Name: Arooj Quiz 1(NLP) 25/25 Roll Number: 21L-5619

Q1. Write down the Vocabulary, number of tokens and number of types in the following text: 23.5+1.5 = 25

Solution: [2+2+1 marks]

Vocabulary:

{

بلی کو سمجھانے آئے، چوہے کئی ہزار
نے اک بات نہ مانی روئے
زار و زار سنو گپ شپ ناؤ میں
ندی ڈوب چلی شیر اور بکری
مل کر بیٹھے گھوڑا گھاس کھائے تینوں
اپنی ضد کے پورے کون کسے سمجھائے

Tokens: 42

Types: 38

بلی کو سمجھانے آنے
چوہے کئی ہزار
بلی نے اک بات نہ مانی
روئے زار و زار
سنو گپ شپ سنو گپ شپ
ناؤ میں ندی ڈوب چلی
شیر اور بکری مل کر بیٹھے
گھوڑا گھاس نہ کھائے
تینوں اپنی ضد کے پورے
کون کسے سمجھائے

consider as 1 token and type

Q2. Assume you have the following training corpus: [10+5 Marks]

<s> I am from Vellore </s>
<s> I am a teacher </s>
<s> Students are good and are from various cities</s>
<s> Students from Vellore do engineering</s>

i) Find the Bigram probability of the given test sentence, including <s> & </s> as a token. ii) Compute the perplexity of this bigram.

Test data:
<s> Students are from Vellore cities</s>

Note: Use Laplace (Add-1) Smoothing if needed. Write down vocabulary as well if you use it.

Vocabulary: <s>, I, am, from, Vellore, a, teacher, Students, are, good, and, various, cities, do, engineering, </s>

|V| = 16

Name: Arooj Quiz 1(NLP) Roll Number: 21L-5619

(a)

Solution: <s> Students are from Vellore cities </s>

We use Add-1 smoothing as.

$P(\text{cities}|\text{Vellore}) = 0$

Formula: $P(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + |V|}$

$\Rightarrow P(\text{students}|\langle s\rangle) = \frac{2+1}{4+16} = \frac{3}{20}$

$\Rightarrow P(\text{are}|\text{students}) = \frac{1+1}{2+16} = \frac{2}{18}$

$\Rightarrow P(\text{from}|\text{are}) = \frac{1+1}{2+16} = \frac{2}{18}$

$\Rightarrow P(\text{Vellore}|\text{from}) = \frac{2+1}{3+16} = \frac{3}{19}$

$\Rightarrow P(\text{cities}|\text{Vellore}) = \frac{0+1}{2+16} = \frac{1}{18}$

$\Rightarrow P(\langle /s\rangle|\text{cities}) = \frac{1+1}{1+16} = \frac{2}{17}$

$P(\langle s\rangle \text{ Students are from Vellore cities } \langle /s\rangle) =$

$\left(\frac{3}{20}\right)\left(\frac{2}{18}\right)\left(\frac{2}{18}\right)\left(\frac{3}{19}\right)\left(\frac{1}{18}\right)\left(\frac{2}{17}\right)$

$= 1.91 \times 10^{-6}$

(b) perplexity $= P(\text{sentence})^{-1/N}$

$= (1.91 \times 10^{-6})^{-1/N}$

$\therefore N = 6+1$

$= 8.97$

<s> </s>

18 + 9.5

13.5 + 18

Name: Avooj Quiz 1(NLP) Roll Number: 21L-5619

Q3: Use byte pair encoding on training corpus to generate suitable vocabulary and apply it on our testing corpus and describe how well it segments our testing data. [3+2 Marks]

Training Corpus:
bat cat mat rat hat pat sat fat bat cat

Testing Set:
bat cat mat lat dat

5

Solution:

Vocabulary :- b, a, t, c, m, r, h, p, s, f, at, at_, bat_, cat_

2 bat_
2 cat_
1 mat_
1 rat_
1 hat_
1 pat_
1 sat_
1 fat_

By merging a and t

By merging at and _

By merging b and at_

By merging c and at_

Testing set: let's assume we have:

bat_, cat_, mat_, lat_, dat_

It segment our testing data as:

bat_ as bat_ , cat_ as cat_ , m at_ > mat_ ,

we don't have lat_ and dat_ in our vocabulary (even not by segmenting)

we don't have l and d in our vocabulary only at_.