PAGE NO.
DATE: / /

N-gram model

Probability of words in sentences

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

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 predicat dinner

" about five minutes from dinner"

Bigram (2-gram): one word history

$$P(w_1, w_2) = \prod_{i=2} P(w_2 | w_1)$$

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

"about five minutes from ----"

Assumption: Next word may be college, class

P(college | about five minutes from)

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

P(class | about five minutes from)

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

count (about five minutes from)
= P(about | <s>) x P(five | about) x
P(minutes | five) x P(from | minutes)

count (about five minutes from college)
= P(about | <s>) x P(five | about) x
P(minutes | five) x P(from | minutes) x
P(college | from)

count (about five minutes from class)
= P(about | <s>) x P(five | about) x
P(minutes | five) x P(from | minutes) x
P(class | from)

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

$$= P(\text{college} \mid \text{from})$$ Next word depends only on previous word

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

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

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

PAGE NO.
DATE : / /

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

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

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

- 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 overcom this problem generally Bi-gram model is used

Exercise1 : Estimating Bigram Probabilities

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

Given corpus

<s> I do like Henry </s>
<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
</s>	7
I	6
am	2
Henry	5
like	5
college	3
do	4

1) <s> do ?

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

Next word	Probability Next word
P(</s> | 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) <s> I like Henry?

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

Next word	Probability
P(</s> | Henry)	3/5
P(I | Henry)	1/5
P(am | Henry)	0
P(like | Henry)	1/5
P(college | Henry)	0
P(Henry | Henry)	0
P(do | Henry)	0

∴ </s> is more probable

3) <s> DO I like?

Use tri-gram

Here w_{i-2} = I and w_{i-1} = 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]
- Use Bi-gram

1. <S> I like college </S>

$= P(I|<S>) \times P(like|I) \times P(college|like) \times P(</S>|college)$

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

$= \frac{9}{70}$

$= 0.13$

2. <S> do I like Henry </S>

$= P(do|<S>) \times P(I|do) \times P(like|I) \times P(Henry|like) \times P(</S>|Henry)$

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

$= 0.0257$

First statement is more probable