NLP - Assignment#1

Yuqian Zhang

(1)D1 = [woof woof meow]

D2 = [woof woof squeak]

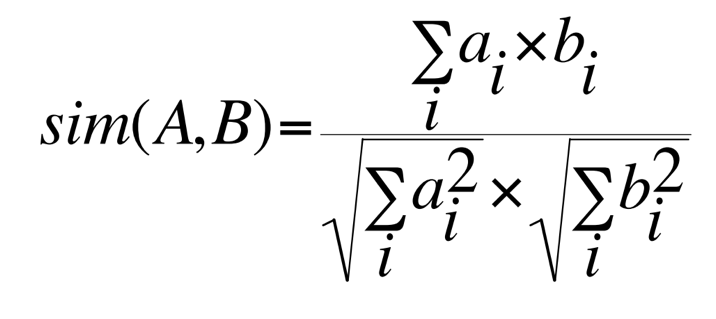
W = [woof meow squeak] (n=3)

V1 = [ 2, 1, 0 ]

V2 = [ 2, 0, 1 ]

(a) What is the cosine similarity of D1 and D2, not using idf weighting?

Solution: By using the common choice:



sim (D1,D2) = ( 2\*2+1\*0+0\*1)/5 = 4/5

(b) What is the cosine similarity if idf weighting is used?

Solution:

idf1=log(2/2) = 0

idf2=log(2/1) = 1

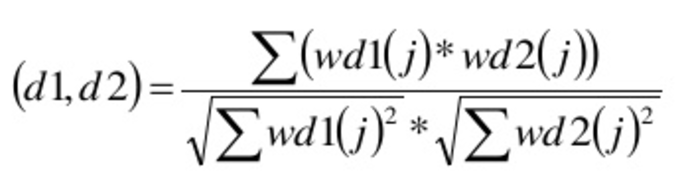
idf3=log(2/1) = 1

D1: D2:

w1= tf1 \*idf1 = (2)\*0=0 w1= tf1 \*idf1 = (2)\*0=0

w2= tf2 \*idf2 = (1)\*1=1 w2= tf2 \*idf2 = (0)\*1=0

w3= tf3 \*idf3 = (0)\*1=0 w3= tf3 \*idf3 = (1)\*1=1



Applying all data in D1 and D2 to the formula above,

sim (D1,D2)=(0+0+0)/((1)\*(1))=0

(c) D3= [meow squeak] added

Solution:

D1 = [woof woof meow]

D2 = [woof woof squeak]

D3= [ meow squeak]

W = [woof meow squeak] (n=3)

V1 = [ 2, 1, 0 ]

V2 = [ 2, 0, 1 ]

V3 = [ 0, 1, 1 ]

idf1=log(3/2)

idf2=log(3/2)

idf3=log(3/2)

D1:

w1= tf1 \*idf1 = (2)\* log(3/2)

w2= tf2 \*idf2 = (1)\* log(3/2)

w3= tf3 \*idf3 = (0)\* log(3/2)=0

D2:

w1= tf1 \*idf1 = (2)\* log(3/2)

w2= tf2 \*idf2 = (0)\* log(3/2)=0

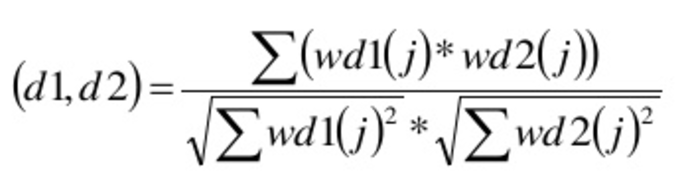
w3= tf3 \*idf3 = (1)\* log(3/2)

D3:

w1= tf1 \*idf1 = (0)\* log(3/2)=0

w2= tf2 \*idf2 = (1)\* log(3/2)

w3= tf3 \*idf3 = (1)\* log(3/2)



