

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO PROJECT 1

**PHÂN LOẠI VĂN BẢN TIẾNG VIỆT SỬ DỤNG
THUẬT TOÁN NAÏVE BAYES**

Giảng viên hướng dẫn: TS. NGÔ XUÂN BÁCH

Sinh viên : LÊ THỊ NGỌC CHÂM

Lớp : D13CN6

Mã SV : B13DCCN305

Hà Nội, ngày 28 tháng 7 năm 2016

Mục Lục

I. Đặt vấn đề	1
1. Tổng quan	1
2. Mục tiêu	1
II. Nội dung nghiên cứu	1
1. Phát biểu bài toán phân loại văn bản:	1
2. Phương pháp phân loại Naïve Bayes đơn giản.	1
3. Phân loại văn bản Tiếng Việt sử dụng phân loại Bayes đơn giản	3
3.1. Dữ liệu huấn luyện:	3
3.2. Tiền xử lý	3
3.3. Thư viện đặc trưng	5
3.4. Áp dụng thuật toán huấn luyện bộ phân lớp.	5
4. Thực nghiệm.	5
4.1. Phương pháp thực nghiệm	5
4.2. Kết quả thực nghiệm	6
4.3. Nhận xét	10
5. Chương trình Demo	10
5.1. Giao diện chương trình.	10
5.2. Các lớp quan trọng	11
III. Kết luận:	11

I. ĐẶT VẤN ĐỀ

1. Tổng quan

Từ trước đến nay, phân loại văn bản tự động đã có rất nhiều công trình nghiên cứu và đạt được kết quả đáng khích lệ. Dựa trên các thống kê của Yang & Xiu (1999)^[1] một số phương pháp phân loại thông dụng hiện nay là: Support Vector Machine -Joachims, 1998^[2], Linear Least Squares Fit -Yang and Chute, 1994^[3]...vv. Các phương pháp trên đều dựa vào xác suất thống kê hoặc thông tin về trọng số của từ trong văn bản.

Vấn đề phân loại văn bản tiếng Việt được nhiều cơ sở nghiên cứu trong cả nước quan tâm trong những năm gần đây. Một số công trình nghiên cứu cũng đạt được những kết quả khả quan. Tuy vậy để đi đến những biện pháp triển khai khả thi thì vẫn cần đẩy mạnh nghiên cứu.

2. Mục tiêu

Tìm hiểu thuật toán Naïve Bayes, ứng dụng vào xây dựng một chương trình phân loại văn bản tiếng Việt cụ thể là phân loại các bài báo thuộc lĩnh vực : chính trị, kinh tế, y tế, giáo dục, văn hóa, giải trí, thể thao, khoa học.

II. NỘI DUNG NGHIÊN CỨU

1. Phát biểu bài toán phân loại văn bản.

Cho x là một văn bản, biết x là một trong các loại y , y có thể nhận giá trị từ một tập nhãn hữu hạn C , hãy tìm loại đúng nhất của x .

2. Phương pháp phân loại Bayes đơn giản.

Mỗi văn bản x là một đối tượng cần phân loại, x được chuyển thành một biểu diễn véc-tơ thực n chiều: $\mathbf{x} = (x_1, x_2, \dots, x_n)$ $x_j \in \mathbb{R}$ các thành phần x_j , $j = 1, 2, \dots, n$ được gọi là các đặc trưng hay thuộc tính của x .

Theo lý thuyết học Bayes, nhãn phân loại được xác định bằng cách tính xác suất điều kiện của nhãn khi quan sát thấy tổ hợp giá trị thuộc tính $\langle x_1, x_2, \dots, x_n \rangle$. Thuộc tính được chọn, ký hiệu là c_{MAP} là thuộc tính có xác suất điều kiện cao nhất (MAP là viết tắt của maximum a posterior).

$$y = c_{MAP} = \arg \max_{c_j \in C} P(c_j | x_1, x_2, \dots, x_n)$$

Sử dụng quy tắc Bayes, biểu thức trên được viết lại như sau

$$c_{MAP} = \arg \max_{cj \in C} \frac{P(x_1, x_2, \dots, x_n | cj) P(cj)}{P(x_1, x_2, \dots, x_n)}$$

