**TF-IDF Assignment Questions and Answers:**

**Question -1: Why are we adding “1” in the numerator and denominator in the IDF formula?**

**Answer: sklearn module uses this formula to avoid zero division**

**Question -2: What is meant by Normalization? What is the difference between L1 and L2 normalization?**

**Answer:**

1. **Finding the length of the vector from origin is called normalization.**
2. **L1 normalization means the sum of absolute values of x, y, z and so on(n-dimensional) components of the vector**
3. **L2 normalization means the square root of sum of the squared values of x,y,z and so on(n-dimensional) components of the vector**

**Question -3: What is sparsity of a given matrix? Why do we use sparse matrix representation when we have a lot of zeros in the given matrix?**

**Answer:**

1. **For all the non zero values of the matrix, storing the row number, column number and non zero value is called sparsity of a given matrix.**
2. **This way we can save lot of memory if we deal with huge matrices. If an element in the matrix is zero then no need to preserve zero in the memory, it consumes lot of memory and time while dealing with huge matrix operations.**

**Question-4: Which of the following will be True, regarding two sentences S1 and S2**

**S1 - He asked why the government should be allowed to set up a committee .**

**( tf value of word ‘committee’ in S1 = t1, idf value of word ‘committee’ in S1=i1)**

**S2 - The committee will work according to the rules defined by the previous committee.**

**( tf value of word ‘committee’ in S2 = t2, idf value of word ‘committee’ in S2=i2)**

**a) t1>t2, i1>i2**

**b) t1<t2, i1<i2**

**c) t1>t2, i1=i2**

**d) t1<t2, i1=i2**

**Answer: d**

**Question- 5: Can you show mathematically why you are getting the value of IDF for top 50 features as 6.92 in TASK-2 ?**

**Hint - Find and substitute the value of N and n given in equation of que 1.**

**Answer: For example, the below calculation is for the the unique word “dysfunction”**

**Total no of documents in the corpus = 746**

**Th no of documents the word “dysfunction” contains = 1**

**As per sklearn module, idf = 1+lne((1+no of docs in corpus)/(1+no of docs that word exists))**

**idf for “dysfunction” = 1+lne((1+746)/(1+1))**

**= 1+lne(373.5)**

**= 1+5.9229**

**=6.9229**

**So, idf for “dysfunction” in the corpus = 6.9229**