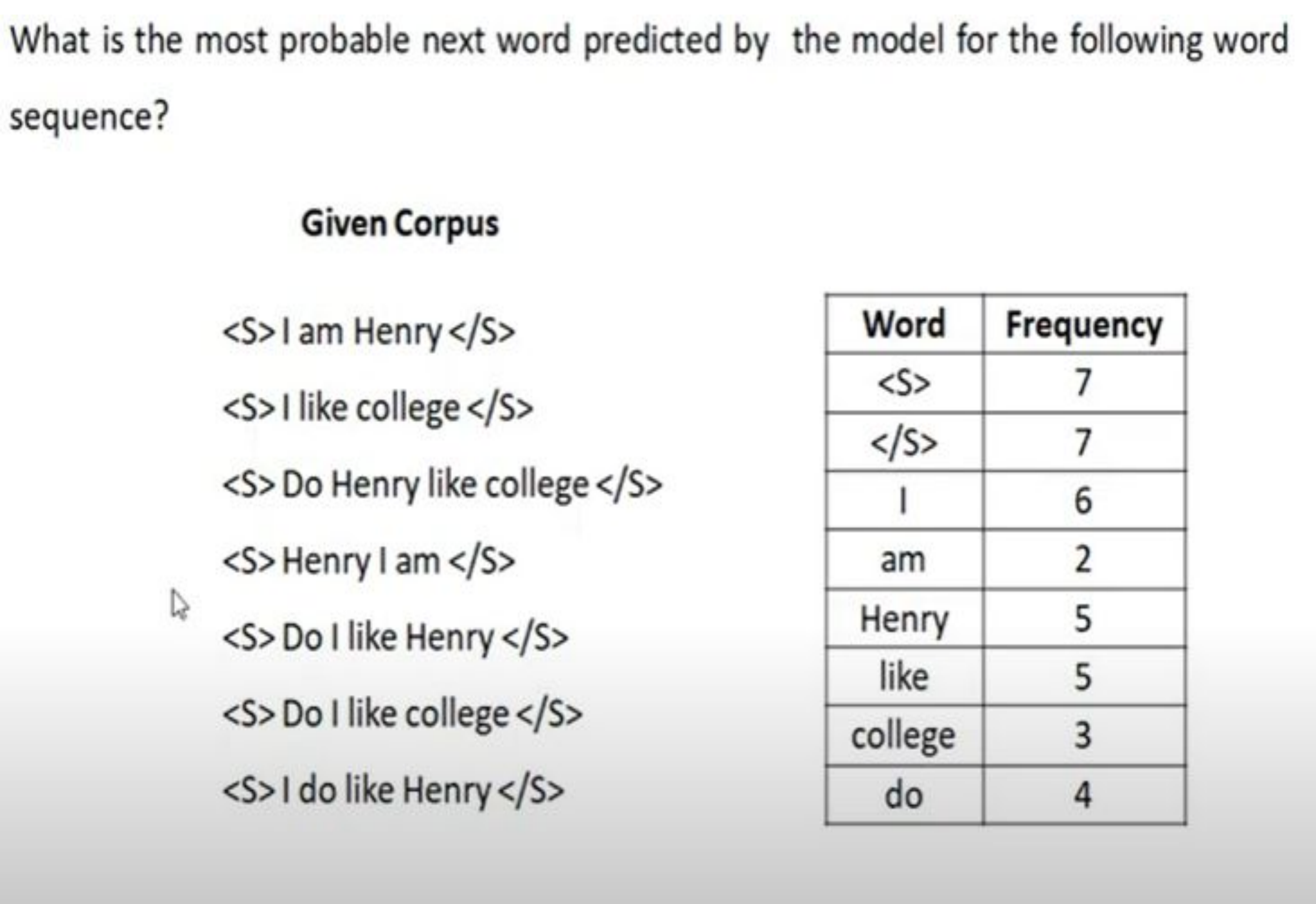
# N-grams

Exercises



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>

<S> I do like Henry </S>

| Word | Frequency |
|---|---|
| <S> | 7 |
| </S> | 7 |
| I | 6 |
| am | 2 |
| Henry | 5 |
| like | 5 |
| college | 3 |
| do | 4 |

1) <S> Do ?