Trong vế phải của biểu thức này, mẫu số không phụ thuộc vào c_j và ta có thể viết lại như sau:

$$y = c_{MAP} = \arg \max_{cj \in C} P(x_1, x_2, \dots, x_n | cj) P(cj)$$

Với giả thiết về tính độc lập xác suất có điều kiện, có thể viết:

$$P(x_1, x_2, \dots, x_n | cj) = P(x_1 | cj) P(x_2 | cj) \dots P(x_n | cj)$$

Thay vào công thức ở trên ta được bộ phân loại Bayes đơn giản (có đầu ra ký hiệu là c_{NB}) như sau:

$$c_{NB} = \arg \max_{cj \in C} P(cj) \prod_i P(x_i | cj)$$

Nếu sử dụng loga, ta có quy tắc phân loại tuyến tính:

$$y = \arg \max_{cj \in C} \left(\sum_i^n \log P(x_i | cj) + \log P(cj) \right)$$

❖ Mô hình Bayes đơn giản dạng đa thức^[4].

- Trong mô hình Bayes dạng đa thức, ta xét tần số xuất hiện của từng từ x_j trong x .
- Gọi $f(c_j, x_i)$ là số lần từ x_i xuất hiện trong loại văn bản c_j . Khi đó ước lượng hợp lý cực đại của các tham số là:

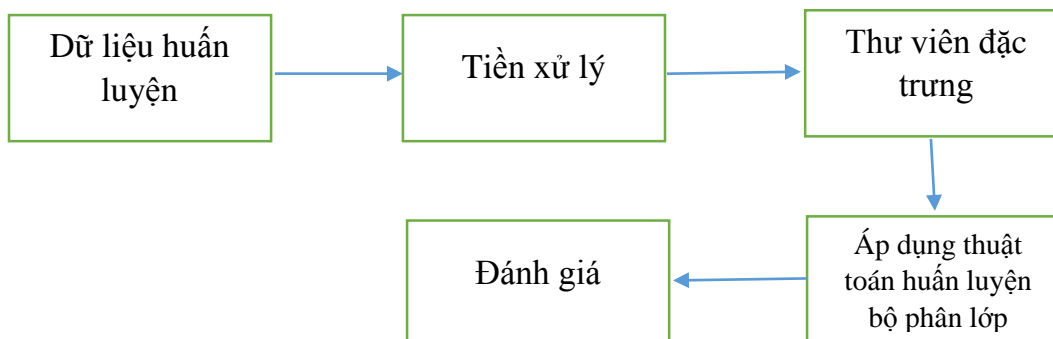
$$P(x_i | c_j) = \frac{f(c_j | x_i)}{\sum_i^n f(c_j | x_i)}$$

- Sử dụng phương pháp làm trơn Laplace để xử lý trường hợp $P(x_i | c_j) = 0$.

$$P(x_i | c_j) = \frac{f(c_j | x_i) + 1}{\sum_i^n f(c_j | x_i) + n}$$

3. Phân loại văn bản Tiếng Việt sử dụng phân loại Bayes đơn giản

Phân loại văn bản bao gồm các bước xử lý chung được thể hiện như sau:



3.1. Dữ liệu huấn luyện:

Kho dữ liệu được thu thập từ nhiều nguồn báo mạng khác nhau. Bao gồm 12000 bài báo theo chủ đề: chính trị, kinh tế, y tế, giáo dục, văn hóa, giải trí, thể thao, khoa học với tỷ lệ 1500 bài/chủ đề tại các trang:

- <http://vnexpress.net/>
- <http://tuoitre.vn/>
- <http://dantri.com.vn/>
- <http://news.zing.vn/>
- <http://laodong.com.vn/>
- <http://thethaovanhoa.vn/>
- <http://www.tienphong.vn/>
- <http://hanoimoi.com.vn/>
- <http://www.nhandan.org.vn/>

3.2. Tiền xử lý

Chuyển đổi tài liệu trong kho dữ liệu thành một hình thức phù hợp để phân loại, theo các bước: *tách câu, tách từ, gán nhãn từ loại, bỏ stopwords, cutoff*.

3.2.1. Tách câu

Sử dụng thư viện có sẵn vnSentDetector để tách một file văn bản thành các câu hoàn chỉnh. Cụ thể là dùng phương thức `String[] detectSentences(String fileName)` và `detectSentences(String fileName, String fileName)`.

