Mã thông báo

Mã thông báo là quá trình nhận dạng mã thông báo từ hoặc trích xuất từ văn bản đang chạy. Nó thường là bước đầu tiên trong tiền xử lý văn bản. Khoảng trắng và dấu chấm câu thường đóng vai trò là dấu phân cách từ đáng tin cậy; tuy nhiên, các cách tiếp cận đơn giản có khả năng gặp phải các trường hợp ngoại lệ như "USA" và tương tự. Mã thông báo là công cụ NLP được tối ưu hóa cao cho nhiệm vụ mã hóa từ và chúng có thể dựa vào các biểu thức chính quy được tạo cẩn thận hoặc có thể được đào tạo bằng thuật toán học máy.

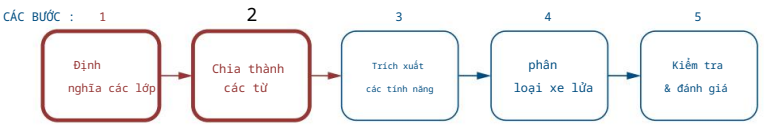
Để kiểm tra mức độ hiểu biết của bạn về những bước mã thông báo đạt được, hãy thử mã hóa chuỗi văn bản theo cách thủ công trong bài tập 2.3 trước khi xem xét giải pháp ở cuối chương. Sau đó, bạn cũng sẽ có thể kiểm tra xem các giải pháp của mình có trùng khớp với các giải pháp được trả về bởi trình mã thông báo hay không:

Bài tập 2.3 Bạn

sẽ mã hóa các chuỗi sau thành từ như thế nào? (Có thể tìm thấy giải pháp ở cuối chương này.)

```
1 Cách tốt nhất để nấu một chiếc bánh piz z a là gì?
2 Chúng ta sẽ sử dụng đá nướng.
3 Tôi chưa từng sử dụng đá nướng trước đây.
```

Bây giờ chúng ta hãy xác định bước 2 của thuật toán của bạn như sau: áp dụng mã thông báo để chia văn bản đang chạy thành các từ sẽ đóng vai trò là tính năng (hình 2.7).



Hình 2.7 Ở bước 2 (được đánh dấu), tách văn bản đang chạy thành các từ.

2.2.3 Bước 3: Trích xuất và chuẩn hóa các tính năng

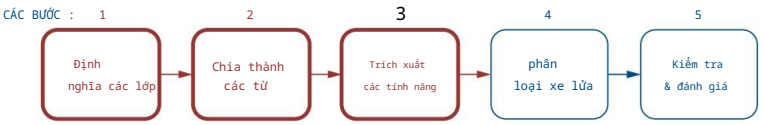
Bây giờ, chúng tôi xem xét kỹ các từ được trích xuất và xem liệu tất cả chúng có tốt như nhau để được sử dụng làm tính năng hay không—nghĩa là liệu chúng có biểu thị như nhau về nội dung liên quan đến spam hay không. Giả sử hai email sử dụng định dạng khác nhau: một email nói

Thu thập tiền trúng xổ số của bạn

trong khi một người khác nói

Thu thập tiền thắng xổ số của bạn

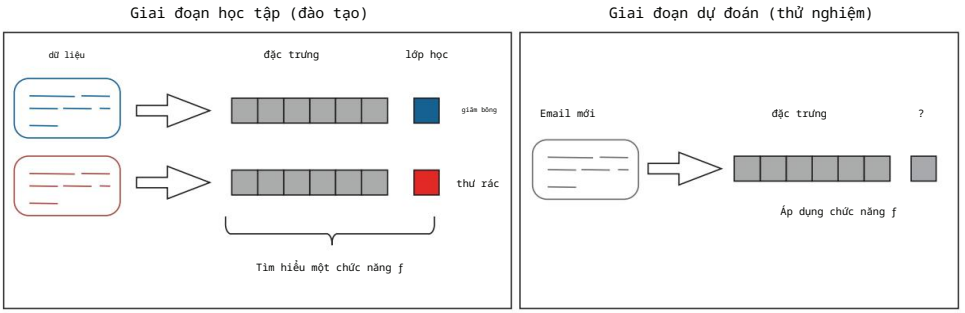
Thuật toán phân tách các thông báo này thành các từ sẽ kết thúc bằng các danh sách từ khác nhau vì, ví dụ, xổ số Xổ số, nhưng liệu nó có khác nhau về ý nghĩa không? Để loại bỏ các vấn đề về định dạng như chữ hoa so với chữ thường, bạn có thể đặt tất cả các từ được trích xuất thành chữ thường bằng chức năng Python. Do đó, bước 3 trong thuật toán của bạn nên được xác định như sau: trích xuất và chuẩn hóa các tính năng; ví dụ, bằng cách đặt tất cả các từ thành chữ thường (hình 2.8).



