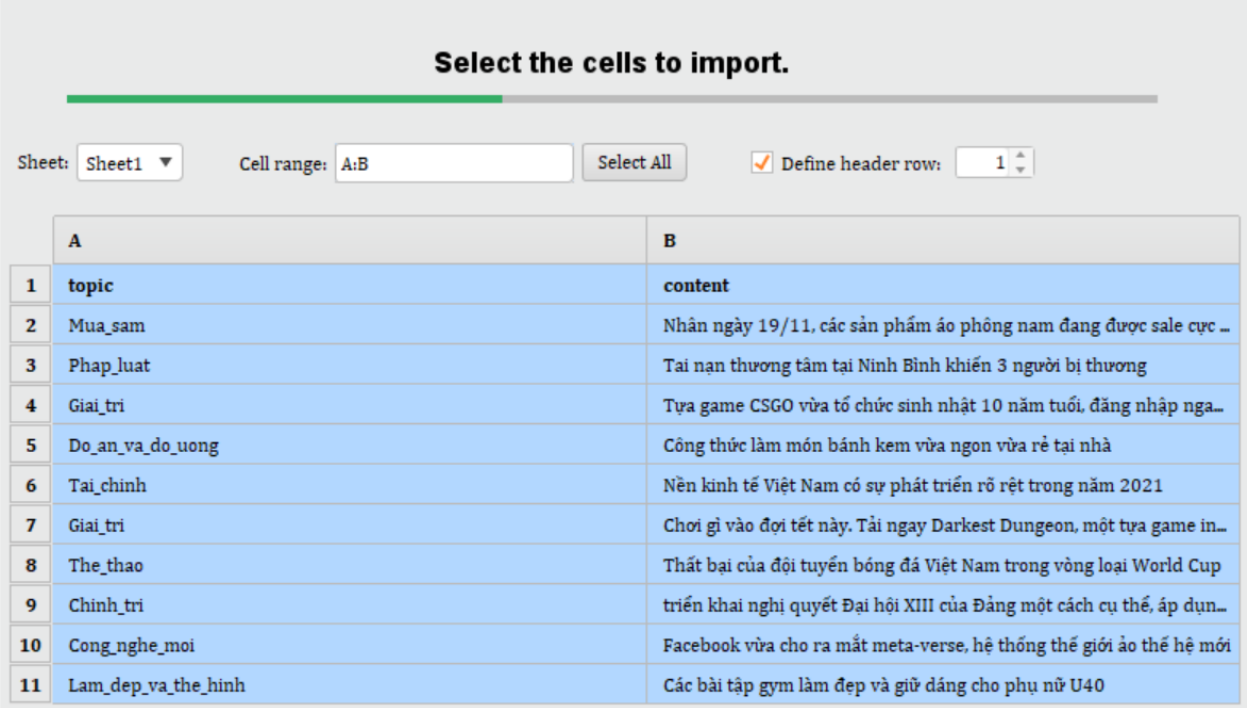
+ Với đầu vào là tập dữ liệu Test, ta vẫn áp dụng những bước như chuyển kiểu dữ liệu từ Nominal to Text, Set role cho cột topic để được phân loại và sau đó đem đi xử lí dữ liệu

+ Sau khi đã lựa chọn những dữ liệu không bị thiếu sử dụng Operator Select Attribute, ta bắt đầu áp dụng Model tự động phân loại đã được xây dựng lên tập dữ liệu Test

+ Sử dụng Operator Apply model, với đầu vào là tập dữ liệu Test và Model đã được xây dựng, đầu ra sẽ là dữ liệu được dự đoán bởi RapidMiner.

2. Tiến hành kiểm thử trên bảng dữ liệu Test

Ở đây, nhóm em đã tự chuẩn bị 1 tập dữ liệu Test gồm 2 cột là content và topic. Vì đây là khâu kiểm thử nên nhóm em vẫn giữ cột topic để so sánh với dự đoán của RapidMiner, còn khi áp dụng với tập dữ liệu khác ta có thể xóa hoặc để trống cột topic này. Tập dữ liệu sẽ có dạng như sau :



