1. What is a corpus? Describe the features of a corpus from NLP perspective.
2. Describe steps in “Text Normalization”
3. Discuss things to be considered during tokenization.
4. Discuss with an example “Byte-pair encoding” for tokenization
5. With examples illustrate how case folding may or may not be useful during “Word Normalization”
6. Consider transformation from ‘chat’ to ‘had’. Compute MED and backtrace the corresponding alignment. Assume the following:
   - Cost of insertion = cost of deletion = 1
   - Cost of substitution (in case of mismatch) = 2

Regular Expression

7. Write the regex to match these strings
   (a) {gray, grey}
   (b) {babble, bebble, bibble, bobble, bubble}
   (c) {ggle, gogle, google, gooogle, goooogle, ...}
   (d) {google, googoogle, googoogoogle, googoogoogoogle, ...}
   (e) {zzz, zzzz, zzzzz, zzzzzz}
   (f) {zzz, zzzz, zzzzz, ...}
   (g) {0,1,2,3,4,5,6,7,8,9}

8. Write the regex of following string.
   (a) String contains an 11-digit string starting with a 1
   (b) String contains an integer in the range 2…36 inclusive
   (c) String contains a positive integer or floating-point number with exactly two characters after the decimal point.
   (d) String begins with “Btech”
   (e) String ends with “Btech”
   (f) String exactly matches with “Btech”
   (g) a or b or c
   (h) any character except a, b, or c

9. Match the given string with regex
   (a) abcdef42skjhfskjfhsjdfs
   (b) Match the Water in water botte but no the water in water pump
   (c) Word not starting with Um
   (d) A word following a hyphen
   (e) Digits not preceded by a digit, +, or –
   (f) Check whether email address contain ‘@’.