28.7 TF-IDF (Term Frequency - Inverse Document Frequency)

Term Frequency (TF)

Let us assume we have the words 'w_1', 'w_2', 'w_3', 'w_4', 'w_5' and 'w_6' and there are 'N' documents in the corpus.

Term Frequency(w_i, r_j) = (Number of times 'w_i' occurs in 'r_j')/(Total Number of Words in 'r_j')

Let us assume we have 2 reviews 'r_1' and 'r_2' and the words in them be

r_1: w_1 w_2 w_3 w_2 w_5

r_2: w_1 w_3 w_4 w_2 w_6 w_5

So $TF(w_2,r_1) = 2/5$

$TF(w_2,r_2) = 1/6$

The term frequency of any word in general lies in between 0 and 1 (inclusive). So as this value lies in between 0 and 1, we can interpret it as probability. So $TF(w_i, r_j)$ can also be called as probability of occurrence of the word 'w_i' in 'r_j'.

Inverse Document Frequency (IDF)

Let $D_c \rightarrow$ Data of corpus $\rightarrow \{r_1, r_2,.....,r_n\}$

Inverse Document Frequency of a word is defined over the corpus, but not over a document.

$IDF(wi,D_c) = \log_e$(Total Number of Documents(N)/Total Number of reviews containing 'w_i')

So $IDF(w_i,D_c) = \log_e(N/n_i)$

We know that $n_i <= N$, so $N/n_i >= 1$.

So it means $\log_e(N/n_i) >= 0$

If 'n_i' increases, 'N/n_i' decreases and ultimately $\log_e(N/n_i)$ decreases.

$\log_e(N/n_i)$ is a monotonically decreasing function in 'n_i'. The more the word 'w_i' across the reviews in the corpus, the lesser is the IDF score.

If 'w_i' is more frequent, $IDF(w_i,D_c)$ will be low.

If 'w_i' is a rare word, $IDF(w_i,D_c)$ will be more.

Vector Creation for each document using TF-IDF

In vector creation using TF-IDF, the magnitude of each dimension 'w_i' is given by

Magnitude of 'w_i' = $TF(w_i, r_i)$*$IDF(w_i, D_c)$

In a nutshell, we are giving more priority to the words which occur most frequently (whose TF value is high), and the words which occur rarely (whose IDF value is very low). By this, we can maximize the value of the product TF*IDF.

Drawbacks of TF-IDF

One of the drawbacks of TF-IDF is, it also doesn't take the semantic meaning into consideration. (ie., words like (cheap, affordable), (tasty, delicious), (valuable, precious) are considered as separate dimensions)

Difference between BOW and TF-IDF

In both BOW and TF-IDF, we convert each of the reviews into a d-dimensional vector where 'd' is the number of unique words in the total text across all the reviews.

The key difference between BOW and TF-IDF is that instead of using frequency counts as the values in the d-dimensional vector for each word as in BOW, we use the TF-IDF score for the words in TF-IDF vector representation.