Hình 2.8 Trong bước 3 (được đánh dấu), trích xuất và chuẩn hóa các tính năng của bạn (ví dụ: từ).

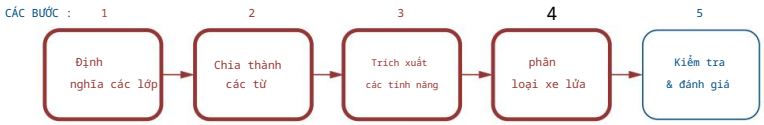
2.2.4 Bước 4: Huấn luyện bộ phân loại

Tại thời điểm này, bạn sẽ có hai bộ dữ liệu—một bộ được liên kết với **lớp** thư rác và một bộ khác được liên kết với **lớp** ham . Mỗi dữ liệu được xử lý trước theo cùng một cách trong bước 2 và 3 và các tính năng được trích xuất. Tiếp theo, bạn cần cho phép máy sử dụng dữ liệu này để xây dựng kết nối giữa tập hợp các tính năng (thuộc tính) mô tả từng loại email (thư rác hoặc ham) và các nhãn được đính kèm với từng loại. Ở bước 4, một thuật toán học máy cố gắng xây dựng một mô hình thống kê, một hàm, giúp nó phân biệt giữa hai lớp. Đây là những gì xảy ra trong giai đoạn học tập (đào tạo). Hình 2.9 là hình ảnh cập nhật về đào tạo và kiểm tra
quy trình.



Hình 2.9 Các giai đoạn học tập (đào tạo) và dự đoán của quá trình lọc thư rác

Vì vậy, bước 4 của thuật toán nên được xác định như sau: xác định mô hình học máy và đào tạo nó trên dữ liệu với các tính năng được xác định trước trong các bước trước đó (hình 2.10).

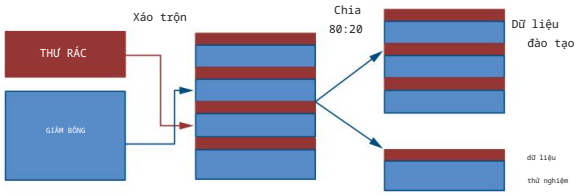


Hình 2.10 Ở bước 4 (được đánh dấu), xác định mô hình học máy và huấn luyện mô hình đó trên dữ liệu.

Thuật toán của bạn hiện đã học được một chức năng có thể ánh xạ các tính năng từ mỗi loại email tới các nhãn thư rác và ham . Trong quá trình đào tạo, thuật toán của bạn sẽ tìm ra tính năng nào quan trọng hơn và nên được tin cậy trong quá trình dự đoán; ví dụ: nó có thể phát hiện ra rằng sự xuất hiện của xổ số từ trong email có liên quan chặt chẽ với nhãn thư rác, trong khi sự xuất hiện của cuộc gặp gỡ từ sẽ gợi ý rõ ràng về nhãn giăm bông . Bước cuối cùng trong quy trình này là đảm bảo rằng thuật toán thuật toán đang thực hiện tốt các dự đoán đó. Bạn sẽ làm nó thế nào?

