Speech and Text Processing

- Dom Jurafsky James Hartin

Chapter 2 : Regular Enpressions, Tent Normalization, Edit Distance · Tent Normalization tokenizing

lammatization

Stemming

Sentence segmentation

Exercises (92.1) (22.2) (a) (·+) 1 (a) 16 [a-z A-z] + 16 (b) \b[a-z]*b\b (b) 1[0-9]+.*[A-Za-z]*\. (c) \b b + (ab+) + \b (c) bgrotto b. * braven b 15 raven 16 * 15 grotto 16 (Q2.4) # deal # 0 1 2 3 l 1 1 2 3 4 e 2 2 1 2 3 232223 043333 (Q2.5) # b r i e f # d i v c 8 s #012 3 # 0 1 2 3 4 5 6 5 d 1012345 0 1 2 3 4 5 6 7 2 1 2 3 4 5 6 Y 2 3 2 3 4 5 i 3 2 1 2 3 4 5 3 4 3 2 3 4 v 4 5 4 3 4 0 4 3 2 1 2 3 4 e 5 4 3 2 1 2 3 e 5 6 5 4 5 6

Chapter 3: N-gram language Models Exercises $(\omega_{n-2}\omega_{n-1}\omega_n)$ (93.1) P(wn | wn-2 wn-1) = C (WN-2 WN-1) (5) (5) I am Sam </5> <5><5> Sam I am <15> <5><5> I do not like green eggs and ham <15> p(I|<5><5>) = 2/3 $P(am) \leq S = 1/2$ (83.2) P(i want chinese food) = P(il<s>) P(want Ii) P(Chinese | want) P(food | chinese) b (<12>/ feog) = 0.25 x .33 x 0 0065 x .52 x 0.68 = 0.0001896 P(i want chinese food) = P(i <<>>) P(want | i) P(dinese | want) P(food | chinese) P(<18> | food) = .19 x 0.21 x 0.0029 x 0.052x .4 = 0.00000 2406 (82.3) The unsmoothed probability is higher because the bigrams used in the sentences are very common and has probablities. nowever, in the smoothed case, their probablities are distributed among net - so-common bigrans which are not used in our test statement.

(03.4)	<5) I	am	Sam	do	not	0.0	green		anel	<15>
(\$3 1)	1 4	1	2	1	1	1	green 1	388	anux 1	1
I	1 1	4	エ	2	1	1	1	1	1	1
am	1 1	1	3	1	ュ	1	1	1	ュ	2_
Sam	1 2	1	ュ	1	1	1	1	ᅱ	3	4
do	" 1	٠,	١,	"	2	۳	"	n	17	1
not	е "	14	1.	41	7	2	11	V	' -	(1)
üke	/1 \	(,	١,	"	**	1	2_	,	'n	7
green	^ ~	11	11	"	.,	11	1	2_	,	11
ब्रुष्ट	n 4	11	17	"	<i>\</i> \	12	"	1	2	"
and	^ "	^	2	y	"	**	"	į	1	4
<15>	" "	1,	1	\ 	,,	11	"	"	11	~
P (Sam	n lam)	٤	C (our		1) =	3	=	.214	1	
(Q 3.5) <s< th=""><th>a</th><th>ط</th><th></th><th></th><th></th><th></th><th></th><th></th><th></th><th></th></s<>	a	ط								
<s> 0</s>	2_	2								
a 0	1	1								
ЪО	1	1								
	-	<u></u>								
• P(a			a <5	>) f	>(a)a	·) =	0 · 5 ×	: 0.5	= 0.	25
• P(a	a) =	P (_ al <s< td=""><td>>)</td><td>PLbia</td><td>L) =</td><td>0.5×</td><td>0.5</td><td>= 0</td><td>25</td></s<>	>)	PLbia	L) =	0.5×	0.5	= 0	25
P (al	a) =	P (_ al <s< td=""><td>>)</td><td>PLbia</td><td>L) =</td><td>0.5×</td><td>0.5</td><td>= 0</td><td>25</td></s<>	>)	PLbia	L) =	0.5×	0.5	= 0	25
p (a b	a) =	P (P (_ al <s b \ < s</s 	>) >)	P (6 1 1	(a) =	0.5	x 0.5	= 0	25 ·25
P (b b	a) = c) = c) =	P (P (P (al <s b < s b < s</s 	>) >)	P (6 1 1	(a) =	0.5	x 0.5	= 0	25 ·25