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

| Word | Frequency |
|---|---|
| <S> | 7 |
| </S> | 7 |
| I | 6 |
| am | 2 |
| Henry | 5 |
| like | 5 |
| college | 3 |
| do | 4 |

**Next word prediction probability $W_{i-1}$=do**

| **Next word** | **Probability Next Word=** $\frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$ |
|---|---|
| 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 ?

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

| Word | Frequency |
|---|---|
| <S> | 7 |
| </S> | 7 |
| I | 6 |
| am | 2 |
| Henry | 5 |
| like | 5 |
| college | 3 |
| do | 4 |

**Next word prediction probability** **$W_{i-1}$=Henry**

| Next word | Probability Next Word= $\frac{N}{D} = \frac{\mathrm{count}(w_{i-1}, w_i)}{\mathrm{count}(w_{i-1})}$ |
|---|---|
| **P(</S> \| Henry)** | **3/5** |
| P(<I> \| Henry) | 1/5 |
| P(<am>\| Henry) | 0 |
| P(<Henry>\| Henry) | 0 |
| P(<like \| Henry) | 1/5 |
| P(<college \| Henry) | 0 |
| P(do \| Henry) | 0 |

**</S> is more probable**

3) <S> Do I like ?

Use Tri-gram

P<I like>=3

<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>

<S> I do like Henry </S>

Next word prediction probability

$W_{i-2}$=I and $W_{i-1}$=like

| Next word | Probability Next Word= $\frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$ |
|---|---|
| P(</S> \|I like) | 0/3 |
| P(<I> \| I like) | 0/3 |
| P(<am>\| I like) | 0/3 |
| P(<Henry>\| I like) | 1/3 |
| P(<like \| I like) | 0/3 |
| P(<college \| I like) | 2/3 |
| P(do \| I like) | 0/3 |

College is probable

4) <S> Do I like college ?

Use Four-gram

<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>

<S> I do like Henry </S>

**Next word prediction probability**

$W_{i-3}$=I, $W_{i-2}$=like $W_{i-1}$=college

| Next word | Probability Next Word= $\frac{\text{count}(w_{i-3,} w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-3}, w_{i-2}, w_{i-1})}$ |
|---|---|
| P(</S> \|I like college) | 2/2 |
| P(<I> \| I like college) | 0/2 |
| P(<am>\| I like college) | 0/2 |
| P(<Henry>\| I like college) | 0/2 |
| P(<like \| I like college) | 0/2 |
| P(<college \| I like college) | 0/2 |
| P(do \| I like college) | 0/2 |

</S> is more probable

Bi-gram(2-gram): **One word history**

$$P(w_1, w_2) = \prod_{i=2} P(w_2 \mid w_1) \qquad P(w_i \mid w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$$

"about five minutes from....."

Assumption: Next word may be college, class

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

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

“about five minutes from.....”

Count(about five minutes from)= P(about|<S>) × P(five| about) × P(minutes | five) × P(from| minutes)

Count(about five minutes from **college**)= P(about|<S>) × P(five| about) × P(minutes | five) × P(from| minutes) × P(college| from)

Count(about five minutes from **class**)= P(about|<S>) × P(five| about) × P(minutes | five) × P(from| minutes) × 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(college| from)**

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

**=P(class| from)**

Tri-gram(2-gram): **Two words history**

$$P(w_1, w_2, w_3) = \prod_{i=3} P(w_3 \mid w_1, w_2) \qquad P(w_i \mid w_{i-1}, w_{i-2}) = \frac{\text{count}(w_{i-2}, w_{i-1}, w_i)}{\text{count}(w_{i-2}, w_{i-1})}$$

Count(about five minutes from)= P(five|<S>, about) × P(minutes| about, five) ×
P(from| five , minutes)

Count(about five minutes from **college**)= P(five|<S>, about) × P(minutes| about, five) ×
P(from| five , minutes) × P(college| minutes from)

Count(about five minutes from **class**)= P(five|<S>, about) × P(minutes| about, five) ×
P(from| five , minutes) × P(class| minutes 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(college| minutes from)**

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

**=P(class| minutes from)**