Text Classification

Who is the real author of Hamlet?

Jisen Fang
CS550 - Machine Learning and Business Intelligence
2023 Spring

https://hc.labnet.sfbu.edu/~henry/sfbu/course/mllib/naive_bayes/slide/exercise_naive_bayes.html

Q12 ==> Who is the real author of Hamlet?

Table of contents

Text Classification

Text Classification is the process of categorizing text from one or more different classes to organize, structure, and filter into parameters.

Applications:

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language identification
- Sentiment analysis
- And more

Who is the Real Author of Hamlet

	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	?

Total classes : 3 (C, W, F)

Number of different words: 6 (W1, W2, W3, W4, W5, W6)

Number of words in 3 Class C: 12

Number of words in 2 Class W: 8

Number of words in 2 Class F: 9

Training - Priors

P(X) = The probability of a class X = Number of class X / total number of classes = N_X / N

P(C) = The probability of a class C = 3 C-classes/7 total classes = 3/7

P(W) = The probability of a class W = 2 W-classes/7 total classes = 2/7

P(F) = The probability of a class F = 2 F-classes/7 total classes = 2/7

Training - Conditional Probabilities

P(w|x) = If a document belongs to class x, the probability that the document has word w.

= The probability that the word w appears on the class x document.

= (count(w, x) + 1) / (count(x)+|V|)

count(x) = Number of words in Class x

|V| = 6, the number of different words (W1, W2, W3, W4, W5, W6)

$P(W1|C) = (count(W1, C) + \underline{1}) / (count(C)+|V|) = (4+1)/(12+6) = 5/18$

$P(W1|W) = (count(W1, W) + \underline{1}) / (count(W)+|V|) = (1+1)/(8+6) = 2/14 = 1/7$

$P(W1|F) = (count(W1, F) + \underline{1}) / (count(F)+|V|) = (0+1)/(9+6) = 1/15$

$P(W2|C) = (count(W2, C) + \underline{1}) / (count(C)+|V|) = (2+1)/(12+6) = 3/18 = 1/6$

$P(W2|W) = (count(W2, W) + \underline{1}) / (count(W)+|V|) = (1+1)/(8+6) = 2/14 = 1/7$

$P(W2|F) = (count(W2, F) + \underline{1}) / (count(F)+|V|) = (2+1)/(9+6) = 3/15 = 1/5$

$P(W3|C) = (count(W3, C) + \underline{1}) / (count(C)+|V|) = (2+1)/(12+6) = 3/18 = 1/6$

$P(W3|W) = (count(W3, W) + \underline{1}) / (count(W)+|V|) = (1+1)/(8+6) = 2/14 = 1/7$

$P(W3|F) = (count(W3, F) + \underline{1}) / (count(F)+|V|) = (2+1)/(9+6) = 3/15 = 1/5$

P(W1|C)	5/18
P(W1|W)	1/7
P(W1|F)	1/15
P(W2|C)	1/6
P(W2|W)	1/7
P(W2|F)	1/5
P(W3|C)	1/6
P(W3|W)	1/7
P(W3|F)	1/5

P(W4C) = (count(W4, C) + 1) / (count(C)+|V|) = (2+1)/(12+6) = 3/18

P(W4|W) = (count(W4, W) + 1) / (count(W)+|V|) = (1+1)/(8+6) = 2/14 = 1/7

P(W4|F) = (count(W4, F) + 1) / (count(F)+|V|) = (2+1)/(9+6) = 3/15 = 1/5

P(W5|C) = (count(W5, C) + 1) / (count(C)+|V|) = (2+1)/(12+6) = 3/18 = 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) = 3/15 = 1/5

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

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

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

P(W4|C)	3/18
P(W4|W)	1/7
P(W4|F)	1/5
P(W5|C)	1/6
P(W5|W)	3/14
P(W5|F)	1/5
P(W6|C)	1/18
P(W6|W)	3/14
P(W6|F)	2/15

Test - Analysis

Decide whether d8 (Hamlet) belongs to class C, class W or class F.

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

a. The probability that the document d8 belongs to class C

$P(C|d8) \propto P(C) * P(W1|C) * P(W4|C) * P(W6|C) * P(W5|C) * P(W3|C)$

$= 3/7 * 5/18 * 3/18 * 1/18 * 1/6 * 1/6$

$\cong 0.00003062$

b. The probability that the document d8 belongs to class W

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

$= 2/7 * 1/7 * 1/7 * 3/14 * 3/14 * 1/7$

$\cong 0.00003825$

c. The probability that the document d8 belongs to class F

$P(F|d8) \propto P(F) * P(W1|F) * P(W4|F) * P(W6|F) * P(W5|F) * P(W3|F)$

$= 2/7 * 1/15 * 1/5 * 2/15 * 1/5 * 1/5$

$\cong 0.00002032$

Conclusion

Based on the probability in the analysis,

$P(W|d8) > P(C|d8) > P(F|d8)$, $0.00003825 > 0.00003062 > 0.00002032$

the document d8 should belong to class W (William Stanley)