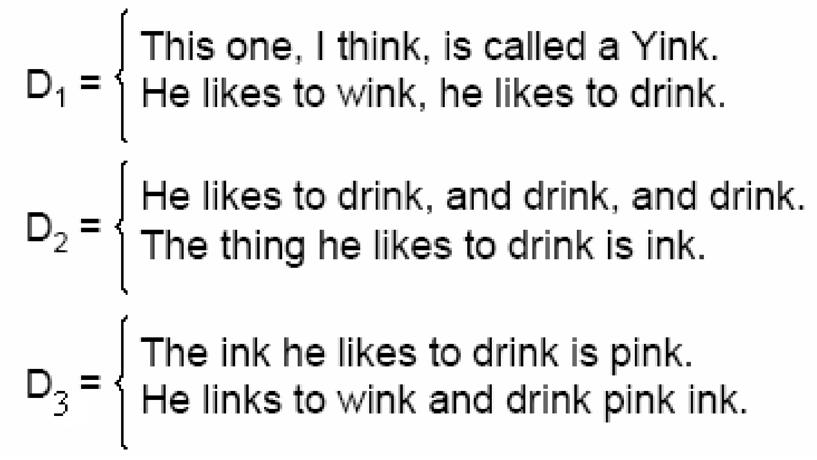
IR-HW02

Page 1



D_1 = This one, I think, is called a Yink.
He likes to wink, he likes to drink.

D_2 = He likes to drink, and drink, and drink.
The thing he likes to drink is ink.

D_3 = The ink he likes to drink is pink.
He links to wink and drink pink ink.

Q1 = drink　Q2=wink drink　Q3=pink ink

This one,I think, is called a Yink.
He likes to wink, he likes to drink.

- D1 长度为 16

词表

this	one	I	think	is	called	a	yink	he	likes	to	wink	drink	pink	ink

共 15 个词

修正之前：

P(drink|D1)=1/16

P(wink|D1)=1/16

P(pink|D1)=0

P(ink|D1)=0

P(Q1|D1) = 1/16

P(Q2|D1) = 1/16 * 1/16 = 1/256 = 0.004

P(Q3|D1) = 0*0=0

修正后：

(1)Laplace correction：

P(drink|D1)=(1+1)/(16+15)=2/31

P(wink|D1)=(1+1)/(16+15)=2/31

P(pink|D1)=(0+1)/(16+15)=1/31

P(ink|D1)= (0+1)/(16+15)=1/31

P(Q1|D1) = 2/31

P(Q2|D1) = 2/31 * 2/31 =4/961

P(Q3|D1) =1/31 * 1/31 =4/961

(2)Lindstone corrention:

P(drink|D1)=(1+0.001)/(16+15*0.001)=1001/16015

P(wink|D1)=(1+0.001)/(16+15*0.001)= 1001/16015

P(pink|D1)=(0+0.001)/(16+15*0.001)= 1/16015

P(ink|D1)= (0+0.001)/(16+15*0.001)= 1/16015

P(Q1|D1) =1001/16015

P(Q2|D1) = 1001/16015* 1001/16015 = 0.004

P(Q3|D1) =1/16015 * 1/16015 =4*10^-9

(3)Absolute Discounting:

P(drink|D1)=(1-0.001)/16=999/16000

P(wink|D1)=(1-0.001)/16=999/16000

P(pink|D1)=(0+0.001)/16=1/16000

P(ink|D1)= (0+0.001)/16=1/16000

P(Q1|D1) =999/16000

P(Q2|D1) = 999/16000 * 999/16000 = 0.004

P(Q3|D1) =1/16000 * 1/16000=4*10^-9

- D2 长度为 16

词表

he	likes	to	drink	and	the	thing	is	ink	wink	pink

共 11 个词

修正之前:

P(drink|D2)=4/16

P(wink|D2)=0

P(pink|D2)=0

P(ink|D2)=1/16

P(Q1|D2) = 4/16

P(Q2|D2) = 0 * 4/16 = 0

P(Q3|D2) = 0 * 1/16=0

修正后:

(1)Laplace correction:

P(drink|D2)=(4+1)/(16+11)=5/27

P(wink|D2)=(0+1)/(16+11)=1/27

P(pink|D2)=(0+1)/(16+11)=1/27

P(ink|D2)= (1+1)/(16+11)=2/27

P(Q1|D2) = 5/27

P(Q2|D2) = 1/27 * 5/27 =5/729

P(Q3|D2) =1/27 * 2/27= 2/729

(2)Lindstone corrention:

P(drink|D2)=(4+0.001)/(16+11*0.001)=4001/16011

P(wink|D2)=(0+0.001)/(16+11*0.001)= 1/16011
P(pink|D2)=(0+0.001)/(16+11*0.001)= 1/16011
P(ink|D2)= (1+0.001)/(16+11*0.001)= 1001/16011

