Name: Harshit Yadav
Semester - 2:
Section - CSE
Roll No - 190913017

course name = Mtech. CSE
Signature = Harsh
Date = may 22, 2020

# NLP Assignment-3

Q consider the following toy example training data:-

<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>

Assume that we use a bigram language model based on the above training data

1) what is the most probable next word predicted by the model for the following word sequence ?

| | <S> | I | am | Sam | like | do | </S> |
|---|---|---|---|---|---|---|---|
| <S> | 0 | 0.2 | 0 | 0.6 | 0 | 0.2 | 0 |
| I | 0 | 0 | 2/5 | 0 | 2/5 | 1/5 | 0 |
| am | 0 | 0 | 0 | 0 | 0 | 0 | 1/2 |
| Sam | 0 | 3/5 | 0 | 0 | 0 | 0 | 0.4 |
| like | 0 | 0 | 0 | 1/3 | 0 | 0 | 0.667 |
| do | 0 | 1/2 | 0 | 0 | 1/2 | 0 | 0 |

a) $P(I/sam) = 3/5$

→ <s> sam ... will be followed by "I" next word

---

b) <s> Sam I do ...

$P(I/do) = 1/2$ $P(like/do) = 1/2$

→ I and Like next word

---

c) <s> Sam I am

$P(sam/am) = 1/3$

→ Sam is next word

---

d) <s> do I ---

$P(like/I) = 2/5$, $P(am/I) = 2/5$ $P(do/I) = 1/5$

, like , am next word

---

e) <s> Sam like ---

$P(sam/like) = 1/3 * 0$

next word cannot be calculated without smoothing.

---

2. calculate the probability after and after applying Laplacian smoothing

a) <s> do sam I like $|V| = 5$

$P(do|<s>) * P(sam|do) * P(I|sam) * P(like|I)$

$= \frac{2}{10} * \frac{1}{7} * \frac{4}{10} * \frac{3}{10}$

$= 0.34285 \times 10^{-2}$

b) <s> sam do I like

$= P(sam|<s>) * P(do|sam) * P(I|do) * P(like|I)$

$= \frac{4}{10} * \frac{1}{10} * \frac{2}{7} * \frac{3}{10}$

$= 0.34285 \times 10^{-2}$

Both sequence have equal probability