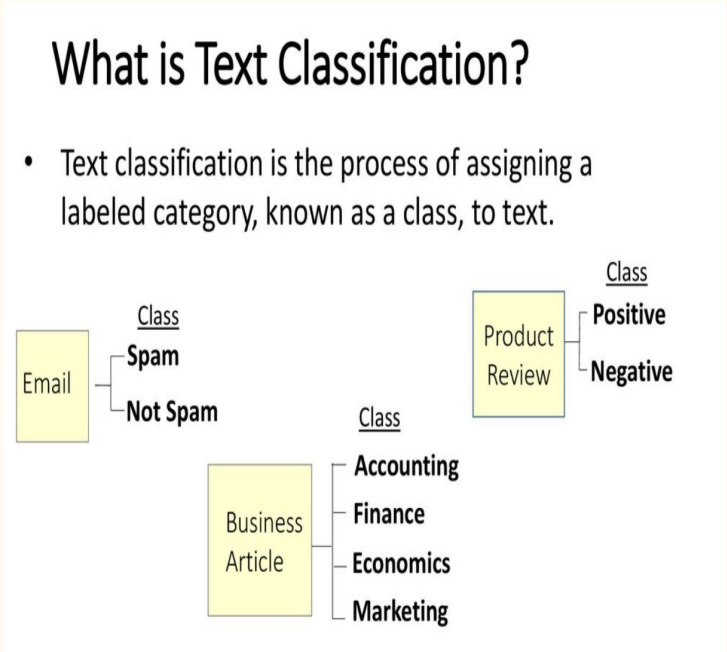
Text Classification

Norina Akhtar

Table of Contents

Introduction

Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web.



Design

	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)

Implementation

	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	?

- Does d8 belong to C or W or F?

$$\hat{P}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$$

Implementation

Please clearly shows the results :

P(C) : The probability of class c =3/7 (i.e., 3 c-classes / total classes)

P(W): The probability of class w =32/7 (i.e., 2 w-classes / total classes)

P(F) : The probability of class F =2/7 (i.e., 2 f-classes / total classes)

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)
= 2/14

P(W1|F) : (count(w1, F) + 1) / (count(F)+|V|)
= (0+1) / (9+6)
= 1/15

Implementation

P(W3|C) : (count(w3, C) + 1) / (count(C)+|V|)
= (2+1) / (12+6)
= 3/18

P(W3|W) : (count(w3, W) + 1) / (count(w)+|V|)
= (1+1) / (8+6)
= 2/14

P(W3|F) : (count(w3, F) + 1) / (count(F)+|V|)
= (2+1) / (9+6)
= 3/15

P(W4|C) : (count(w3, 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

P(W4|F) : (count(w4, F) + 1) / (count(F)+|V|)
= (2+1) / (9+6)
= 3/15

P(W5|C) : (count(w5, C) + 1) / (count(C)+|V|)
= (2+1) / (12+6)
= 3/18

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

P(W6|C) : (count(w6, C) + 1) / (count(C)+|V|)
= (0+1) / (12+6)
= 1/18

Implementation

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

Test

Decide whether d8 (i.e., document 8) belongs to class C or class W or class W.

P(C|d8)

The probability that the document d8 belongs to class C

P(C|d8) = P(C) * P(d8|C) / P(d5)

==> Applying Bayes Theorm

= P(C) * P(W1 ∩ W4 ∩ W6∩ W5 ∩ W3|c) / P(d8)

==> Applying Naive Bayes Theorm

∝ (P(c) * (P(W1|c) * P(W4 |c) * P(W6|c) * P(W5|c) * P(W3|c) |c))) / P(d5)

==> Applying Compare Model

P(c|d8) ∝ P(c) * (P(W1|c) * P(W4 |c) * P(W6|c) * P(W5 * P(W3|c)

= (3/7) * (5/18) * (3/18) * (1/18) * (3/18) * (3/18)

≅ 3.061924×10^{-5}

Test

P(W|d8):

The probability that the document d8 belongs to class w

P(W|d8) = P(W) * P(d8|W) / P(d8)

==> Applying Bayes Theorm

= P(W) * P(W1 ∩ W4 ∩ W6∩ W5 ∩ W3|c) / P(d8)

