Implementation of Text Classifier

By Akanksha Singh(19745)

Introduction

This includes a Project in which we implement a Text Classifier and test it to predict that who is the real author of Hamlet?

Table of contents

Problem Question with Training and Test Data

	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	?

Prior and Conditional Probabilities formulas used

Training

Priors:

P(X) = The probability of a class X

= Number of class X / total number of classes

- = N_X / N

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|)

Results for probabilities

- P(C)=3/7
- P(W)=2/7
- P(F)=2/7
- P(W1|C)=4+1/12+6=5/18
- P(W4|C)=2+1/12+6=1/6
- P(W6|C)=0+1/12+6=1/18
- P(W5|C)=2+1/12+6=1/6
- P(W3|C)=2+1/12+6=1/6
- P(W1|W)=1+1/8+6=1/7
- P(W4|W)=1+1/14 = 1/7
- P(W6|W)=2+1/14=3/14
- P(W5|W)=3/14
- P(W3|W)=1/7
- P(W1|F)=0+1/9+6=1/15
- P(W4|F)=2+1/15=1/5
- P(W6|F)=2/15
- P(W5|F)=1/5
- P(W3|F)=1/5

Applying Compare model on all the classes and finding results

- 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*1/18*1/6*1/6 = 0.0000306

- 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*3/14*3/14*1/7=**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*2/15*1/5*1/5=0.0000203

Conclusion

- Does d8 belong to C or W or F?

Ans : It belongs to W

William Stanley has the highest probability of being the author of Hamlet.

Bibliography

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

Github Link

https://github.com/codeyogg/Machine_learning/tree/main/Text_Classification