$$(Q_3 \cdot 7) = P(sam \mid am) = d_1 P(sam) + d_2 (sam \mid am)$$

$$= \frac{1}{2} \times \frac{42}{25} + \frac{1}{2} \times \frac{2}{3}$$

$$= \frac{2}{25} + \frac{1}{3} = 0.41$$

 $\rho(0) = \frac{91}{400}$

P(1) = P(2) P(9) = 1

(03.6) $P(\omega_3|\omega_1\omega_2) = C(\omega_1\omega_2\omega_3) + 1$

(93.12) $PP(\omega) = \sqrt{\frac{N}{11}} \frac{1}{P(\omega_i)}$

= / (100) to / (100)

= 1.726

Chapter 4: Naive Bayes and Sentiment Classification

(94.1)
$$S =$$
 "I always like foreign films"
$$P(\text{neg }|S) = \frac{P(S|\text{neg})}{P(S)}$$

$$P(pos(s) = \frac{P(s|pos)P(pos)}{P(s)}$$

The naive bayes will assign "nez" class to the sentence because
$$P(nez|S) > P(pos|S)$$

= ignow common

(94.2)
$$P(\text{camedy}) = 2/5$$
 $P(\text{action}) = 3/5$
 $P(\text{fast} | \text{camedy}) = \frac{\text{Court}(\text{fast}, \text{camedy}) + 1}{\text{E}(\text{count}(\text{w}, \text{camedy}) + 1)}$
 $= \frac{2}{3+7} = \frac{2}{16}$
 $P(\text{fast} | \text{camedy}) = \frac{3}{16} P(\text{shoot} | \text{camedy}) = \frac{1}{16}$
 $P(\text{couple} | \text{camedy}) = \frac{3}{16} P(\text{shoot} | \text{camedy}) = \frac{1}{16}$
 $P(\text{couple} | \text{action}) = \frac{1}{18} P(\text{shoot} | \text{action}) = \frac{5}{18}$
 $P(\text{fly} | \text{correcty}) = \frac{2}{16}$
 $P(\text{fly} | \text{correcty}) = \frac{2}{16}$
 $P(\text{correcty} | D) = \frac{P(D(\text{comedy})) P(\text{correcty})}{16}$
 $P(\text{correcty} | D) = \frac{2}{16} P(\text{correcty}) P(\text{correcty}) = \frac{2}{16} P(\text{action}(0)) = \frac{3}{18} P(\text{correcty}) = \frac{2}{16} P(\text{action}(0)) = \frac{3}{18} P(\text{acti$

(9 43) • Binarized maive Bayes

$$P(neg) = 0.6$$
 $P(pos) = 0.4$
 $P(grad | neg) = 3/9$
 $P(grad | pos) = 2/7$
 $P(grad | neg) = 4/9$
 $P(port | pos) = 2/7$
 $P(great | neg) = 2/3$
 $P(great | pos) = 3/7$
 $P(neg(0) = \frac{3}{3} \times \frac{1}{9} \times \frac{2}{3} \times 0.6 = 0.0197$
 $P(pos | 0) = \frac{2}{3} \times \frac{1}{7} \times \frac{3}{7} \times 0.4 = 0.0139$

(lawified as "neg" by BNB.

• Multinomial bouve Bayes

 $P(good | pos) = 4/12$
 $P(good | neg) = 3/17$
 $P(port | pos) = 2/12$
 $P(good | neg) = 11/17$
 $P(good | neg) = 11/17$
 $P(good | neg) = 11/17$
 $P(good | neg) = 3/17$
 $P(good | neg) = 3/17$

(Chapter	2 (0:																						
C	Q (0·1)		REF 14P2													m, 10)				
			nigra																					. `	
		9,6	jr a vv	<u>.</u>	rve.t	TV		in	•	pa	ue			λi	ιτ.	EN	ne	29	3,5			ρ.,	Zvavv	us /	
			gran								U	igr	øvn.	R	:	11/1	8								
		Вк	gram	ı P	=	9 [D				B	gra	m	R	:	3/1	7								
			c	hr f	> =	(u/	/u +	9)	lo)	1	2.	=	•	95										
											Ů														
			d	hof	ર દ		Ċι	ı /18	+	3/1	7)	/2	=	C	9•	57									
			Ch	~f2	2, 2		5	L	ne f	·*	d	&R				5	ו	95	×ε) ·S=	-				
								ı	1 ch	r P	+	chr	R			_	4 x •	g <i>s</i>	+ 0	· 5 7	-				
															=	0.	6 l	99							

