Sai kalyan
Koshike.

Assignment 2

Sampling

Token	probability
a	0.32
b	0.22
c	0.45

2.1 Greedy Sampling.

Greedy sampling selects the token with the highest probability. Among the given distributing.

$C = 0.45$, $a = 0.32$, $b = 0.22$.

The token C would be generated using greedy Sampling.

2.2 Temperature Sampling.

Temperature is essentially a diversity and creative factor. The lower temperature implies less chance of picking lower prob distribution token.

Convert the probability to logits. This ensures the probability adjustment to temperature factor.

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Using the given probabilities.

Token a: $p = 0.32$.

$$\text{logit}(0.32) = \log\left(\frac{32}{1-032}\right) = \log\left(\frac{32}{68}\right) \approx -0.754$$

token b: $p = 0.22$

$$\text{logit}(0.22) = \log\left(\frac{22}{1-22}\right) = \log\left(\frac{0.22}{0.78}\right) \approx -1.270$$

token c: $p = 0.45$

$$\text{logit}(0.45) = \log\left(\frac{0.45}{0.55}\right) \approx \log(0.8182) \approx -0.201$$

Scale logit by temperature.

Scaled logit of a: $\frac{-0754}{0.2} = -3.75$

Scaled logit of b: $\frac{-1.270}{0.2} = -6.35$

Scaled logit of c: $\frac{-0.201}{0.2} = -1.06$

Apply Softmax

Softmax function converts the scaled logits into a probability distribution

$$P(x) = \frac{e^{\text{logit}(x)}}{\sum e^{\text{logit}(x)}}$$

$e^{-3.75} \approx 0.0235$

$e^{-6.35} \approx 0.0017$

$e^{-1.00} \approx 0.3679$

$$\sum e^{\text{logit}(x)} = 0.0235 + 0.0017 + 0.3679 = 0.3931$$

For a: $\frac{0.0235}{0.3931} \approx 0.0597$

b: $\frac{0.0017}{0.3931} \approx 0.0043$

c: $\frac{0.3679}{0.3931} \approx 0.9340$

For temperature 0.2, the new sampling distribution is

a	0.0597
b	0.0043
c	0.9340

The token is selected randomly, however there is high probability the token C is selected, followed by a the b.

23 Top k sampling

In top k sampling we selected the top k number of tokens and randomly select one from that top k tokens.

Given probabalities

$P(a) = 0.32$
$P(b) = 0.22$
$P(c) = 0.45$

Ordering in descending order

$P(c) = 0.45$
$P(a) = 0.32$
$P(b) = 0.22$

Top $K=2$. we Relect $P(c)=0.45$ and $P(a)=0.32$.

Resultant $\Rightarrow$ $P(c) = 0.45$
top k $P(a) = 0.32$.

Apply Softmax

$$D(x) = \frac{e^{P(x)}}{\sum e^{P(x)}}$$

$a \Rightarrow e^{0.32} = 1.377$
$c \Rightarrow e^{0.45} = 1.568$

$$\sum e^{P(x)} = 1.377 + 1.568 \approx 2.945$$

$$P(c) = \frac{1.568}{2.945} \approx 0.532$$

$$D(a) = \frac{1.377}{2.945} \approx 0.468$$

New probability distribution.

Token	probability
c	0.532
a	0.468

This means token c has a slightly higher chance of being selected, but a still has a notable probability.

3. Byte pair Encoding.

Corpus ["hug", "pug", "pun", "bun", "hugs"]

base vocabulary ['b', g, h, n, p, s, u]

Frequency distribution

(hug, 5) (pug, 15), (pun, 20), (bun, 2), (hugs, 4)

⇒ List of tokens

(h, u, g, 5) (p, u, g, 15) (p, u, n, 20), (b, u, n, 2) (h, u, g, s, 4)

Pair Frequency.

hu ⇒ 5 (hug) + 4 (hugs) = 9

ug ⇒ 5 (hug) + 15 (pug) + 4 (hugs) = 24

pu ⇒ 15 (pug) + 20 (pun) = 35

un ⇒ 20 (pun) + 2 (bun) = 22

gs ⇒ 4 (hugs) = 4

bu ⇒ 2 (bun) = 2.

First Merge Rule

The most frequent pair is pu. we merge pu into a new token

Vocabulary [b, g, h, n, p, s, u, pu]

corpus [(h, u, g, 5), (p, u, g, 15), (pu, n, 20), (b, u, n, 2) (h, u, g, s, 4)
(pu, g, 15)

Now we have some pairs that result in a token longer than two characters: the pair ("pu", "g"), for instance (20 times)

The most frequent pair at this stage is ("u", "g"), however, present "24" times in the corpus. So the second merge rule learned is ("u", "g") ⇒ (ug)

ug ⇒ 5(hug) + 15(Pug) + 4(hugs) = 24.

* Second Merge

Vocabulary [b, g, h, n, p, s, u, pu, ug]

corpus [("h", "ug", 5), ("pu", "g", 15), ("pu", "n", 20), ("b", "un", 2), ("hu", "gs", 4)]

Now

"pun" ⇒ 20(pun) = 20

"pug" ⇒ 15(pug) = 15

"hug" ⇒ 5(hug) + 4(hug) = 9

The most frequent pair is ("pu", n) i.e 20. So we learn merge rule "pun".

* third Merge "pu" + "n" = "pun".

Vocabulary [b, g, h, n, p, s, u, pu, ug, pun]

corpus [('h', "ug", 5) ("pu", "g", 15), ("pun", 20), ("b", "un", 2), ("hu", "gs", 4)]