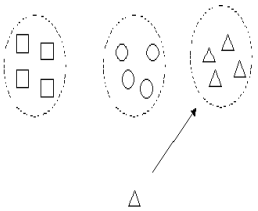
Phân loại văn bản

Lê Thanh Hương
Bộ môn Hệ thống thông tin
Viện CNTT&TT

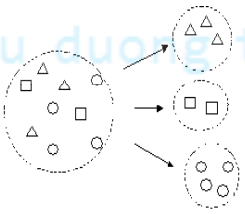
Phân loại văn bản

- **Phân loại:** (Text Categorization)
  Đầu vào của bài toán là tập các văn bản đã được phân lớp sẵn, cho một văn bản mới vào, ứng dụng phải chỉ ra văn bản đó thuộc chủ đề nào trong các chủ đề ban đầu.



Phân nhóm văn bản

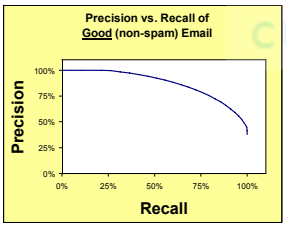
- **Phân nhóm**: (Text Clustering)
  Là bài toán cho một tập văn bản chưa được phân lớp gì cả, ứng dụng phải chia tập văn bản này thành các nhóm dựa trên độ tương đồng giữa chúng.



Tại sao cần PLVB?

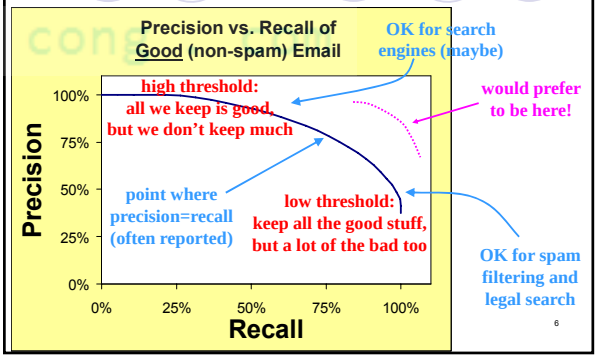
- Là tiếng Việt?
- Lọc tin
- Chuyển hướng cuộc gọi
- Phân loại thư (cuộc hẹn, công việc, khẩn, bạn bè, thư rác, …)

Đo độ chính xác



- Precision = $\frac{\text{các thư được giữ (đúng)}}{\text{tất cả các thư giữ}}$
- Recall = $\frac{\text{các thư được giữ (đúng)}}{\text{các thư đúng}}$

Đo độ chính xác



Các trường hợp đo độ chính xác phức tạp hơn

- Phân lớp nhiều lớp
  - Độ chính xác trung bình (hoặc precision hoặc recall) của các phân lớp 2 lớp: thể thao hoặc không, tin tức hoặc không
  - Tốt hơn, đánh giá chi phí của các lớp lỗi
    - vd, đánh giá ảnh hưởng của các vấn đề sau:
      - đặt các bài về Thể thao vào mục Tin tức
      - đặt các bài về Mốt vào mục Tin tức
      - đặt các bài về Tin tức vào mục Mốt
    - điều chỉnh hệ thống để giảm thiểu tổng chi phí
- Với các hệ thống xếp hạng:
  - Mức độ liên quan đến xếp hạng của con người
  - Lấy các phản hồi tích cực từ người dùng

7

Cách phân loại

```
Subject: would you like to . . . .

. . drive a new vehicle for free ? ? ? this is not hype or a
hoax , there are hundreds of people driving brand new cars ,
suvs , minivans , trucks , or rvs . it does not matter to us
what type of vehicle you choose . if you qualify for our
program , it is your choice of vehicle , color , and options
. we don ' t care . just by driving the vehicle , you are
promoting our program . if you would like to find out more
about this exciting opportunity to drive a brand new vehicle
for free , please go to this site : http : / / 209 . 134 . 14
. 131 / ntr to watch a short 4 minute audio / video
presentation which gives you more information about our
exciting new car program . if you do n't want to see the
short video , but want us to send you our information package
that explains our exciting opportunity for you to drive a new
vehicle for free , please go here : http : / / 209 . 134 . 14
. 131 / ntr / form . htm we would like to add you the group
of happy people driving a new vehicle for free . happy
motoring .
```

8

Cách phân loại? (có giám sát)

1. Xây dựng mô hình n-gram cho mỗi lớp, sử dụng lý thuyết Bayes
2. Biểu diễn mỗi tài liệu như 1 vector
   (cần chọn cách biểu diễn và độ đo khoảng cách ; sử dụng SVD?)
   - Cách 1: Đưa vào lớp mà tài liệu gần với trung tâm của lớp nhất (có thể ko phù hợp nếu các thành phần trong lớp cách xa nhau)
   - Cách 2: Chia mỗi lớp thành các nhóm con (sau đó sử dụng cách 1 để lấy 1 lớp, trả về lớp chứa nhóm con. Phương pháp này cũng có thể dùng cho mô hình n-gram)
   - Cách 3: Chỉ nhìn vào các nhãn của các tài liệu luyện (vd, sử dụng k láng giềng gần, có thể láng giềng gần hơn có trọng số lớn hơn)