Hãy nhớ rằng ban đầu bạn được cung cấp một tập hợp các email được gắn nhãn sẵn cho bạn là thư rác và ham. Điều này có nghĩa là bạn biết câu trả lời chính xác cho những email này. Tại sao không sử dụng một số trong số chúng để kiểm tra xem thuật toán của bạn hoạt động tốt như thế nào? Trên thực tế, đây chính xác là cách nó được thực hiện trong học máy—bạn sử dụng một số dữ liệu được gắn nhãn của mình để kiểm tra hiệu suất của bộ phân loại. Bit dữ liệu này được gọi là tập kiểm tra. Tuy nhiên, có một lưu ý: nếu bạn đã sử dụng dữ liệu này để huấn luyện bộ phân loại (tức là để cho nó tìm ra sự tương ứng giữa các đặc điểm và các lớp), thì nó đã biết câu trả lời đúng. Để tránh điều đó, bạn cần đảm bảo rằng bit dữ liệu bạn đã sử dụng trong bước 4 để huấn luyện là riêng biệt và không trùng lặp với tập kiểm tra. Bit dữ liệu này được gọi là tập huấn luyện. Do đó, trước khi đào tạo trình phân loại của bạn ở bước 4, bạn cần chia tập dữ liệu đầy đủ của mình thành các tập kiểm tra và đào tạo. Đây là bộ quy tắc cho điều đó (hình 2.11):

Xáo trộn dữ liệu của bạn để tránh mọi sai lệch. Chia ngẫu nhiên thành một tỷ lệ lớn hơn cho giai đoạn huấn luyện và dành phần còn lại cho giai đoạn kiểm tra. Tỷ lệ điển hình cho các bộ là 80% cho huấn luyện và 20% cho kiểm tra. Huấn luyện bộ phân loại của bạn ở bước 4 chỉ sử dụng tập huấn luyện. Bộ thử nghiệm của bạn ở đó để cung cấp cho bạn ước tính thực tế và hợp lý về hiệu suất của trình phân loại, vì vậy đừng để trình phân loại của bạn xem trộm. Chỉ sử dụng nó ở bước cuối cùng để đánh giá.



Hình 2.11 Trước khi huấn luyện bộ phân loại, hãy xáo trộn dữ liệu và chia thành các tập huấn luyện và kiểm tra.

Phân tách dữ liệu cho học máy có giám sát Trong học máy có giám sát, thuật toán được đào tạo trên một tập hợp con của dữ liệu được gắn nhãn gọi là tập huấn luyện. Nó sử dụng tập hợp con này để tìm hiểu chức năng ánh xạ dữ liệu đầu vào sang nhãn đầu ra. Tập kiểm tra là tập con của dữ liệu, tách rời khỏi tập huấn luyện, trên đó thuật toán có thể được đánh giá. Việc phân chia dữ liệu điển hình là 80% cho đào tạo và 20% cho thử nghiệm. Lưu ý rằng điều quan trọng là hai bộ không trùng nhau. Nếu thuật toán của bạn được đào tạo và thử nghiệm trên cùng một dữ liệu, bạn sẽ không thể biết nó thực sự học được gì thay vì ghi nhớ.

2.2.5 Bước 5: Đánh giá bộ phân loại

Giả sử bạn đã đào tạo trình phân loại của mình ở bước 4 và sau đó áp dụng nó vào dữ liệu thử nghiệm. Làm thế nào bạn sẽ đo lường hiệu suất? Một cách tiếp cận là kiểm tra xem thuật toán phân loại chính xác tỷ lệ email kiểm tra là bao nhiêu—nghĩa là gán nhãn thư rác cho email rác và phân loại email ham là ham. Tỷ lệ này được gọi là độ chính xác và cách tính của nó khá đơn giản:

$$Accuracy = \frac{num(\text{correct predictions})}{num(\text{all test instances})}$$

Bây giờ hãy kiểm tra sự hiểu biết của bạn với bài tập 2.4.

Bài tập 2.4

Giả sử thuật toán của bạn dự đoán các nhãn sau cho một số tập dữ liệu nhỏ của các ví dụ thử nghiệm. (Có thể tìm thấy giải pháp ở cuối chương này.)

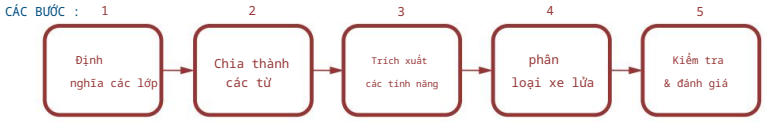
Nhãn chính xác	Nhãn dự đoán
Thư rác	giăm bông
Thư rác	Thư rác
giăm bông	giăm bông
giăm bông	Thư rác
giăm bông	giăm bông

1. Độ chính xác của bộ phân loại của bạn trên tập dữ liệu nhỏ này là bao nhiêu?
2. Đây có phải là một độ chính xác tốt? Nghĩa là, nó có gợi ý rằng bộ phân loại hoạt động tốt không? Điều gì sẽ xảy ra nếu bạn biết rằng tỷ lệ giữa ham và thư rác trong nhóm email của bạn là 50:50? Điều gì sẽ xảy ra nếu đó là 60% email ham và 40% thư rác—điều đó có thay đổi đánh giá của bạn về mức độ hoạt động của trình phân loại không?
3. Nó có hoạt động tốt hơn trong việc xác định email ham hoặc email spam không?

