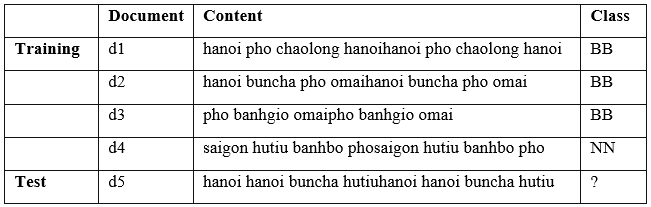
**Buổi thực hành 4: Naïve Bayes**

**Bài 2**: Giả sử trong tập training có các văn bản d1, d2, d3, d4d1, d2, d3, d4 như trong bảng dưới đây. Mỗi văn bản này thuộc vào 1 trong 2 classes: B (Bắc) hoặc N (Nam). Hãy xác định class của văn bản d5.



**BÀI LÀM:**

Ta có: PBB = 3/4 = 0.75, PNN = 1/4 = 0.25

Tập hợp toàn bộ các dữ liệu từ bảng là: V = {hanoi, pho, chaolong, buncha, omai, banhgio, saigon, hutiu, banhbo}. Tổng cộng số phần tử trong từ điển là |V|=9

Hình dưới đây minh hoạ quá trình Training và Test cho bài toán này khi sử dụng Multinomial Naive Bayes, trong đó có sử dụng Laplace smoothing với α=1.

Class = BB

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| d1: x1  d2: x2  d3: x3 | hanoi | pho | chaolong | buncha | omai | banhgio | saigon | hutiu | banhbo |
| 4 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 0 | 2 | 2 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 2 | 2 | 0 | 0 | 0 |
| Total | 6 | 5 | 2 | 2 | 4 | 2 | 0 | 0 | 0 |
| = 1 | 13/30 | 11/30 | 5/30 | 5/30 | 9/30 | 5/30 | 1/30 | 1/30 | 1/30 |

d = |V| = 9

* NBB = 21
* NBB + |V| = 30

Class = NN

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| d4: x4 | hanoi | pho | chaolong | buncha | omai | banhgio | saigon | hutiu | banhbo |
| 0 | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 2 |
| Total | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 2 |
| = 1 | 1/17 | 5/17 | 1/17 | 1/17 | 1/17 | 1/17 | 5/17 | 5/17 | 5/17 |

d = |V| = 9

* NBB = 8
* NBB + |V| = 17

p(BB|d5) = 0.75 \* (13/30)4 \* (5/30)2 \* (1/30)2 = 8.162 \* 10-7

p(NN|d5) = 0.25 \* (1/17)4 \* (1/17)2 \* (5/17)2 = 8.96 \* 10-10

p(BB|d5) = = 0.998

p(NN|d5) = 1 0.998 = 0.002