Trường Đại học Khoa học tự nhiên
Đại học Quốc gia TP. Hồ Chí Minh

TOÁN ỨNG DỤNG & THỐNG KÊ

Bài tập tuần 10

Phan Đình Kha - 18120127

Ngày 21 tháng 6 năm 2021

1 Yêu cầu

Thống kê (tự động) trên N file text, xây dựng$< S, P >$. Hỏi nếu X_n = 'a' thì X_{n+1} =?.

2 Bài làm

2.1 Tìm X_{n+1} =?

Chúng ta xét dữ liệu text sau:

> *Lunar New Year Festival often falls between late January and early February; it is among the most important holidays in Vietnam.*

Ta sẽ xác định xích Markov bằng cách tìm $< S, P >$. Để đơn giản, chúng ta chọn S là tập gồm kí tự 'a' và các kí tự liền sau nó. Như vậy, S = 'a', 'r', 'l', 't', 'n'.

Gọi X_i là kí tự thứ n trong chuỗi được xét (nhận giá trị là các phần tử trong tập S), $\pi^{(i)}$ là vector xác suất của kí tự thứ i (trong đó, $\pi^{(i)} \in R^5$, $|S| = 5$).

Dựa trên tập S và dữ liệu ở trên, ta thống kê số trường hợp mà liền sau của kí tự 'x' là kí tự 'y', với 'z', 'y' thuộc S.

Từ đó, ta có bảng sau:

	a	r	l	t	n
a	0	5	2	1	3
r	0	0	1	1	0
l	1	0	1	0	0
t	1	0	0	0	1
n	2	0	0	1	0

Nhận xét: Dòng 1 bắt đầu là 'a', với mỗi ô trên dòng đó là số trường hợp mà liền sau của kí tự 'a' là kí tự cột tương ứng. Tương tự cho các cột còn lại.

Chuyển thành xác suất ta có:

	r	l	t	n	n
a	0	0.454545	0.181818	0.090909	0.272727
r	0	0	0.5	0.5	0
l	0.5	0	0.333333	0	0
t	0.5	0	0	0	0.5
n	0.666667	0	0	0.333333	0

Nhận xét: tổng xác suất trên từng dòng đều bằng 1.

Do đó, ma trận xác suất chuyển trạng thái là:

$$P = \begin{bmatrix} 0 & 0.454545 & 0.181818 & 0.090909 & 0.272727 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.333333 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \\ 0.666667 & 0 & 0 & 0.333333 & 0 \end{bmatrix}$$

Theo đề, X_n = 'a'. Khi đó, vector xác suất của kí tự thứ n sẽ là $\pi^{(n)}$ = [1;0;0;0;0].

Vector xác suất của kí tự thứ $n + 1$ (X_{n+1}) là:

$$
\begin{aligned}
\pi^{(n+1)} &= \pi^{(n)}.P \\
&= [1;0;0;0;0] \begin{bmatrix} 0 & 0.454545 & 0.181818 & 0.090909 & 0.272727 \\ 0 & 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.333333 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0.5 \\ 0.666667 & 0 & 0 & 0.333333 & 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 0.454545 & 0.181818 & 0.090909 & 0.272727 \end{bmatrix}
\end{aligned}
$$

Phân tích kết quả:

Xác suất để X_{n+1} nhận giá trị là 'a', 'r', 'l', 't', 'n' tương ứng là $\begin{bmatrix} 0 & 0.454545 & 0.181818 & 0.090909 & 0.272727 \end{bmatrix}$.

2.2 Ứng dụng

Bài toán trên có thể ứng dụng vào việc thông báo lỗi chính tả.

Hướng thực hiện:

- Chúng ta chọn S là tập 24 kí tự alphabet và kí tự khoảng trắng ' '.
- Chuẩn bị dữ liệu là N file text đủ lớn. Thực hiện thống kê để tìm được P như ví dụ trên.
- Cuối cuối cùng là bước kiểm tra. Chúng ta sẽ duyệt qua từng từ, với mỗi kí tự trong từ, ta sẽ đoán kí tự tiếp theo có thể là những kí tự nào (có xác suất) dựa trên xích markov. Nếu kí tự tiếp theo trong từ đang được xét không nằm trong những kí tự ta dự đoán, ta sẽ thông báo lỗi chính tả.