P(Q1|D2) = 4001/16011
P(Q2|D2) = 1/16011* 4001/16011 = 1.6*10^-5
P(Q3|D2) = 1/16011* 1001/16011 =4*10^-6

(3)Absolute Discounting:
P(drink|D2)=(4-0.001)/16=3999/16000
P(wink|D2)=(0+0.001)/16=1/16000
P(pink|D2)=(0+0.001)/16=1/16000
P(ink|D2)= (1-0.001)/16=999/16000

P(Q1|D2) =3999/16000
P(Q2|D2) = 1/16000 * 3999/16000 = 1.6*10^-5
P(Q3|D2) =1/16000 *999/16000=4*10^-6

- D3 长度为 16

词表

the	ink	he	likes	to	drink	is	pink	links	wink	and

共 11 个词
修正之前:
P(drink|D3)=2/16
P(wink|D3)=1/16
P(pink|D3)=2/16
P(ink|D3)=2/16

P(Q1|D3) = 2/16
P(Q2|D3) = 1/16 * 2/16 = 2/256
P(Q3|D3) = 2/16 * 2/16=4/256
修正后:
(1)Laplace correction:
P(drink|D3)=(2+1)/(16+11)=3/27
P(wink|D3)=(1+1)/(16+11)=2/27
P(pink|D3)=(2+1)/(16+11)=3/27
P(ink|D3)= (2+1)/(16+11)=3/27

P(Q1|D3) =3/27
P(Q2|D3) = 2/27 * 3/27 =6/729
P(Q3|D3) =3/27 * 3/27= 9/729

(2)Lindstone corrention:
P(drink|D3)=(2+0.001)/(16+11*0.001)=2001/16011

P(wink|D3)=(1+0.001)/(16+11*0.001)= 1001/16011
P(pink|D3)=(2+0.001)/(16+11*0.001)= 2001/16011
P(ink|D3)= (2+0.001)/(16+11*0.001)= 2001/16011

P(Q1|D3) = 2001/16011
P(Q2|D3) = 1001/16011* 2001/16011 =0.0078
P(Q3|D3) = 2001/16011* 2001/16011 =0.016

(3)Absolute Discounting:
P(drink|D3)=(2-0.001)/16=1999/16000
P(wink|D3)=(1-0.001)/16=999/16000
P(pink|D3)=(2-0.001)/16=1999/16000
P(ink|D3)= (2-0.001)/16=1999/16000

P(Q1|D3) =1999/16000
P(Q2|D3) = 999/16000 * 1999/16000 = 0.0078
P(Q3|D3) =1999/16000 * 1999/16000=0.016

Page 2

1. 在使用文档的交叉熵进行检索时，为什么要用$H(M_Q||M_D)$而不是$H(M_D||M_Q)$?
首先，从公式上看，这两个公式是不对称的：
$$H(M_Q||M_D) = \sum_w P(w|M_Q) log P(w|M_D)$$
$$H(M_D||M_Q) = \sum_w P(w|M_D) log P(w|M_Q)$$
因而，采取不同的形式计算的结果并不相同。
从语义上看，$H(M_Q||M_D)$表示用M_D编码M_Q所需要的平均编码长度。对于一次检索而言，意指确定一个查询 Query 及其语言模型M_Q,使用文档集中的各个文档的语言模型M_D对M_Q进行编码，并由短到长进行排序，实现一次检索。这是$H(M_Q||M_D)$的实际意义。
反之，若使用$H(M_D||M_Q)$，其含义为对某一篇确定文档及其语言模型M_D，用多个 Query 的语言模型M_Q去描述，这不符合检索的定义。

2. 对一个给定的文档集，对每个文档构建其 uni-gram 模型，列出出现概率最高的 10 个词。
DOC1：Computers on display in Fuzhou. Display of a computer in Sidney. Playing a

computer in Sidney. Fuzhou computer store in debt. Sidney Science Fair .(24words)
DOC2: new home sales top forecasts. home sales rise in July. increase in home sales in July. july old home sales rise.(21words)
DOC3: breakthrough drug for schizophrenia. new schizophrenia drug. new approach for treatment of schizophrenia. new hopes for schizophrenia patients .(18words)

DOC1		DOC2		DOC3	
in	4/24	sales	4/21	schizophrenia	4/18
Sidney	3/24	home	4/21	new	3/18
computer	3/24	July	3/21	for	3/18
Fuzhou	2/24	in	3/21	drug	2/18
display	2/24	rise	2/21	treatment	1/18
a	2/24	top	1/21	patients	1/18
store,,science, playing,on, of,fair,debt, computers	1/24	old	1/21	of	1/18
		new	1/21	hopes	1/18
		increase	1/21	breakthrough	1/18
		forecasts	1/21	approach	1/18

部分 DOC 来自教材与互联网。