EXPERIMENT 4

AIM: To implement tf-idf in Information Retrieval.

RESOURCES REQUIRED: Jupyter NoteBook, 4GB RAM and above, i5 Processor and above.

THEORY:

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction. TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be defined as the calculation of how relevant a word in a series or corpus is to a text. The meaning increases proportionally to the number of times in the text a word appears but is compensated by the word frequency in the corpus (data-set).

Terminologies:

Term Frequency: In document d, the frequency represents the number of instances of a given word t. Therefore, we can see that it becomes more relevant when a word appears in the text, which is rational. Since the ordering of terms is not significant, we can use a vector to describe the text in the bag of term models. For each specific term in the paper, there is an entry with the value being the term frequency.

Document Frequency: This tests the meaning of the text, which is very similar to TF, in the whole corpus collection. The only difference is that in document d, TF is the frequency counter for a term t, while df is the number of occurrences in the document set N of the term t. In other words, the number of papers in which the word is present is DF.

Inverse Document Frequency: Mainly, it tests how relevant the word is. The key aim of the search is to locate the appropriate records that fit the demand. Since tf considers all terms equally significant, it is therefore not only possible to use the term frequencies to measure the weight of the term in the paper. First, find the document frequency of a term t by counting the number of documents containing the term:

Term frequency is the number of instances of a term in a single document only; although the frequency of the document is the number of separate documents in which the term appears, it depends on the entire corpus. Now let's look at the definition of the frequency of the inverse paper. The IDF of the word is the number of documents in the corpus separated by the frequency of the text.

Term Frequency-Inverse Document Frequency:

Given a corpus D, a term t_i and a document $d_j \in D$, we denote the number of occurrences of t_i in d_j by tf_{ij}. This is referred as the term frequency.

The inverse document frequency for a term t_i is defined as

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

where,

$| D |$ is the number of documents in our corpus, and

$| \{d : t_i \in d\} |$ is the number of documents in which the term appears.

If the term t_i appears in every document of the corpus, idf_i is equal to 0. The fewer documents the term t_i appears in, the higher the idf_i value.

The measure called term frequency-inverse document frequency (*tf-idf*) is defined as $tf_{ij}*idf_i$ (Salton and McGill, 1986). It is a measure of importance of a term t_i in a given document d_j. It is a term frequency measure which gives a larger weight to terms which are less common in the corpus. The importance of very frequent terms will then be lowered, which could be a desirable feature.

CONCLUSION: Hence we have successfully implemented Term Frequency Inverse Document Frequency.