Link thư viện: <http://mim.hus.vnu.edu.vn/phuonglh/software/vnSentDetector>

3.2.2. Tách từ và gán nhãn từ loại.

Sử dụng thư viện có sẵn vnTagger để tách một câu văn tiếng Việt thành các từ và nhãn từ loại của chúng. Cụ thể là sử dụng phương thức `List<WordTag> tagText2(String sentence)`. WordTag là một đối tượng có sẵn trong thư viện bao gồm từ và nhãn của nó.

Link thư viện: <http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTagger>

3.2.3. Bỏ stopwords.

Danh sách các từ vnstopWord được thu thập từ trang web <http://seo4b.com/thuat-ngu-SEO> và http://xltiengviet.wikia.com/wiki/Danh_s%C3%A1ch_stop_word bao gồm 892 từ phổ biến và mang nghĩa chung chung.(có trong file đính kèm).

Bên cạnh đó, dựa vào nhãn từ loại ta có thể bỏ đi các từ không mang nội dung về lĩnh vực mà câu truyền tải. Là các từ có nhãn:

- Np - Proper noun
- Nc - Classifier
- Nu - Unit noun
- P - Pronoun
- R - Adverb
- L - Determiner
- M - Numeral
- E - Preposition
- C - Subordinating conjunction
- CC - Coordinating conjunction
- I - Interjection
- T - Auxiliary, modal words

3.2.4. Cutoff.

Sử dụng phương pháp *count cutoff* để loại bỏ các cụm từ có tần số thấp trong tập dữ liệu huấn luyện.

Phương pháp *count cutoff* hoạt động như sau: nếu từ hoặc cụm từ nào xuất hiện ít hơn k lần trong tập dữ liệu huấn luyện thì nó bị loại bỏ ra khỏi mô hình ngôn ngữ. Khi tính toán, nếu gặp lại các cụm từ này thì xác suất của chúng sẽ được tính thông qua phương pháp làm trơn ở trên.(Trong phần cài đặt của project này, chọn k = 3).

3.3. Thư viện đặc trưng.

Sau khi tiền xử lý dữ liệu, dữ liệu huấn luyện còn lại những từ ngữ đặc trưng. Tiến hành tính xác suất của từ trong từng nhãn theo công thức:

$$P(x_i|c_j) = \frac{f(c_j|x_i) + 1}{\sum_i f(c_j|x_i) + n}$$

3.4. Áp dụng thuật toán huấn luyện bộ phân lớp.

Mỗi văn bản đầu vào x , tiến hành tiền xử lý để trích chọn những đặc trưng cho x . Biểu diễn x dưới dạng véc-tơ thực D chiều: $\mathbf{x} = (x_1, x_2, \dots, x_n)$ $x_j \in \mathbb{R}$ các thành phần $x_j, j = 1, 2, \dots, n$ được gọi là các đặc trưng hay thuộc tính của x .

Phân loại x vào nhãn y với:

$$y = \arg \max_{c_j \in C} \left(\sum_i^n \log P(x_i|c_j) + \log P(c_j) \right)$$

4. Thực nghiệm.

4.1. Phương pháp thực nghiệm.

Sử dụng phương pháp Cross Validation để đánh giá chương trình.

Chia dữ liệu huấn luyện thành 10 folders bằng nhau. Tiến hành thực nghiệm 10 lần mỗi lần sử dụng 9 folders làm dữ liệu huấn luyện, 1 folder làm dữ liệu cần phân loại. Mỗi lần ghi lại các độ đo:

4.1.1. Độ đo Accuracy.

Độ đo Accuracy là tỉ lệ của số kết quả đúng trên tổng số lần kiểm tra.

$$Accuracy = \frac{ta}{ta + fa}$$

ta: số test đúng

fa: số test sai

4.1.2. Độ đo Precision và Recall.

- Precision đối với nhãn l_i là tổng số lần kiểm tra được phân loại chính xác vào nhãn l_i chia cho tổng số lần kiểm tra được phân loại vào nhãn l_i :

$$Precision = \frac{tp}{tp + fp}$$

tp: số test phân loại đúng vào nhãn l_i

fp: số test phân loại sai vào nhãn l_i

- Recall đối với nhãn l_i là tổng số các lần kiểm tra thuộc lớp l_i được phân loại chính xác chia cho tổng số các lần kiểm tra thuộc lớp l_i :

$$Recall = \frac{tr}{tr + fr}$$

tr: số test thuộc lớp l_i được phân loại đúng

fr: số test thuộc lớp l_i bị phân loại sai

4.1.3. Độ đo F1.

