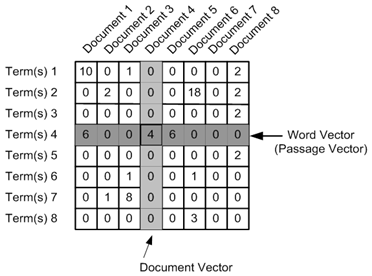
A continuación se obtiene la frecuencia de término (term frequency o tf ), y se representa en una matriz llamada matriz término-documento (document-term matrix o dtm)[figura], básicamente es una especie de ”tabla” en la que aparece una fila por cada término existente, y una columna por cada documento. Cada “fila” representa un vector de palabras, ya que en cada posición contiene la cantidad de veces que esa palabra aparece en cada documento. Cada “columna” representa un vector de documento, ya que cada posición contiene la cantidad de veces que aparece cada término.



Por otra parte se obtiene la frecuencia inversa de documento (inverse document frequency o idf), una medida que representa si determinado término es común o no en una colección de documentos. Se obtiene dividiendo el número total de documentos por el número de documentos que contienen el término, y se toma el logaritmo de ese cociente. El resultado es un vector con una posición por cada término, cuyos valores serán el resultado de la operación mencionada anteriormente.

Finalmente se obtiene la matriz tf-idf, el resultado de multiplicar cada valor tf por su idf, en otras palabras, multiplicar los valores de cada “columna” de la dtm por el vector idf.

Como el cociente dentro de la función logaritmo del idf es siempre mayor o igual que 1, el valor del idf (y del tf-idf) es mayor o igual que 0. Cuando un término aparece en muchos documentos, el cociente dentro del logaritmo se acerca a 1, ofreciendo un valor de idf y de tf-idf cercano a 0.

En resumen, un peso alto en tf-idf se alcanza con una elevada frecuencia de término en el documento dado y una pequeña frecuencia de ocurrencia del término en la colección completa de documentos.