DS504: Natural Language Processing
Quiz 1 - Solutions

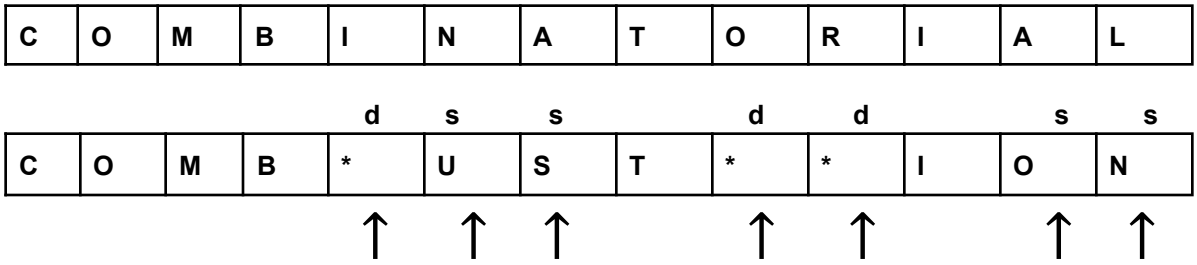
1. *The process of mapping from a word form to its morphological root is called* **Lemmatization**. (1)

2. *Zipf's law gives the relationship between the frequency of a word and* **its position in the list.** (1)

3. *Find the edit distance between the strings "Combinatorial" and "Combustion"? Only mention the distance and operations. No need to show the matrix. You may assume all operations have the same cost.*

   (2)

   Ans: **7**



d : deletion s: substitution

4. *Using Dynamic programming, show the edit distance calculation for strings "tea" and "len"? Please show the matrix. You may assume all operations have the same cost. (3)*

   Ans: 2

		T	E	A
	0	1	2	3
L	1	1	2	3
E	2	2	1	2
N	3	3	2	2

5. *Suppose, we use symmetric delete operations and preprocess the dictionary entries for a maximum edit distance of 2. Given that the dictionary contains 100,000 entries with an average word length of 6, how many entries will be there after the preprocessing? Show the calculations.*

   (2)

   Ans - **2, 200, 000**
   Total symmetric delete combinations for each word= 6C2 + 6C1= 21
   So, for 100,000 entries, we'll have 100,000 * 21 + 100,000 (original words) = 2,200,000 dictionary entries.

6. *What would be the output of applying the regular expression -*
   "[A-z]{1,2}[0-9R][0-9A-Z]? [0-9][ABD-HJLNP-UW-Z]{2}", on the string "BBC News Centre, W70 9RJ, London". (2)

   Ans - The regular expression matches **'W70 9RJ'**

   1. [A-z]{1, 2} matches any one of the characters, uppercase or lowercase at least once and maximum twice - here it matches 'W'
   2. [0-9R] matches any digit or 'R' - here it matches '7'
   3. [0-9A-Z]? matches any digit or capital letter zero times or maximum one time - here it matches '0'
   4. ' ' (space) matches ' ' (space)
   5. [0-9] matches any digit - here it matches '9'
   6. [ABD-HJLNP-UW-Z]{2} matches any two letter pattern where the letters belong to the set specified - here it matches 'RJ'

Note: In the exam, the printing was such that it was not clear whether there is space after '?' or not and if we consider the space is not there, the regular expression does not match anything in the given string. Hence, we will also give marks to the students who wrote **"it matches nothing"**.

However, other answers apart from these two ("**it matches W70 9RJ**" or **"it matches nothing"**), are not accepted.