COMSATS University Islamabad, Wah Campus

Department of Computer Science

Course : **TICS-II**	Total Marks : **10**
Class : **BCS-8 (A/D)**	Dated : **19/09/2025**
Semester: **FALL 2025**	**Submission Deadline: (22-Sep-2025) In Class**

Assignment (01) (CLO 1)

Name: ______________________

Reg no: _____________________

Format:

- **Make sure this page is the first page of your assignment file.**
- **All answers should start from page 2.**
- **Assignment should be submitted in handwritten & physical format.**
- **Mention both name and roll number on this page.**
- **Give justifications where necessary!**
- **MARKS WILL BE DEDUCTED ON NOT FOLLOWING THE FORMAT.**

Assignment Tasks:

Task 1: Apply Byte Pair Encoding (BPE) Tokenization for the Urdu Language. The corpus is given below. (10)

1. کتاب
2. کتابیں
3. کتابی
4. پڑھنا
5. پڑھا
6. لکھنا
7. لکھا
8. گھر
9. سکول
10. ہم

Task 2. During the merges, some whole words became tokens, while others remained split into characters. What does this show about the relationship between word frequency and tokenization in BPE?

Task 3. If you kept merging pairs for 50 steps instead of 5, what might happen to your vocabulary and how would that affect the balance between memory efficiency and handling rare words?