==> Applying Naive Bayes Theorm

∝ (P(W) * (P(W1|W) * P(W4 |W) * P(W6|W) * P(W5|W * P(W3|W) |W))) / P(d8)

==> Applying Compare Model

P(w|d8) ∝ P(W) * (P(W1|W) * P(W4 |W) * P(W6|W) * P(W5|W) * P(W3|W)

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

≅ 3.824937×10-5

P(F|d8):

The probability that the document d8 belongs to class F

P(F|d8) = P(F) * P(d8|F) / P(d8)

==> Applying Bayes Theorm

= P(F) * P(W1 ∩ W4 ∩ W6∩ W5 ∩ W3|c) / P(d8)

==> Applying Naive Bayes Theorm

∝ (P(F) * (P(W1|F) * P(W4 |F) * P(W6|F) * P(W5|F) * P(W3|F) |F))) / P(d8)

==> Applying Compare Model

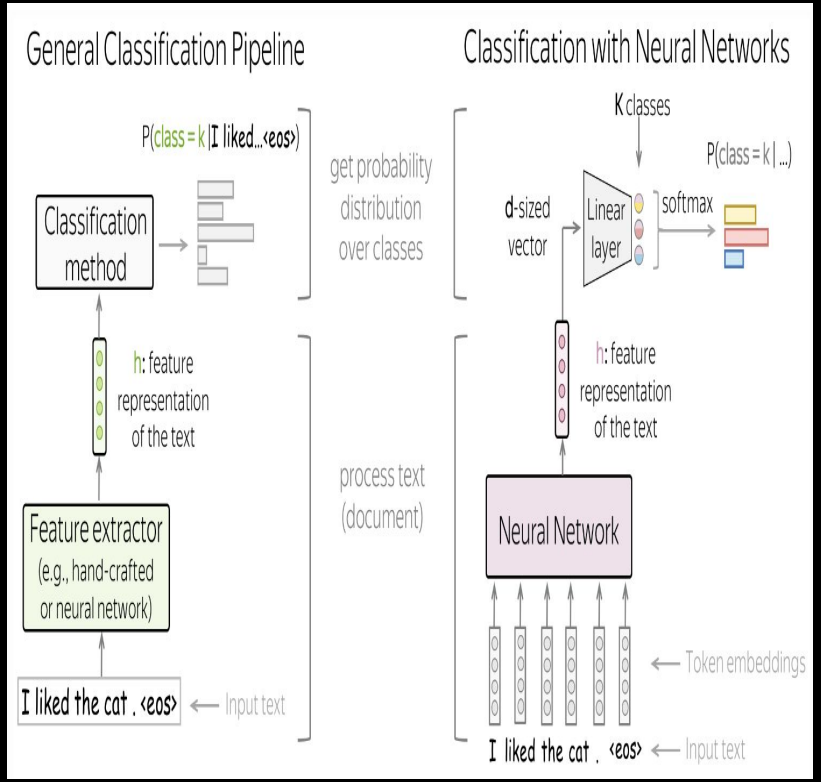
P(F|d8) ∝ P(F) * (P(W1|F) * P(W4 |F) * P(W6|F) * P(W5|F) * P(W3|F)

= (2/7) * (1/15) * (3/15) * (2/15) * (3/15) * (3/15)

≅ 2.031746×10-5

Enhancement Ideas

- Model can be improved by adding bigrams and trigrams as features
- Topic modeling can also used to enhance text classification
- Use Recurrent Neural network architecture(mostly the LSTM versions)
- SVM



Conclusion

	Doc	Words	Author
Training	1	W1 W2 W3 W4 W5	C
	2	W1 W1 W4 W3	C
	3	W1 W2 W5	C
	4	W5 W6 W1 W2 W3	W
	5	W4 W5 W6	W
	6	W4 W6 W3	F
	7	W2 W2 W4 W3 W5 W5	F
Test	8 (Hamlet)	W1 W4 W6 W5 W3	W

Document 8 should belong to the class W. This is because while comparing bayes and naive bayes theorem we identified that the probability of Hamlet belonging to william is greater than other authors.

So the real author of hamlet is **William Stanley.**

References

- https://hc.labnet.sfbu.edu/~henry/sfbu/course/mllib/naive_bayes/slide/text_classifier.html
- https://lena-voita.github.io/nlp_course/text_classification.html
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a