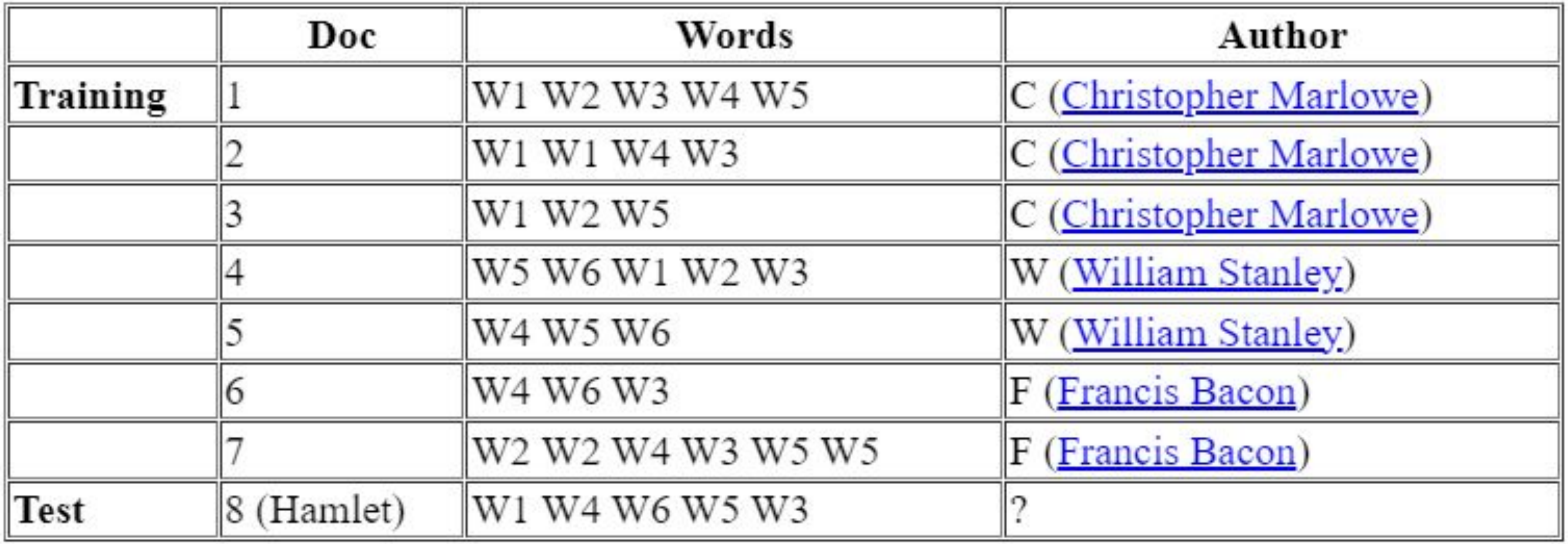
Text classifier
Who is the real author of Hamlet?

Khoi Duong

Prof. Chang

CS550

Dataset



	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C (Christopher Marlowe)
	2	W1 W1 W4 W3	C (Christopher Marlowe)
	3	W1 W2 W5	C (Christopher Marlowe)
	4	W5 W6 W1 W2 W3	W (William Stanley)
	5	W4 W5 W6	W (William Stanley)
	6	W4 W6 W3	F (Francis Bacon)
	7	W2 W2 W4 W3 W5 W5	F (Francis Bacon)
Test	8 (Hamlet)	W1 W4 W6 W5 W3	?

Calculation

There are total of 7 rows with 3 classes (C, W, and F)

P(C) = 3/7

P(W) = 2/7

P(F) = 2/7

|V| = Number of vocab = 6

Calculation

$P(w1|C) = (count(w1, C) + 1) / (count(C) + |V|) = (4+1) / (12+6) = 5/18$

$P(w1|W) = (count(w1, W) + 1) / (count(W) + |V|) = (1+1) / (8+6) = 1/7$

$P(w1|F) = (count(w1, F) + 1) / (count(F) + |V|) = (0+1) / (9+6) = 1/15$

$P(w3|C) = (count(w3, C) + 1) / (count(C) + |V|) = (2+1) / (12+6) = 1/6$

$P(w3|W) = (count(w3, W) + 1) / (count(W) + |V|) = (1+1) / (8+6) = 1/7$

$P(w3|F) = (count(w3, F) + 1) / (count(F) + |V|) = (2+1) / (9+6) = 1/5$

Calculation

$P(w4|C) = (count(w4, C) + 1) / (count(C) + |V|) = (2+1) / (12+6) = 1/6$

$P(w4|W) = (count(w4, W) + 1) / (count(W) + |V|) = (1+1) / (8+6) = 1/7$

$P(w4|F) = (count(w4, F) + 1) / (count(F) + |V|) = (2+1) / (9+6) = 1/5$

$P(w5|C) = (count(w5, C) + 1) / (count(C) + |V|) = (2+1) / (12+6) = 1/6$

$P(w5|W) = (count(w5, W) + 1) / (count(W) + |V|) = (2+1) / (8+6) = 3/14$

$P(w5|F) = (count(w5, F) + 1) / (count(F) + |V|) = (2+1) / (9+6) = 1/5$

Calculation

$P(w6|C) = (\text{count}(w6, C) + 1) / (\text{count}(C) + |V|) = (0+1) / (12+6) = 1/18$

$P(w6|W) = (\text{count}(w6, W) + 1) / (\text{count}(W) + |V|) = (2+1) / (8+6) = 3/14$

$P(w6|F) = (\text{count}(w6, F) + 1) / (\text{count}(F) + |V|) = (1+1) / (9+6) = 2/15$

Calculation

P(C|d8) = P(C) * P(w1|C) * P(w4|C) * P(w6|C) * P(w5|C) * P(w3|C)

= 3/7 * 5/18 * 1/6 * 1/6 * 1/6 * 1/18 = 5/163296 = 0.000031

P(W|d8) = P(W) * P(w1|W) * P(w4|W) * P(w6|W) * P(w5|W) * P(w3|W)

= 2/7 * 1/7 * 1/7 * 1/7 * 3/14 * 3/14 = 9/235298 = 0.000038

P(F|d8) = P(F) * P(w1|F) * P(w4|F) * P(w6|F) * P(w5|F) * P(w3|F)

= 2/7 * 1/15 * 1/5 * 1/5 * 1/5 * 2/15 = 4/196875 = 0.000020

Therefore, given d8 (Hamlet), it belongs to W (William Stanley)

Program

Dataset: Text_Classifier.csv

Source code and Google Colab: https://colab.research.google.com/drive/1FcgdkyNxwwNp2ynDXytYHdp5u7oF39vE?usp=sharing