9

Cách phân loại? (có giám sát)

3. Coi như bài toán giải quyết nhập nhằng từ
   a) Mô hình vector – sử dụng tất cả các đặc trưng
   b) Danh sách quyết định – chỉ sử dụng đặc trưng tốt nhất
   c) Naive Bayes – sử dụng tất cả các đặc trưng, đánh trọng số dựa trên tác động của nó trong việc phân biệt các lớp
   d) Cây quyết định – sử dụng một số đặc trưng theo trình tự

10

Mô hình vector

2 tài liệu sau tương tự nhau:
Sau khi chuẩn hóa độ dài vector thành 1,
giống không gian Euclidean (similar endpoint)
High dot product (similar direction)

aardvark	abacus	abandoned	abbot	abduct	above	...	zygote	zymurgy
(0,	0,	3,	1,	0,	7,	...	1,	0)
(0,	0,	1,	0,	0,	3,	...	0,	1)

Khi tạo vector, có thể:
loại bỏ từ chức năng hoặc giảm trọng số của nó
Sử dụng các đặc trưng khác so với unigrams

11

Danh sách quyết định

slide courtesy of D. Yarowsky (modified)

Để phân giải nhập nhằng của từ *lead* :
- Duyệt danh sách các ứng cử viên
- Dấu hiệu đầu tiên tìm thấy là dấu hiệu quyết định
- Không tốt bằng cách kết hợp các dấu hiệu, nhưng hoạt động tốt cho WSD

Đánh giá trọng số của dấu hiệu:
log [p(cue | sense A) [smoothed] / p(cue | sense B)]

Position	Collocation	led	li:d
+1 L	lead *level/N*	219	0
-1 W	*narrow* lead	0	70
+1 W	lead *in*	207	898

LogL	Evidence	Pronunciation
11.40	*follow/V* + lead	⇒ li:d
11.20	*zinc* (in ±*k* words)	⇒ led
11.10	lead *level/N*	⇒ led
10.66	*of* lead *in*	⇒ led
10.59	*the* lead *in*	⇒ li:d
10.51	lead *role*	⇒ li:d
10.35	*copper* (in ±*k* words)	⇒ led
10.28	lead *time*	⇒ li:d
10.24	lead *levels*	⇒ led
10.16	lead *poisoning*	⇒ led
8.55	*big* lead	⇒ li:d
8.49	*narrow* lead	⇒ li:d
7.76	*take/V* + lead	⇒ li:d
5.99	lead , *NOUN*	⇒ led
1.15	lead *in*	⇒ li:d

slide courtesy of D. Yarowsky (modified)

Kết hợp các dấu hiệu và Naive Bayes

Authorship ID: Who Wrote a Student's Term Paper?

Word in Text	Frequency as Student A	Frequency as Student B
optimally	97	1
certainly	84	3
typically	46	4
perspicuous	26	0
actually	13	4
whilst	6	0
the	241	229
awesome	0	63
totally	0	40
wonderful	0	26
incredibly	0	13

các giá trị này được tính từ các bài của các tác giả đã biết trước (học có giám sát)

$$\frac{P(optimally|StudentA)}{P(optimally|StudentB)} = \frac{97}{1} \qquad \frac{P(the|StudentA)}{P(the|StudentB)} = \frac{1.1}{1}$$

13

slide courtesy of D. Yarowsky (modified)

Kết hợp các dấu hiệu và Naive Bayes

Combining Evidence - One (Bayesian) Approach

$$\frac{P(optimally|StudentA)}{P(optimally|StudentB)} = \frac{97}{1} \qquad \frac{P(the|StudentA)}{P(the|StudentB)} = \frac{1.1}{1}$$

$$\frac{P(awesome|StudentA)}{P(awesome|StudentB)} = \frac{0}{63}$$

$$\frac{P(StudentA)}{P(StudentB)} = \frac{P(w_{-3}|StudentA)}{P(w_{-3}|StudentB)} \times \frac{P(w_{-2}|StudentA)}{P(w_{-2}|StudentB)} \times \ldots$$

Mô hình "Naïve Bayes" cho phân lớp văn bản

(Chú ý giả thiết độc lập)

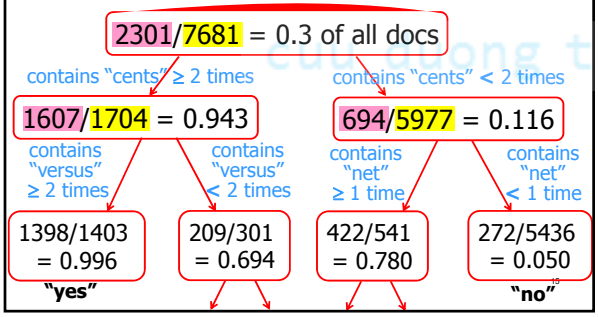
Câu này là câu của sinh viên A hay B?

14

example from Manning & Schütze

Cây quyết định

Bài báo Reuters này thuộc lĩnh vực Lợi nhuận?



