Course: **Natural Language Processing [A]** 15-June-2023 **(Spring 2023)**

Resource Person: **Muhammad Shakeel** **QUIZ – 1 (Text Processing)**

Total Points: 10

SOLUTION

Suppose we have a corpus that has the words (after pre-tokenization based on space) — *old*, *older*, *finest*, and *lowest*, and we count the frequency of occurrence of these words in the corpus. Suppose the frequency of these words is as follows:

{"old": 7, "older": 3, "finest": 9, "lowest": 4}

Now, create a complete vocabulary using the Byte Pair Encoding algorithm for the above given corpus.

Corpus	**Vocabulary**
7 o l d _	_, d, e, f, i, l, n, o, r, s, t, w
3 o l d e r _	23 10 16 9 9 14 9 14 3 13 13 4
9 f i n e s t _	
4 l o w e s t _	

1. Most frequent pair: **e** and **s** (Frequency: 9+4 = 13)

7 o l d _	_, d, e, f, i, l, n, o, r, s, t, w, es
3 o l d e r _	
9 f i n es t _	
4 l o w es t _	

2. Most frequent pair: **es** and **t** (Frequency: 9+4 = 13)

7 o l d _	_, d, e, f, i, l, n, o, r, s, t, w, es, est
3 o l d e r _	
9 f i n est _	
4 l o w est _	

3. Most frequent pair: **est** and **_** (Frequency: 9+4 = 13)

7 o l d _	_, d, e, f, i, l, n, o, r, s, t, w, es, est, est_
3 o l d e r _	
9 f i n est_	
4 l o w est_	

4. Most frequent pair: **o** and **l** (Frequency: 7+3 = 10)

7	ol d _	_, d, e, f, i, l, n, o, r, s, t, w, es, est, est_, ol
3	ol d e r _	
9	f i n est_	
4	l o w est_	

5. Most frequent pair: **ol** and **d** (Frequency: 7+3 = 10)

7	old _	_, d, e, f, i, l, n, o, r, s, t, w, es, est, est_, ol, old
3	old e r _	
9	f i n est_	
4	l o w est_	

No more merges are possible after this iteration.

END OF QUIZ SOLUTION