F1 là độ đo kết hợp giữa Precision và Recall:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4.2. Bảng kết quả.

Các độ đo được tính theo đơn vị %.

Lần 1	tp	fp	tr	fr	Precision	Recall	F1
Cultural	107	34	107	43	75.89	71.33	73.54
Economy	120	26	120	30	82.19	80	81.08
Education	133	6	133	17	95.68	88.67	92.04
Entertainment	131	42	131	19	75.72	87.33	81.11
Medical	139	22	139	11	86.34	92.67	89.39
Politic	124	28	124	26	81.58	82.67	82.12
Science	114	27	114	36	80.85	76	78.35
Sport	147	0	147	3	100	98	98.99

Accuracy = 84.58

Lần 2	tp	fp	tr	fr	Precision	Recall	F1
Cultural	107	43	107	43	71.33	71.33	71.33
Economy	113	24	113	37	82.48	75.33	78.74
Education	131	12	131	19	91.61	87.33	89.42
Entertainment	132	40	132	18	76.74	88	81.99
Medical	140	15	140	10	90.32	93.33	91.8
Politic	128	25	128	22	83.66	85.33	84.49
Science	122	30	122	28	80.26	81.33	80.79
Sport	137	1	137	13	99.28	91.33	95.14

Accuracy = 84.17

Lần 3	tp	fp	tr	fr	Precision	Recall	F1
Cultural	108	31	108	42	77.7	72	74.74
Economy	128	15	128	22	89.51	85.33	87.37
Education	138	9	138	12	93.88	92	92.93
Entertainment	142	47	142	8	75.13	94.67	83.78
Medical	141	12	141	9	92.16	94	93.07
Politic	126	15	126	24	89.36	84	86.6
Science	123	22	123	27	84.83	82	83.39
Sport	138	5	138	12	96.5	92	94.2

Accuracy = 87

Lần 4	tp	fp	tr	fr	Precision	Recall	F1
Cultural	101	34	101	49	74.81	67.33	70.87
Economy	134	23	134	16	85.35	89.33	87.29
Education	130	4	130	20	97.01	86.67	91.55
Entertainment	135	46	135	15	74.59	90	81.57
Medical	133	15	133	17	89.86	88.67	89.26
Politic	131	33	131	19	79.88	87.33	83.44
Science	124	15	124	26	89.21	82.67	85.82
Sport	142	1	142	8	99.3	94.67	96.93

Accuracy = 85.83

Lần 5	tp	fp	tr	fr	Precision	Recall	F1
Cultural	111	26	111	39	81.02	74	77.35
Economy	132	29	132	18	81.99	88	84.89
Education	135	5	135	15	96.43	90	93.1
Entertainment	135	43	135	15	75.84	90	82.32
Medical	138	15	138	12	90.2	92	91.09
Politic	130	14	130	20	90.28	86.67	88.44
Science	127	17	127	23	88.19	84.67	86.39
Sport	142	1	142	8	99.3	94.67	96.93

Accuracy = 87.5

Lần 6	tp	fp	tr	fr	Precision	Recall	F1
Cultural	113	25	113	37	81.88	75.33	78.47
Economy	134	26	134	16	83.75	89.33	86.45
Education	140	6	140	10	95.89	93.33	94.59
Entertainment	129	35	129	21	78.66	86	82.17
Medical	138	18	138	12	88.46	92	90.2
Politic	130	17	130	20	88.44	86.67	87.55
Science	126	19	126	24	86.9	84	85.43
Sport	144	0	144	6	100	96	97.96

Accuracy = 87.83

Lần 7	tp	fp	tr	fr	Precision	Recall	F1
Cultural	113	27	113	37	80.71	75.33	77.93
Economy	138	23	138	12	85.71	92	88.74
Education	130	3	130	20	97.74	86.67	91.87
Entertainment	135	35	135	15	79.41	90	84.37
Medical	144	20	144	6	87.8	96	91.72
Politic	132	20	132	18	86.84	88	87.42
Science	117	15	117	33	88.64	78	82.98
Sport	145	3	145	5	97.97	96.67	97.32