15

Các đặc trưng ngoài Unigrams

- Các cách tiếp cận trên (trừ mô hình n-gram) có thể sử dụng các đặc trưng khác, không chỉ unigrams.
- Vấn đề lựa chọn đặc trưng
  - Sử dụng tập lớn các đặc trưng lưu trong 1 template
  - Có thể tìm các đặc trưng có ích khi xét 1 cách độc lập?
  - Thêm lần lượt các đặc trưng
    - Đo hoặc đoán khả năng cải thiện của mỗi đặc trưng
  - Cuối cùng, loại bỏ các đặc trưng làm giảm tính chính xác của hệ thống khi tiến hành thử nghiệm trên bộ dữ liệu mới
- Chương trình SpamAssassin sử dụng các đặc trưng gì

16

Các đặc trưng trong SpamAssassin

100	From: địa chỉ trong danh sách đen
4.0	Người gửi trong danh sách www.habeas.com Habeas Infringer
3.994	Ngày không hợp lệ: tiêu đề (timezone không tồn tại)
3.970	Viết bằng 1 ngôn ngữ lạ
3.910	Liệt kê trong Razor2, xem http://razor.sf.net/
3.801	Tiêu đề là các ký tự lấp đầy 8-bit
3.472	Thông báo tuân theo Senate Bill 1618
3.437	exists:X-Precedence-Ref
3.371	Ngày đảo ngược
3.350	Thông báo bạn có thể bị loại khỏi danh sách
3.284	Tài sản bí mật
3.283	Thông báo yêu cầu rời khỏi danh sách
3.261	Có chứa từ "Stop Snoring"
3.251	Received: chứa tên với địa chỉ IP giả
3.250	Nhận được qua chuyển tiếp trong list.dsbl.org
3.200	Tập ký tự chỉ một ngôn ngữ lạ

17

Các đặc trưng trong SpamAssassin

3.198	Forged eudoramail.com 'Received:' header found
3.193	Free Investment
3.180	Received via SBLed relay, seehttp://www.spamhaus.org/sbl/
3.140	Character set doesn't exist
3.123	Dig up Dirt on Friends
3.090	No MX records for the From: domain
3.072	X-Mailer contains malformed Outlook Expressversion
3.044	Stock Disclaimer Statement
3.009	Apparently, NOT Multi Level Marketing
3.005	Bulk email software fingerprint (jpfree) found inheaders
2.991	exists:Complain-To
2.975	Bulk email software fingerprint (VC_IPA) found inheaders
2.968	Invalid Date: year begins with zero
2.932	Mentions Spam law "H.R. 3113"
2.900	Received forged, contains fake AOL relays
2.879	Asks for credit card details

18

Cách phân loại? (không giám sát)

Nếu không có dữ liệu luyện

Thực hiện lặp đi lặp lại:

1. Nhóm các tài liệu
2. Luyện mô hình n-gram, Naive Bayes, hoặc danh sách quyết định để phân biệt các nhóm
3. Sử dụng mô hình để gán lại các tài liệu vào các nhóm (chỉ có 1 số ít thay đổi)
4. Quay lại bước 2 đến khi hội tụ

19

Cách phân loại? (bán giám sát)

Nếu chỉ có một ít dữ liệu luyện?

1. Bắt đầu với các lớp nhỏ và chính xác
2. Luyện mô hình n-gram, Naive Bayes, hoặc danh sách quyết định để phân biệt các nhóm
3. Thêm vào mỗi lớp các tài liệu mới mà mô hình phân loại được một cách chắc chắn (cũng có thể loại bớt một số tài liệu)
4. Quay lại bước 2 đến khi hội tụ

20

Cách phân loại? (thích nghi)

Nếu dữ liệu luyện được tăng cường theo thời gian?

- Sử dụng phản hồi (tích cực hoặc thụ động) về việc phân lớp hiện có
- Các hệ thống mới phân lớp hoặc điều chỉnh
  - Thêm các tài liệu mới vào dữ liệu luyện
  - Nếu chúng chưa được gán nhãn (không giám sát), gán chúng một cách tự động

Mô hình được điều chỉnh theo thời gian

- Vd., thay đổi trung tâm của nhóm hoặc các tham số của n-gram
- Muốn tăng trọng số của dữ liệu mới
  - Vd., tài liệu k ngày trước có trọng số 0.9^k (k=0,1,2, ...)
  - Mô hình hiện tại = dữ liệu hiện tại + 0.9 * mô hình cũ

21

Cách phân loại? (phân cấp)

Đưa 1 tài liệu vào Yahoo! category?

- Có hàng nghìn lớp – quá khó
- Chọn 1 trong 14 lớp ở mức trên cùng, vd., khoa học
- Sau đó sử dụng bộ phân lớp cho lĩnh vực Khoa học để chọn 1 trong 54 lớp mức 2 của lớp Khoa học
- Tiếp tục đi xuống các mức dưới
- Khi không thể phân lớp với độ chắc chắn cao, hỏi con người (sử dụng câu trả lời của con người như là dữ liệu luyện mới)

22