Course: **Natural Language Processing [A]** 10-August-2023 **(Spring 2023)**

Resource Person: **Muhammad Shakeel** **QUIZ – 4 (Text Similarity and VSM)**

Total Points: 10

SOLUTION

Consider the following corpus with two documents:

Document 1: **The car is driven on the road.**

Document 2: **The truck is driven on the highway.**

Find the TF-IDF values of all words in both documents. Which words are termed significant by TF-IDF?

Word	**TF**		**IDF**	**TF*IDF**	
	Doc 1	**Doc 2**		**Doc 1**	**Doc 2**
The	2/7	2/7	log(2/2) = 0	0	0
car	1/7	0	log(2/1) = 0.3	0.043	0
truck	0	1/7	log(2/1) = 0.3	0	0.043
is	1/7	1/7	log(2/2) = 0	0	0
driven	1/7	1/7	log(2/2) = 0	0	0
on	1/7	1/7	log(2/2) = 0	0	0
road	1/7	0	log(2/1) = 0.3	0.043	0
highway	0	1/7	log(2/1) = 0.3	0	0.043

From the above table, we can see that TF-IDF of common words was zero, which shows they are not significant. On the other hand, the TF-IDF of "car", "truck", "road", and "highway" are non-zero. These words have more significance.