Accuracy = 87.83

Lần 8	tp	fp	tr	fr	Precision	Recall	F1
Cultural	104	23	104	46	81.89	69.33	75.09
Economy	134	16	134	16	89.33	89.33	89.33
Education	140	5	140	10	96.55	93.33	94.91
Entertainment	139	46	139	11	75.14	92.67	82.99
Medical	137	12	137	13	91.95	91.33	91.64
Politic	134	23	134	16	85.35	89.33	87.29
Science	121	21	121	29	85.21	80.67	82.88
Sport	142	3	142	8	97.93	94.67	96.27

Accuracy = 87.58

Lần 9	tp	fp	tr	fr	Precision	Recall	F1
Cultural	116	24	116	34	82.86	77.33	80
Economy	128	16	128	22	88.89	85.33	87.07
Education	127	7	127	23	94.78	84.67	89.44
Entertainment	138	28	138	12	83.13	92	87.34
Medical	141	17	141	9	89.24	94	91.56
Politic	134	25	134	16	84.28	89.33	86.73
Science	126	22	126	24	85.14	84	84.57
Sport	145	7	145	5	95.39	96.67	96.03

Accuracy = 87.92

Lần 10	tp	fp	tr	fr	Precision	Recall	F1
Cultural	108	23	108	42	82.44	72	76.87
Economy	125	13	125	25	90.58	83.33	86.8
Education	138	0	138	12	100	92	95.83
Entertainment	142	48	142	8	74.74	94.67	83.53
Medical	137	15	137	13	90.13	91.33	90.73
Politic	137	23	137	13	85.63	91.33	88.39
Science	124	21	124	26	85.52	82.67	84.07
Sport	145	1	145	5	99.32	96.67	97.98

Accuracy = 88

Giá trị trung bình các độ đo sau 10 lần thực nghiệm:

Mean Accuracy = 86.82%

Mean F1	
Cultural	75.62
Economy	85.78
Education	92.57
Entertainment	83.12
Medical	91.05
Politic	86.25
Science	83.47
Sport	96.78

Từ đó ta được kết quả độ chính xác của chương trình.

Mean Value = 86.82%

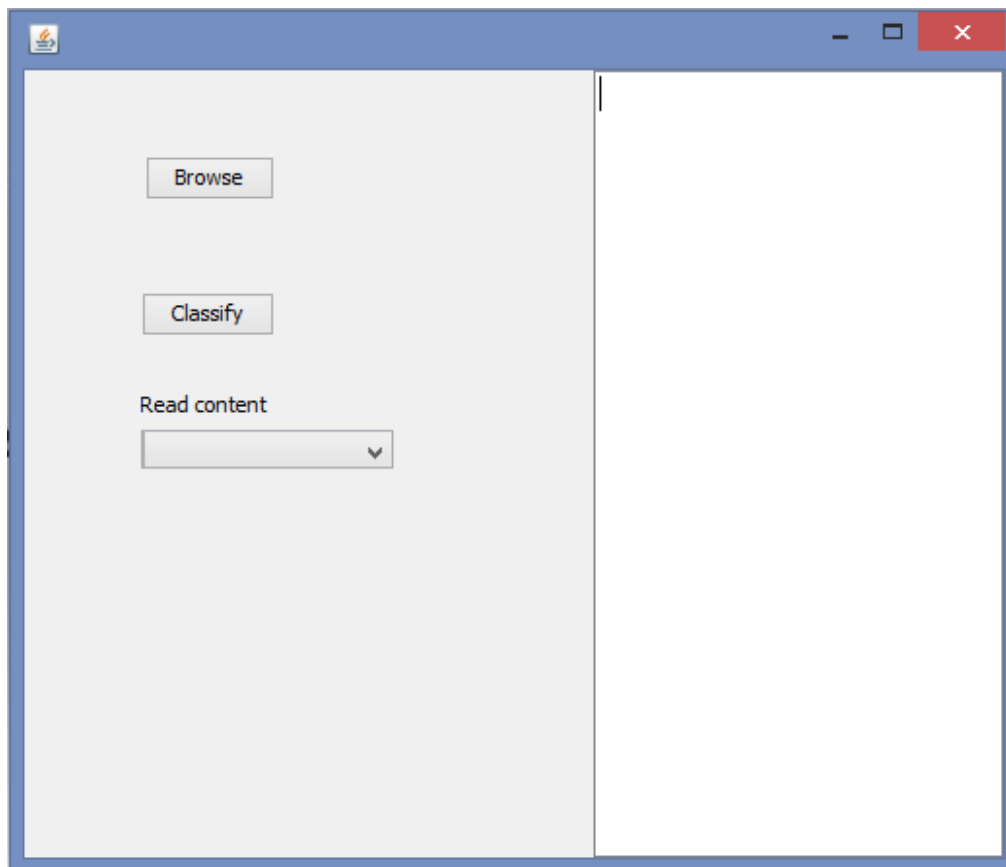
4.3. Nhận xét:

