HANDOUT5

Example: Calculate cosine similarity for words I and like

vec(I) = [0,2,1,0,0,0,0,0]
vec(like) = [2,0,0,1,0,1,0,0]

cos(vec(I),vec(like)) =
sum(0*2+2*0+1*0+0*1+0*0+0*1+0*0+0*0)/det(vec(I))*det(vec(like))

= 0

Example: L = About three years ago, he nearly gave up because he had nothing to sell; now his shelves are full, and towels and clothes hang from a line overhead.

1- Give a collocational feature vector for the word line in L, given a window size of 3 words to the left and 3 words to the right.(Â±3)
Collocational features: position specific information about the words and collocations in window

[wi-3,wi-2,wi-1,wi+1]
[hang,from,a,overhead]

We can also use pos tags in the feature set. First we need to tag these words.

[wi-3,pi-3,wi-2,pi-2,wi-1,pi-1,wi+1,pi+1]
[hang,pos_hang,from,pos_from,a,pos_a,overhead,pos_overhead]

2- Give a bag-of-words feature vector for the word linein L, given the following word feature list: [written, school, speech, row, major, hang, sell, nothing, rope, words]

[written, school, speech, row, major, hang, sell, nothing, rope, words]
[0,0,0,0,0,0,1,1,0,0]

Example:Give a collocational feature vector for the word accident in L, given a window size of 5 words to the left and 3 words to the right.(Â±5)

[wi-5,pi-5,wi-4,pi-4,wi-3,pi-3,wi-2,pi-2,wi-1,pi-
1,wi+1,pi+1,wi+2,pi+2,wi+3,pi+3,wi+4,pi+4,wi+5,pi+5]

[to,TO,restart,VB,some,DT,sort,NN,of,IN,and,CC,emergency,NN,service,NN,appear,VB,to
,TO]

Example: Bag-of-words vector for "watch"

What is the correct binary vector representation fro sentences(Â±10)

[likes,movies,time,escape,football,wrist,prison,night]
sent1: [1,0,0,0,1,0,0,0] (Mary,also,likes,to,football,games)
sent2: [0,0,1,0,0,0,0,0] (John's new,shows,the,t,me,in,five,locations)
sent3: [0,0,0,0,0,0,1,0] (clearly,shows,that,no-
one,has,escaped,the,prison,during,my)
sent4: [0,1,0,0,0,0,0,0]
(scene,where,the,actors,go,out,of,frame,and,you,their,shadows,fighting,at,least,in,
cliche,movies)

Example: Cosine Similarity

cat:
The cat was playing in the garden. (the,was,playing,in)
The owner feed her cat every morning. (owner,feed,her,every,morning)
You can find cat food in the markets. (you,can,find,food,in,the)
The cat often eats in the morning. (the,often,eats,in)
They were fighting like a cat and a dog. (fighting,like,a,and,a,dog)
How much shoul I feed my cat? (i,feed,my)
Her cat was always sleeping. (her,was,always)

dog:
The family's cat and dog are playing in the garden. (family's,cat,and,are playing,in)
Encourage your dog to play in the garden. (encourage,your,to,play,in)
Dog food is not sold here. (food,is,not)
His dog does not eat meat. (his,does,not,eat)
The dog was hit by a car. (the,was,hit,by)
I never feed my dog raw meat. (never,feed,my,raw,meat)

Compute cosine similarity(Â±3)

The vector:

[the,was,playing,in,owner,feed,her,every,morning,you,can,find,food,often,eats,fighting,like,a,and,dog,i,my,always,family's,cat,are,encourage,your,to,play,is,not,his,does,eat,hit,by,never,raw,meat]

vec(cat):

[3,2,1,3,1,2,2,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0]

vec(dog):

[1,1,1,2,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,1,0,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1]

Example: Naive Bayes

Letâ€™s walk through an example of training and testing naive Bayes with add-one smoothing. Weâ€™ll use a sentiment analysis domain with the two classes positive(+) and negative(-), and take the following miniature training and test documents simplified from actual movie reviews.

Training

(-) just plain boring
(-) entirely predictable and lacks energy
(-) no surprise and very few laughs
(+) very powerful
(+) the most fun film of the summer

Test
(?) predictable with no fun

p(-)=3/5 and p(+)=2/5
|V|=20,N- = 14 and N+ = 9

p(class|predictable with no fun) = argmax(p(predictable with no

fun|-),p(predictable with no fun|+))

P.S. We apply add-one smoothing

p(predictable with no fun|-) = p(-)*p(predictable|-)*p(with|-)*p(no|-)*p(fun|-) = 3/5 * 2/34 * 1/34 * 2/34

p(predictable with no fun|+) = p(+)*p(predictable|+)*p(with|+)*p(no|+)*p(fun|+) = 2/5 * 1/29 * 1/29 * 1/29 * 2/29