Week 7: Homework 1: Text Classification

BY ADELEYE ADEBODUN 19747

Question?

Using text classifier, predict the author of hamlet

The table

	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	?

Training

The probability of Author P(C) = 3/7 (number of C classes over total classes)

The probability of Author P(W) = 2/7

The probability of Author P(F) = 2/7

Conditional probabilities

Author C

P(W1|C) = (4 + 1)/ (12 + 6) = 5/18

P(W2|C) = (2 + 1)/ (12 + 6) = 3/18

P(W3|C) = (2 + 1)/ (12 + 6) = 3/18

P(W4|C) = (2 + 1)/ (12 + 6) = 3/18

P(W5|C) = (2 + 1)/ (12 + 6) = 3/18

P(W6|C) = (0 + 1)/ (12 + 6) = 1/18

Author W

P(W1|W) = (1 + 1)/ (8 + 6) = 2/14

P(W2|W) = (1 + 1)/ (8 + 6) = 2/14

P(W3|W) = (1 + 1)/ (8 + 6) = 2/14

P(W4|W) = (1 + 1)/ (8 + 6) = 2/14

P(W5|W) = (2 + 1)/ (8 + 6) = 3/14

P(W6|W) = (2 + 1)/ (8 + 6) = 3/14

Author F

P(W1|F) = (0 + 1)/ (9 + 6) = 1/15

P(W2|F) = (2 + 1)/ (9 + 6) = 3/15

P(W3|F) = (2 + 1)/ (9 + 6) = 3/15

P(W4|F) = (2 + 1)/ (9 + 6) = 3/15

P(W5|F) = (2 + 1)/ (9 + 6) = 3/15

P(W6|F) = (1 + 1)/ (9 + 6) = 2/15

Test

Decide whether d8 belongs to Author C or W or F

Analysis

A: Probability that d8 belongs to Author C

Applying compare model

$P(C|d8) \propto P(C) * P(W1|C)^3 * P(W2|C)^2 * P(W3|C)^2 * P(W4|C)^2 * P(W5|C)^2$

$= 3/7 * (5/18)^3 * (3/18)^2 * (3/18)^2 * (3/18)^2 * (3/18)^2$

$= 0.00000019$

Note: $P(C) = 3/7$

There are 5 words in d8: W1 W4 W6 W5 W3

$P(W1|C) = 5/18$

$P(W4|C) = 3/18$

$P(W6|C) = 1/18$

$P(W5|C) = 3/18$

$P(W3|C) = 3/18$

. Probability that d8 belongs to Author W

Applying compare model

$P(W|d8) \propto P(W) * P(W5|W) ^2 * P(W6|W) ^2 * P(W1|W) * P(W2|W) * P(W3|W) * P(W4|W)$

= 2/7 * (3/14) ^2 * 3/14) ^2 * 2/14 * 2/14 * 2/14 * 2/14

= 0.00000025

Note: P(W) = 2/7

There are 5 words in d8: W1 W4 W6 W5 W3

P(W1|W) = 2/14

P(W4|W) = 2/14

P(W6|W) = 3/14

P(W5|W) = 3/14

P(W3|W) = 2/14

Probability that d8 belongs to Author F

Applying compare model

$P(F|d8) \propto P(F) * P(W4|F)\ ^2 * P(W6|F) * P(W3|F)\ ^ 2 * P(W2|F)\ ^ 2 * P(W5|F)\ ^ 2$

$= 2/7 * (3/15)\ ^2 * 2/15 * (3/15)\ ^2 * (3/15)\ ^2 * (3/15)^2$

$= 0.0000001$

Note: P(F) = 2/7

There are 5 words in d8: W1 W4 W6 W5 W3

P(W1|F) = 1/15

P(W4|F) = 3/15

P(W6|F) = 2/15

P(W5|F) = 3/15

P(W3|F) = 3/15

Conclusion: D8 should belong to Author C