Numericals based on N-gram model

Q1. Find the total count of unique bigrams for which likelihood will be estimated.

Alice went to the café

Bob was waiting for Alice

Alice and Bob went to the museum

Total Bigrams = 20

Unique Bigrams = 17

Soln – <s> Alice went to the cafe </s>

<s> Bob was waiting for Alice </s>

<s> Alice and Bob went to the museum </s>

<s> Alice	<s> Bob	<s> Alice
Alice went	Bob was	Alice and
went to	was waiting	and Bob
to the	waiting for	Bob went
the cafe	for Alice	went to
cafe </s>	Alice </s>	to the
		the museum
		museum </s>
6	6	5

Q2.Find the probability of the statement:<S> Michael and Zack played at the playground</s> from the following corpus. Assume a trigram Language model.

<S>the school was open</S>

<S>Michael and Zack went to the school</S>

<S>the playground at the school was huge</S>

<S>Bob and Zack played at the playground</S>

<S>Bob, Michael and Zack were friends</S>

Solⁿ – Calculate the trigram's probability

$P(\text{Michael} \mid \langle s \rangle) = \frac{1}{5}$

$P(\text{and} \mid \langle s \rangle\ \text{Michael}) = \frac{1}{1}$

$P(\text{Zack} \mid \text{Michael and}) = \frac{2}{2}$

$P(\text{played} \mid \text{and Zack}) = \frac{1}{3}$

$P(\text{at} \mid \text{Zack played}) = \frac{1}{1}$

$P(\text{the} \mid \text{played at}) = \frac{1}{1}$

$P(\text{playground} \mid \text{at the}) = \frac{1}{2}$

$P(\langle /s \rangle \mid \text{the playground}) = \frac{1}{2}$

$P(\langle s \rangle\ \text{Michael and Zack played at the playground}\ \langle /s \rangle)$

$= \frac{1}{5} \times \frac{1}{1} \times \frac{2}{2} \times \frac{1}{3} \times \frac{1}{1} \times \frac{1}{1} \times \frac{1}{2} \times \frac{1}{2}$

$= \frac{1}{60}$

Q3.What is the perplexity of the statement in Q2

Sol^n

$P(S)$ = Probability $\quad n$ = Total token (including <s> and </s>)

$$\text{Perplexity} = P(S)^{-1/n}$$

$$= \left(\frac{1}{60}\right)^{-1/9}$$

$$= 1.576$$

Q4.For the given corpus,find the bigram probability estimate by laplace model for P(do|<s>),P(do|Sam),P(Sam|<s>),P(Sam|do),P(I|Sam),P(I|do),P(like|I)

<S>I am Sam</S>

<S>Sam I am</S>

<S>Sam I like</S>

<S>Sam I do like</S>

<S>do I like Sam</S>

Soln

Tokens	Frequency
<s>	5
</s>	5
I	5
am	2
Sam	5
like	3
do	2

Unique tokens = 7

$$P(do|<s>) = \frac{1+1}{5+7} = \frac{2}{12} = \frac{1}{6}$$

$$P(do|Sam) = \frac{0+1}{5+7} = \frac{1}{12}$$

$$P(Sam|<s>) = \frac{3+1}{5+7} = \frac{4}{12} = \frac{1}{3}$$

$$P(Sam|do) = \frac{0+1}{2+7} = \frac{1}{9}$$

$$P(I|Sam) = \frac{3+1}{5+7} = \frac{4}{12} = \frac{1}{3}$$

$$P(I|do) = \frac{1+1}{2+7} = \frac{2}{9}$$

$$P(like|I) = \frac{2+1}{5+7} = \frac{3}{12} = \frac{1}{4}$$

Q5.For a corpus ,MLE for bigram “battery life” is 0.27,frequency of “battery” is 800.After applying Laplace smoothening the MLE for “battery life” becomes 0.025.What is the vocabulary size of the corpus?

Soln

$$P(w_i \mid w_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}$$

$$P_{MLE}(\text{Battery Life}) = \frac{f(\text{battery life})}{f(\text{battery})}$$

$$0.27 = \frac{f(\text{battery life})}{800}$$

$$f(\text{battery life}) = 800 \times 0.27 = 216$$

With Laplace smoothening

$$P_{MLE}(\text{life} \mid \text{Battery}) = \frac{f(\text{battery life}) + 1}{f(\text{battery}) + V}$$

$$0.025 = \frac{216+1}{800+V}$$

$$\therefore \boxed{V = 7880}$$

Q7. Find the edit distance between INTENTION and EXECUTION using dynamic programming.

insert – 1
delete – 1
substitute – 1 or 2

$$D(i,j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \begin{cases} 2 & \text{if } S_1(i) \neq S_2(j) \\ 0 & \text{if } S_1(i) = S_2(j) \end{cases} \end{cases}$$

Edit distance

N	9	8	9	10	11	12	11	10	9	8
O	8	7	8	9	10	11	10	9	8	9
I	7	6	7	8	9	10	9	8	9	10
T	6	5	6	7	8	9	8	9	10	11
N	5	4	5	6	7	8	9	10	11	10
E	4	3	4	5	6	7	8	9	10	9
T	3	4	5	6	7	8	7	8	9	8
N	2	3	4	5	6	7	8	7	8	7
I	1	2	3	4	5	6	7	6	7	8
#	0	1	2	3	4	5	6	7	8	9
	#	E	X	E	C	U	T	I	O	N

D(i, j-1)

D(i-1, j-1)

D(i-1, j)

1+1
1+1
0+2

2+1
2+1
1+2

7+1
7+1
6+0

Q8 In the sentence-"The only thing we have to fear is fear itself" , the ratio between total number of word token and word types is:

Soln

Token = 10

Type = 9

Ratio of $\frac{\text{Token}}{\text{Type}} = \frac{10}{9}$

Q9 Let the rank of two words W1 and W2 in a corpus be 1600 and 400 respectively.Let m1 and m2 represent the number of meanings of w1 and w2 respectively. The ratio m1:m2 would be?

Soln

Zipf's Law

$M \propto \sqrt{F}$, $F \propto \frac{1}{\text{rank}}$

$M_1 = \frac{1}{\sqrt{r_1}} = \frac{1}{\sqrt{1600}} = \frac{1}{40}$

$M_2 = \frac{1}{\sqrt{r_2}} = \frac{1}{\sqrt{400}} = \frac{1}{20}$

$\therefore M_1 : M_2 = 20 : 40$

$\boxed{M_1 : M_2 = 1 : 2}$

Q10 What is the size of unique words in a document where total number of words=1200 and k=3.71 , β=0.69

Soln

Heap's Law

$|V|$ = size of vocabulary = $k \cdot N^{\beta}$

$= 3.71(1200)^{0.69}$

$= 494.32$