

## N-gram model

PAGE NO.

DATE: / /

Probability of words in sentences

$$P(w_1, w_2, \dots, w_n) = \prod_i P(w_i | w_1, w_2, w_3, \dots, w_{i-1})$$

Unigram (1-gram) : No history is used

Bi-gram (2-gram) : one word history

Tri-gram (3-gram) : Two word history

Four-gram (4-gram) : Three word history

Extend upto N-grams

→ Generally in practical applications, Bi-gram, Tri-gram, four-gram are used

Unigram (1-gram) :

"about five minutes from" - - - - -

- Assume in corpus dinner word is present with highest probability

- Unigram doesn't take into account probabilities with previous words like from, minutes

- Unigram will predict dinner

"about five minutes from dinner"



Bigram(2-gram): one word history

$$P(w_1, w_2) = \prod_{i=2} P(w_i | w_{i-1})$$

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

"about five minutes from ----"

Assumption: Next word may be college,  
class

$$P(\text{college} | \text{about five minutes from})$$

$$= \frac{\text{count}(\text{about five minutes from college})}{\text{count}(\text{about five minutes from})}$$

$$P(\text{class} | \text{about five minutes from})$$

$$= \frac{\text{count}(\text{about five minutes from class})}{\text{count}(\text{about five minutes from})}$$



count (about five minutes from)

$$= P(\text{about} | \langle S \rangle) \times P(\text{five} | \text{about}) \times P(\text{minutes} | \text{five}) \times P(\text{from} | \text{minutes})$$

count (about five minutes from college)

$$= P(\text{about} | \langle S \rangle) \times P(\text{five} | \text{about}) \times P(\text{minutes} | \text{five}) \times P(\text{from} | \text{minutes}) \times P(\text{college} | \text{from})$$

count (about five minutes from class)

$$= P(\text{about} | \langle S \rangle) \times P(\text{five} | \text{about}) \times P(\text{minutes} | \text{five}) \times P(\text{from} | \text{minutes}) \times P(\text{class} | \text{from})$$

$P(\text{college} | \text{about five minutes from})$

$$= \frac{\text{count}(\text{about five minutes from college})}{\text{count}(\text{about five minutes from})}$$

$$= P(\text{college} | \text{from})$$

Next word depends only on previous word

$P(\text{class} | \text{about five minutes from})$

$$= \frac{\text{count}(\text{about five minutes from class})}{\text{count}(\text{about five minutes from})}$$

$$= P(\text{class} | \text{from})$$

- we will check whose probability is more if probability of class is more then next predicated word will be class



Tri-gram (3-gram) : Two-words history

$$P(w_1, w_2, w_3) = \prod_{i=3} (w_i | w_1, w_2)$$

$$P(w_i | w_{i-1}, w_{i-2}) = \frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$$

- As no of previous state (history) increases, it is very difficult to match that set of words in corpus
- Probability of larger collection of word is minimum. To overcome this problem generally Bi-gram model is used

Exercise 1 : Estimating Bigram Probabilities

What is the most Probable next word predicted by the model for the following word sequence?

Given corpus

<S> I am Henry </S>

<S> I like college </S>

<S> DO Henry like college </S>

<S> Henry I am </S>

<S> DO I like Henry </S>

<S> DO I like college </S>



Word Frequency

<S> 7

<I> 7

I 6

am 2

Henry 5

like 5

college 3

do 4

1) <S> do ?

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

Next word

probability Next word

$P(<I> | do)$

0/4

$P(I | do)$

2/4

$P(am | do)$

0/4

$P(Henry | do)$

1/4

$P(like | do)$

1/4

$P(college | do)$

0/4

$P(do | do)$

0/4

I is more Probable



2)  $\langle s \rangle$  I like Henry?

Next word prediction probability  $w_{i-1} = \text{Henry}$

Next word	Probability	Next word
-----------	-------------	-----------

$P(\langle s \rangle | \text{Henry})$

$3/5$

$P(I | \text{Henry})$

$1/5$

$P(\text{um} | \text{Henry})$

$0$

$P(\text{like} | \text{Henry})$

$0$

$P(\text{college} | \text{Henry})$

$0$

$P(\text{Henry} | \text{Henry})$

$0$

$P(\text{do} | \text{Henry})$

$0$

$\therefore \langle s \rangle$  is more probable

3)  $\langle s \rangle$  DO I like?

Use tri-gram

Here  $w_{i-2} = I$  and  $w_{i-1} = \text{like}$

What is the most probable next word predicted by the model?



Q: Which of the following sentence is better i.e. gets a higher probability with this model

- Use previous corpus [given in Exercise 1]
- Use Bi-gram

1.  $\langle S \rangle$  I like college  $\langle /S \rangle$

$$= P(I|\langle S \rangle) \times P(\text{like}|I) \times P(\text{college}|\text{like}) \times P(\langle /S \rangle|\text{college})$$

$$= \frac{3}{7} \times \frac{3}{6} \times \frac{3}{5} \times \frac{3}{3}$$

$$= \frac{9}{70}$$

$$= 0.13$$

2.  $\langle S \rangle$  do I like Henry  $\langle /S \rangle$

$$= P(\text{do}|\langle S \rangle) \times P(I|\text{do}) \times P(\text{like}|I) \times P(\text{Henry}|\text{like}) \times P(\langle /S \rangle|\text{Henry})$$

$$= \frac{3}{7} \times \frac{2}{4} \times \frac{3}{6} \times \frac{2}{5} \times \frac{3}{5}$$

$$= 0.0257$$

First statement is more probable