	A	B
1	topic	content
2	Mua_sam	Nhân ngày 19/11, các sản phẩm áo phông nam đang được sale cực ...
3	Phap_luat	Tai nạn thương tâm tại Ninh Bình khiến 3 người bị thương
4	Giai_tri	Tựa game CSGO vừa tổ chức sinh nhật 10 năm tuổi, đăng nhập nga...
5	Do_an_va_do_uong	Công thức làm món bánh kem vừa ngon vừa rẻ tại nhà
6	Tai_chinh	Nền kinh tế Việt Nam có sự phát triển rõ rệt trong năm 2021
7	Giai_tri	Chơi gì vào đợi tết này. Tải ngay Darkest Dungeon, một tựa game in...
8	The_thao	Thất bại của đội tuyển bóng đá Việt Nam trong vòng loại World Cup
9	Chinh_tri	triển khai nghị quyết Đại hội XIII của Đảng một cách cụ thể, áp dụn...
10	Cong_nghe_moi	Facebook vừa cho ra mắt meta-verse, hệ thống thế giới ảo thế hệ mới
11	Lam_dep_va_the_hinh	Các bài tập gym làm đẹp và giữ dáng cho phụ nữ U40

Hình 4.2.1 : Bảng dữ liệu Test

Sau khi Import thành công bảng dữ liệu, ta tiến hành khởi chạy. Kết quả thu được như sau :

Row No.	topic	prediction(t...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...	confidence(...
1	Mua_sam	Mua_sam	0	0	1	0	0	0
2	Phap_luat	Phap_luat	0	0	0	0.199	0	0
3	Giai_tri	Giai_tri	0	0	0	0	0	0
4	Do_an_va_do_u...	Do_an_va_do_u...	0	0	0	0	0	0
5	Tai_chinh	Kinh_doanh_va...	0	0.100	0	0.200	0	0
6	Giai_tri	Giai_tri	0	0	0	0	0	0
7	The_thao	The_thao	0	0	0.100	0	0.100	0
8	Chinh_tri	Chinh_tri	0	0	0	0	0	0.100
9	Cong_nghe_moi	Kinh_doanh_va...	0	0	0	0	0.100	0
10	Lam_dep_va_th...	Lam_dep_va_th...	0	0	0	0	0	0

Hình 4.2.2 : Kiểm thử trên tập dữ liệu Test

Ta thấy được 2 cột là topic và dự đoán của RapidMiner, từ đây ta có thể biết được bao nhiêu dự đoán đúng, bao nhiêu dự đoán sai. Các cột còn lại sẽ là tỉ lệ tự tin (confidence) của model khi dự đoán topic. RapidMiner sẽ chọn topic có độ tự tin cao nhất làm dự đoán.

Phần V : Kết luận

1. Kết quả đạt được và hạn chế nếu có

+ Kết quả :

- Hiểu được tổng quan về khai phá dữ liệu cũng như một số kỹ thuật khai phá cơ bản.
- Có thể ứng dụng thuật toán K- láng giềng vào các bộ dữ liệu khác nhau sau này.
- Xây dựng được một mô hình dự đoán và phân loại có độ chính xác tốt

+ Hạn chế :

- Tất cả các kiến thức nắm được ở mức lý thuyết hoặc thực hành sơ qua trên máy tính riêng, chưa được thực hành trên thực tế nhiều để hiểu sau rộng về thuật toán cũng như các kỹ thuật khai phá dữ liệu.

- Chưa thực sự thông thạo trong việc sử dụng các phần mềm hỗ trợ, dẫn đến ảnh hưởng đến năng suất và cả độ chính xác của mô hình.

2. Hướng nghiên cứu và phát triển

Trong quá trình tìm hiểu và thực hành, chúng em nhận thấy thuật toán K-NN là một thuật toán đơn giản, dễ sử dụng. Tuy nhiên, việc ứng dụng thuật toán này trên các bộ dữ liệu lớn còn khá hạn chế. Vì vậy, chúng em sẽ tiếp tục tìm hiểu các kỹ thuật cũng như thuật toán khai phá dữ liệu để có thêm nhiều kiến thức phục vụ công việc cũng như áp dụng vào các ứng dụng tự xây dựng sau này.

Tài liệu tham khảo và sử dụng

+ Thuật toán K- láng giềng gần nhất – Nguyễn Văn Chức

Link : http://bis.net.vn/forums/p/370/635.aspx

+ Các video, bài hướng dẫn sử dụng RapidMiner

Link : https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

+ vietnamese-stopwords – duyet

Link : https://github.com/stopwords/vietnamese-stopwords

+ Folder drive chứa Training data, Testing data, word_list, model và 2 Process

https://drive.google.com/drive/folders/1PJnKTZMVhWLnyVQ5YtRoRWL5xmF3Nf-b?usp=sharing

+ Ngoài ra nhóm chúng em còn tham khảo các bài viết, video hướng dẫn trên google, youtube