- Nhãn sport có giá trị F1 cao nhất với 96.78% do các file sport được lấy với nội dung chuẩn. Ngôn ngữ thể thao thì thường đặc trưng và khá là phân biệt so với ngôn ngữ của các nhãn khác.
- Bên cạnh đó thì nhãn medical, education cũng có giá trị F1 cao trên 90%.
- Nhãn cultural có giá trị F1 thấp nhất 75.62% vì bản tin văn hóa thường rất là đa dạng cả về ngôn ngữ và nội dung. Trong quá trình thực nghiệm thì nhận thấy rằng các file có nhãn cultural thường bị phân loại nhầm thành nhãn entertainment và politic nhất. Khi xem xét các file bị phân loại nhầm thì nhận thấy rằng nội dung của một số văn bản có thể là giải trí hoặc là văn hóa đều hợp lý.

5. Chương trình Demo.

5.1. Giao diện chương trình:

- Nút “Browse” cho phép chọn 1 hoặc nhiều file đầu vào.
- Nút “Classify” tiến hành phân loại các file đầu vào.
- Combobox “Read content” cho phép lựa chọn một file đầu vào để đọc.



5.2. Các lớp quan trọng.

- Lớp Word_pro: mỗi đối tượng Word_pro được đặc trưng bởi 2 thuộc tính: từ và xác suất của từ đó trong 1 nhãn.
- Lớp Pre_Process: lớp tiền xử lý dữ liệu
 - ✓ ArrayList<Word_pro> pre_Process(String inputFile)
- Lớp Calculate_probability: tính xác suất cho từ theo từng nhãn.
- Lớp Classify: lớp phân loại nhãn cho văn bản đầu vào.

Các phương thức:

- ✓ String classify(ArrayList<ArrayList<Word_pro>> dictionary, int total_word[], String inputFile): phân loại một file trả kết quả là nhãn mà file thuộc về dựa trên từ điển dictionary và số số từ của một nhãn total_word[].

III. KẾT LUẬN:

Sau khi thực hiện xong project 1, em đã hoàn thiện một ứng dụng đầu tiên về xử lý ngôn ngữ Tiếng Việt. Chương trình khá là cơ bản nhưng đó là tiền đề giúp em hiểu thêm nhiều thứ về lĩnh vực xử lý ngôn ngữ tự nhiên từ việc tách câu, tách từ, gán nhãn từ loại và đặc biệt là sử dụng Bayes để phân loại văn bản.

Tuy nhiên chương trình chỉ đạt được độ chính xác mức 86.82% do thuật toán, do cách cài đặt và dữ liệu huấn luyện chưa thực sự chuẩn vậy nên còn cần phải cải tiến thêm.

TÀI LIỆU THAM KHẢO

[1] Yang and Xin Liu, A re-examination of text categorization methods, Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99), 1999

[2] Joachims, Text Categorization with Support Vector Machines, Learning with Many Relevant Features, In European Conference on Machine Learning (ECML), 1998

[3] Y. Yang and G.Chute, An example-based mapping method for text categorization and retrieval, ACM Transaction on Information Systems(TOIS), 12(3):252277,1994

[4] Một số giải thuật phân loại văn bản/TS.Lê Hồng Phương/ Đại học quốc gia Hà Nội/ Trường đại học khoa học tự nhiên/ Viện công nghệ nghiên cứu FPT.

Giáo trình TTNT, TS. Từ Minh Phương, Học viện công nghệ bưu chính viễn thông Hà Nội.