Hãy thảo luận về cách giải bài tập này (lưu ý rằng bạn cũng có thể tìm thấy lời giải chi tiết hơn ở cuối chương). Dự đoán của bộ phân loại dựa trên

sự phân bố của các lớp mà bạn đã gặp trong bài tập này được gọi là đường cơ sở. Trong trường hợp phân bổ lớp bằng nhau, đường cơ sở là 50% và nếu trình phân loại của bạn mang lại độ chính xác là 60%, thì nó sẽ vượt trội so với đường cơ sở này. Trong trường hợp chia 60:40, đường cơ sở, còn có thể được gọi là đường cơ sở của nhóm đa số, là 60%. Điều này có nghĩa là nếu một "công cụ phân loại" giả hoàn toàn không học và chỉ dự đoán nhãn ham cho tất cả các email, thì nó sẽ không lọc bất kỳ email spam nào khỏi hộp thư đến, nhưng độ chính xác của nó cũng sẽ là 60% —giống như công cụ phân loại của bạn điều đó thực sự được đào tạo và thực hiện một số phân loại! Điều này làm cho bộ phân loại trong trường hợp thứ hai trong bài tập này kém hữu ích hơn nhiều vì nó không thực hiện đường cơ sở của lớp đa số.

Tóm lại, độ chính xác là một thước đo hiệu suất tổng thể tốt, nhưng bạn cần lưu ý (1) sự phân bố của các lớp để có điểm so sánh về hiệu suất của bộ phân loại và (2) hiệu suất trên mỗi lớp bị ẩn trong một giá trị độ chính xác duy nhất nhưng có thể gợi ý điểm mạnh và điểm yếu của bộ phân loại của bạn. Do đó, bước cuối cùng, bước 5, trong thuật toán của bạn là áp dụng bộ xác định lớp cho dữ liệu thử nghiệm và đánh giá hiệu suất của nó (hình 2.12).



Hình 2.12 Ở bước 5, kiểm tra và đánh giá bộ phân loại của bạn.

2.3 Thực hiện bộ lọc thư rác của riêng bạn

Bây giờ, hãy thực hiện từng bước trong năm bước. Đã đến lúc bạn mở Jupyter và tạo một sổ ghi chép mới để bắt đầu mã hóa bộ lọc thư rác của riêng mình.

> LƯU Ý Xin nhắc lại: chúng tôi đang sử dụng Jupyter Notebooks, vì chúng cung cấp cho người thực hành một môi trường linh hoạt trong đó mã có thể dễ dàng thêm, chạy và cập nhật và có thể dễ dàng quan sát kết quả đầu ra. Ngoài ra, bạn có thể sử dụng bất kỳ IDE Python nào cho các ví dụ mã từ cuốn sách này. Xem https://jupyter.org để biết hướng dẫn cài đặt. Ngoài ra, hãy xem phần phụ lục để biết hướng dẫn cài đặt và kho lưu trữ của sách (https://github.com/ekochmar/Getting-Started-with-NLP) để biết cả hướng dẫn cài đặt và tất cả các ví dụ về mã.

2.3.1 Bước 1: Xác định dữ liệu và lớp

Khá thường xuyên khi làm việc trên các ứng dụng NLP và máy học, bạn có thể phát hiện ra rằng sự cố đã được mô tả trước đó hoặc ai đó đã thu thập một số dữ liệu mà bạn có thể sử dụng để xây dựng phiên bản ban đầu cho thuật toán của mình. Ví dụ: nếu bạn muốn xây dựng một bộ phân loại máy học để phát hiện thư rác, bạn cần cung cấp cho thuật toán của mình một số lượng thư rác và thư rác vừa đủ. Cách tốt nhất để xây dựng một bộ phân loại như vậy là thu thập các email ham và thư rác của riêng bạn và đào tạo