Chapter 12:	Constituency	Grammass		
(Q _{12·1})				
	ar wed:			
2	→ MP VP N	19 49 1 AV 90		
		Proper-Noun Det	t Nominal 1	somenal (Adj
		ral N Adj N		<u> </u>
Yf	P -> Y IU NP	1 1 NP PP Y P	የ	
PI	P -> Prepositio	m NP		
lenic	els used:			
	Adi -> early	all one-war	n I any	
	N → 1·m	all one-war flights redeye	2 forc dela	ys I five
(Oet → a	0 0	U	0 0
	U → arriv			
ρ	reposition -> fo	som I in lafter	1 on	
P	roper-noun -> (enver Dallas	Washington 1 Th	rusday
	roper-Noun -> (enver Dallas	Washington) Th	
(a) Dallas	toper-Noun > (b) from penver	enver Oallas (c) after five	Washington) Th	iving in washington
	roper-Noun -> (enver Dallas	Washington) Th	
(a) Dallas	toper-Noun > (b) from penver g l pp	enver Oallas (c) after five	washington 1 Th	iving in washington
(a) Dallos	toper-Noun > (b) from penver g PP Preposition NP	c) after five keposition NP	p.m (d) are	riving in washington Preposition NP
(a) Dallas	toper-Noun > (b) from Denver g p p preposition NP from hoper-N	c) after five keposition NP	washington The	freposition ISP in Proper-Noun
(a) Dallos	toper-Noun > (b) from penver g PP Preposition NP	c) after fice (c) after fice S PP Reposition NP ofter Non	washington The	riving in washington Preposition NP
(a) Dallas	toper-Noun > (6) from penver g PP Preposition NP From Proper-N Denver	c) after five pp heposition NP ofter Non nonival	washington The p.m (d) are arriving	reposition NP freposition NP freposition NP kashington
(a) Dallos	toper-Noun > (b) from Denver g p p p p p p p p p p p p p p p p p p	c) after five pp heposition NP ofter Non nonival	washington The	freposition ISP in Proper-Noun
(a) Dallas	toper-Noun > (6) from penver g PP Preposition NP From Proper-N Denver	e flight (g) on	Washington The p.m (d) are arriving inal Thursday	freposition ISP in Proper-Isour in Proper-Isour Ch) a one-way - flight
(c) early flights	toper-Noun > (to from penver g pe pe pe pe pe pe pe pe pe	enver Oallos (c) after fice S PP Reposition NP ofter Nominal Nominal Nominal E feight (g) or	washington The pm (d) are arriving inal Thursday 8	Preposition NP
(c) early flights	toper-Noun > (to from penver g pe pe pe pe pe pe pe pe pe	enver Oallos (c) after fice S PP Reposition NP ofter Nominal N five e flight (g) an	washington The p.m (d) are arriving inal the pm	Preposition NP in Proper-Noun washington (h) a one-way-fright S NP
(c) early fights	toper-Noun > (b) from Denver g p p perposition NP from Proper-N Denver (f) all redey NP Non	c) after five (c) after five S PP Reposition NP Ofter Non Nonival e flights (g) an	washington The p.m (d) are arriving inal Thursday s p reposition NP	freposition ISP in Proper-Noun I Noshington (h) a one-way - feight S INP Det Nominal

) oney del	aye in denver						
Adj N	l in	NP Propos-Noun Denver					
12 · 6)	Poss	sessive	→ NP	's	Noun		
	Possers	ive				Possessiu-	e
NP	's	Noun			NP	2	Noun
Proper-	Moun	office		Det	100	minal	mark
fortu	ne			my	u	ncle	
12.8)							
	Step 1:	Remove as	ll termi	hals	from	right ha	nd side it
		they occu	cy with	٥	now-t	erminal.	
	Example:						
		A -> bC	. ⇒			c	
				В	ط حـ		
	Step 2:	on rules	with more	. thoun	. <u>a</u> .	non-termi	nals on RNS,
		vote ne					
						_	
	Example:	A -> BC	0 ->		, → E		