Sadamala Sasi Rekha
23BDS049

Theoretical Assignment - CS458 (NLP)

Question:1

A) Tokens are the individual words obtained after splitting the text

- Tokens:

The, children, are, playing, in, the, playgrounds, A, child, plays, happily, while, the, others, played, earlier, They, have, been, playing, every, afternoon.

B) Number of types and tokens:

- Types (unique words) : 20

"The|the" and "A|a" are considered the same after case normalization.

- Tokens (total words) : 22.

C) Lemmatize all the nouns:

children → child
playgrounds → playground
child → child
others → other
afternoon → afternoon

D) Lemmatization of verbs:

are → be
playing → play
plays → play
played → play
have → have
been → be
playing → play

Sadanala Sasi Rekha
23BDS049

Question 2:

word pair: kitten → sitting
costs: insertion = 1, Deletion = 1, Substitution = 2

	0	S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
k	1	2	3	4	5	6	7	8
i	2	3	2	3	4	5	6	7
t	3	4	3	2	3	4	5	6
t	4	5	4	3	2	3	4	5
e	5	6	5	4	3	4	5	6
n	6	7	6	5	4	5	4	5

Backtracking:

Starting from cell (n, g) = 5

1) Insert G at end (cost = 1)
2) Match (n → n ⇒ Cost = 0)
3) Substitution: e → i (Cost = 2)
4) match: t → t (Cost = 0)
5) match: t → t (Cost = 0)
6) Substitution: k → s (Cost = 2)

✱ Sequence of operations:

Substitution	(k → s	: 2)
Substitution	(e → i	: 2)
insertion	(g	: 1)

Total cost 2 + 2 + 1 = 5