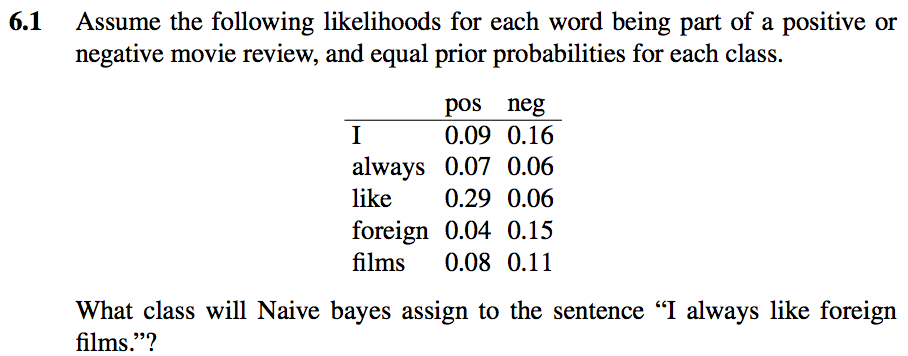
Applying all data in D1,D2,D3 to the formula above,

sim (D1,D2)=(4\*(log(3/2)^2)/(5\*log(3/2)^2)=4/5

sim (D1,D3)=(1\*(log(3/2)^2)/(10^1/2\*(log(3/2)^2)=1/10^1/2

sim (D2,D3)=(1\*(log(3/2)^2)/(10^1/29\*( (log(3/2)^2)= 1/10^1/2

2.



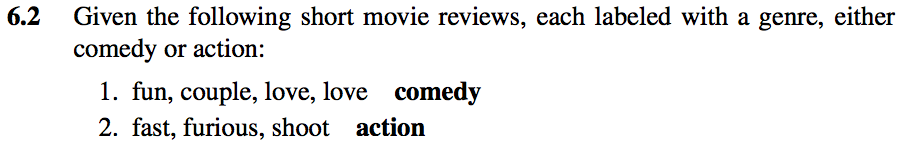
Solution:

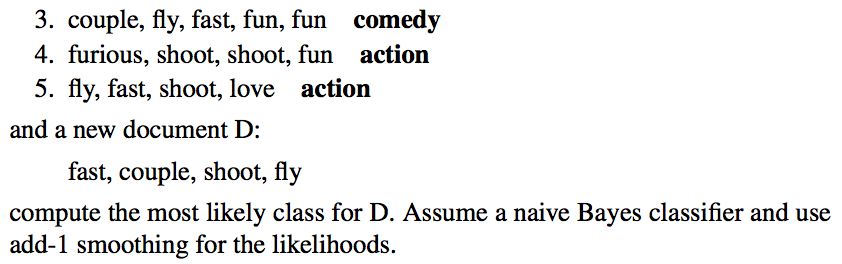
Set test sentence S = ” I always like foreign films”

P(-)P(S|-) = ½ \*0.16\*0.06\*0.06\*0.15\*0.11 = 4.75\*10^(-6)

P(+)P(S|+) = ½\*0.09\*0.07\*0.29\*0.04\*0.08=2.92\*10^(-6)

The model thus predicts the class negative for the test sentence.



Solution:

Training Cat Documents

C fun, couple, love, love

C couple, fly, fast, fun, fun

A fast, furious, shoot

A furious, shoot, shoot, fun

A fly, fast, shoot, love

Test ? fast, couple, shoot, fly

The prior P(c) for the two classes:

P(C) = 2/5 P(A) = 3/5

P(“fast”|C) = (1+1)/(9+7) P(“fast”|A) = (2+1)/(11+7)

P(“couple”|C) = (2+1)/(9+7) P(“couple”|A) = (0+1)/(11+7)

P(“shoot”|C) = (0+1)/(9+7) P(“shoot”|A) = (4+1)/(11+7)

P(“fly”|C) = (1+1)/(9+7) P(“fly”|A) = (1+1)/(11+7)

P(C)P(S|C) = (2/5)\*(2\*3\*1\*2)/16^4 = 7.324\*10^(-5)

P(A)P(S|A) = (3/5)\*(3\*1\*5\*2)/18^4 = 17.147\*10^(-5)

The model thus predicts the class ‘Action’ for the test sentence.

Q3.

